

Artificial Intelligence Risk Certificate



Book By Francisco Bettencourt

Used for AI Risk Certification Exam

Artificial Intelligence Risk Certificate

Contents

Module 1 - AI and Risk – Introduction and Overview	17
1.0 Classical AI	18
1.1 Specific versus General AI	19
1.2 Good Old-Fashioned AI (GOFAI) and Classical AI	21
1.3 Simple Reinforced Learning	23
1.4 Lookahead	24
1.5 Search	25
1.6 Recursion.....	27
1.7 Recursive Adversarial Tree Search	28
1.8 Complexity, heuristics and Refinement by Reinforcing Learning	29
1.9 Limits of Classical Artificial Intelligence	31
2.0 Neural Networks	33
2.1 Artificial Neurons	34
2.2 Connectionism and its Early Challenges	37
2.3 Deep Learning Proves its Potential	39
2.4 Deep Learning Beats Symbolic AI at its Own Game	40
2.5 The inscrutability of Deep Learning	42
2.6 The Dwan of Artificial General Intelligence?	44
3.0 Machine Learning and its Risks	46
3.1 Four Types of Machine Learning	46

Artificial Intelligence Risk Certificate

3.2 Examples of Unsupervised Learning (Principal Component Analysis).....	48
3.3 Risk of Inscrutability.....	51
3.4 Risk of Over Reliance	52
3.5 Risk To individuals, Organizations and Society	53
Questions and Answers Module 1 from GARP	55
Module 2 – Tools and Techniques	57
1.0 Machine Learning	58
1.1.1 ML, Classical Statistics and Econometrics.....	59
1.2 Four Types of Machine Learning	61
1.3 Exploratory Data Analysis.....	63
1.3.1 Data Collection and Preparation	64
1.3.2 Data Cleaning	67
1.3.3 Data Visualization.....	69
1.3.4 Feature Extracting.....	72
1.3.5 Data Scaling.....	75
1.3.6 Data Transformation.....	77
1.4 Dimensionality Reduction Techniques.....	78
1.4.1 Principal Component Analysis	78
1.5 Training Validation and Testing	82
1.5.1 Sample Splitting and Preparation	83
1.6 Software for Machine Learning	84
Questions and Answers Tools and Techniques – Module 2 GARP	86

Artificial Intelligence Risk Certificate

2.0 Unsupervised Learning	91
2.1 K-Means Clustering Algorithm.....	96
2.1.1 Performance Measurement for K-Means	102
2.1.2 Selecting the Starting Positions of the Centroids	102
2.1.3 Selection of K	103
2.1.4 Selection of K Example	104
2.1.5 Advantages of K Means.....	105
2.1.6 Problems with K Means.....	105
2.1.7 Fuzzy K Means	106
2.2 Hierarchical Clustering.....	107
2.3 Density Based Clustering	110
Extra 2.A Different Distance Measurement.....	111
Extra 2.B Silhouette Method	113
Extra 2.C K-Means Clustering Example	114
Questions and Answers Module 2 Chapter 2 from GARP	118
3.0 Supervised Learning for Numerical Data	123
3.1.1. Simple Linear Regression.....	129
3.1.2. Multiple Linear Regression.....	131
3.1.3. Potential problems with Regressions	137
3.1.4. Stepwise Regression Procedures.....	142
3.2. Classification Problems	145
3.2.1. Logistic Regression	146

Artificial Intelligence Risk Certificate

3.2.2. Other Types of limited Dependent Variable Models	148
3.3 Linear Discriminant Analysis	149
Appendix 3.A The Heckman 2 stage Procedure.....	152
Appendix 3.B Fisher Discriminant Analysis	153
Appendix 3.C Linear Discriminant Analysis Example	155
Questions and Answers Module 2 Chapter 3 from GARP	159
4.0 Supervised Learning part 2: Machine Learning Techniques.	
.....	163
4.1 Decision Trees.....	164
4.1.1 Regression Trees	165
4.1.2 Classification Trees.....	167
4.1.3 Classification Trees Example	168
4.1.4 Pruning.....	173
4.1.5 Ensemble Techniques	175
4.2 K Nearest Neighbors	178
4.3 Support Vector Machines	181
4.3.1 Support Vector Machines Example	184
4.3.2 Support Vector Machines Extensions	186
4.4 Neural Networks.....	186
4.4.1 The choice of Activation Function.....	189
4.4.2 A numerical Example	191
4.4.3 Backpropagation	193
4.4.4 Architectural Issues.....	194

Artificial Intelligence Risk Certificate

4.4.5 Overfitting	196
4.4.6 Advanced Neural Network Structures.....	197
4.5 Autoencoders.....	201
Appendix 4.A Technical details of how SVM are Determined...	206
Questions and Answers Module 2 Chapter 4 from GARP – Machine Learning Techniques.....	209
5. Semi supervised Learning	216
5.1 Introduction to Semi Supervised Learning	216
5.2.1 Semi Supervised Learning Assumptions	217
5.2.2 Semi Supervised Learning Techniques.....	220
5.2.3 Self-Training.....	221
5.2.4 Co-Training	225
5.2.5 Unsupervised Pre processing.....	230
Appendix 5.A Semi Supervised Learning Assumptions.....	231
Questions and Answers Module 2 Chapter 5 from GARP – Semi Supervised Learning.....	233
6. Reinforcement Learning	234
6.1 The Principles of Reinforcement Learning	235
6.2. The Multi-Arm Bandit Problem.....	237
6.2.1. Terminology in MABs.....	237
6.2.2. Strategy in MAB	240
6.3 Markov Decision Processes.....	245
6.4 Approaches to Reinforcement Learning.....	248

Artificial Intelligence Risk Certificate

6.5.1 The Bellman Equations	249
6.5.2 The Monte Carlo Method	250
6.5.3 The Temporal Difference TD Method.....	252
6.5.4 Illustrative Example	253
6.5.5 Curse of Dimensionality and Neural Network Approximation	254
6.6 Chapter Summary	255
Appendix 6.A Markov Transition Probabilities	256
Appendix 6.B Detailed Reinforcement Learning Example- The game of Nim	258
Questions and Answers Module 2 Chapter 6 from GARP – Reinforcement Learning	268
7.0 Supervised Learning – Model Estimation	272
7.1.1 Ordinary Least Squares	275
7.1.2 Nonlinear Least Squares	280
7.1.3 Hill Climbing.....	282
7.1.4 The gradient Descent Method	283
7.1.5 Illustration of the Gradient Descent Method.....	287
7.1.6 Backpropagation	288
7.1.7 Computational Issues	290
7.2 Maximum Likelihood	292
7.3 Overfitting, Underfitting and Bias Variance trade off.....	294
7.3.1 Overfitting	294

Artificial Intelligence Risk Certificate

7.3.2 Underfitting	295
7.3.3 Bias variance Trade Off	296
7.3.4 Prediction Accuracy vs. Interpretability	300
7.4 Regularization	301
7.4.1 Ridge Regression.....	302
7.4.2 LASSO.....	302
7.4.3 Elastic Net.....	304
7.4.4 Regularization Example	304
7.5 Cross Validation and Grid search	306
7.5.1 Cross Validation	306
7.5.2 Stratified Cross Validation	308
7.5.3 Bootstrapping	310
7.5.4 Grid Searches	311
Appendix 7.A Genetic Algorithms.....	313
Questions and Answers Module 2 Chapter 7 from GARP – Model estimation.....	316
8.0 Supervised Learning – Model Performance Evaluation....	320
8.1 Model Evaluation when the Output is continuous	320
8.1.1 An example of Continuous Variable Model Performance Comparison	325
8.2 Model Evaluation: Classification	326
8.2.1 An Example of Model Evaluation: Classification	331

Artificial Intelligence Risk Certificate

Questions and Answers Module 2 Chapter 8 from GARP – Model Performance Evaluation	334
9.0 Natural language processing	339
9.1 Data Pre-Processing.....	341
9.2 NLP Models	345
9.2.1 Feature Extraction	345
9.2.2 Vector Normalization	348
9.3 Dictionary Comparison approaches	349
9.3.1 Advantages and Disadvantages of the Dictionary Approach	351
9.3.2 N-Grams	352
9.3.3 Term Frequency-Inverse document Frequency.....	353
9.4 Machine learning approaches	357
9.4.1 The Naïve Bayes Classifier	358
9.4.2 The Naïve Bayes Example	361
9.4.3 Words Meanings.....	369
9.5 NLP Evaluation	370
Appendix 9.A Naïve Bayes application Problem	371
Questions and Answers Module 2 Chapter 9 from GARP – Natural language processing	375
10. Generative Artificial Intelligence.....	378
10.1 Introduction.....	379
10.2 A simple taxonomy for GenAI	379

Artificial Intelligence Risk Certificate

10.3 Word Embedding, Word2Vec and RNNs	383
10.3.1 Word2Vec	384
10.3.2 RNNs	388
10.4 Transformers and LLMs	391
10.4.1 Large Language Models	394
10.4.2 Cloud Based LLMs	398
10.4.3 Chatbots	400
10.4.4 Using LLMs - Prompt engineering and Temperature.....	401
10.5 Applications of Generative AI and LLMs.....	403
10.6 Chapter Summary	406
Appendix 10.A Operations with word Embeddings	407
Questions and Answers Module 2 Chapter 10 from GARP – Generative Artificial Intelligence.....	409
Module 3 – Risk and Risk Factors	413
1.0 Introduction.....	413
2.0 Algorithmic Bias and Fairness.....	415
2.1 What is Bias?	415
2.2 What is fair?.....	416
2.2.1 Individual Fairness	417
2.2.2 Group Fairness	419
2.3 Source of Unfairness.....	427
2.3.1 Problem Specification and Feature selection	428
2.3.2 Data Collection and Data Composition.....	433

Artificial Intelligence Risk Certificate

2.3.3 Model development	436
2.3.3 Model deployment	437
3.0 Explainability, Interpretability and Transparency.....	438
3.1 Black Box problem	439
3.2 Opaqueness	440
3.3 Explainable AI (XAI).....	441
4.0 Autonomy and Manipulation	443
4.1 Autonomy	443
4.2 Manipulation	444
5.0 Safety and Well-Being	445
6.0 Reputational Risk	447
6.1 Causes for AI- related reputational Damage	447
6.2 Types of AI-related criticism companies face	448
6.3 Management and Mitigation Strategies	449
7.0 Existential Risks	450
8.0 Global Challenges and Risks.....	452
8.1 Economic risks from AI	452
8.2 Global Inequality.....	453
8.3 Misinformation Campaigns	454
8.4 Privacy and Surveillance	454
9.0 Conclusion	454
Questions and Answers Module 3 from GARP -Risk and Risk Factors	455

Artificial Intelligence Risk Certificate

Module 4 – Responsible and Ethical AI	461
1.0 Introduction.....	462
2.0 Practical ethics	462
2.1 Why might a firm consider a Practical Ethics Framework..	464
3.0 Ethical Frameworks.....	465
3.1 Consequentialism	466
3.2 Deontology.....	468
3.3 Virtue Ethics	469
4.0 What can AI ethics learn from medical Ethics?	471
5.0 Principles of AI ethics	473
5.1 Nonmaleficence	473
5.2 Beneficence	474
5.3 Justice.....	475
5.4 Autonomy	477
5.5 Explainability	478
6.0 Bias, Discrimination and Fairness	480
6.1 Problematic Biases	481
6.1.1 Biases in Problem Specification	482
6.1.2 Biases in data	482
6.1.3 Biases Modeling, Validation and Algorithm Design.....	484
6.1.4 Biases in Deployment	484
6.2 When does bias count as discrimination?.....	485

Artificial Intelligence Risk Certificate

6.3 Fairness	486
6.4 Avoiding problematic Biases and Unfairness	487
7.0 Privacy and Cybersecurity.....	489
7.1 Why is Privacy an Ethical Issue?.....	490
7.2 Principles and Good practices.....	492
8.0 Governance Challenges	494
8.1 Power Asymmetries	494
8.2 Institutional opaqueness.....	494
8.3 Algorithm Opaqueness	495
8.4 lack of AI ethic structures.....	496
8.5 Lack of National and International Regulation	497
8.6 Unpredictability Issues	497
8.7 Lack of Truth Tracking Abilities	498
8.8 Privacy and Copyright	498
9.0 Regulatory Landscape	499
9.1 The relationship between Ethics and Law	499
9.2 Europe	500
9.3 United States.....	505
9.4 China.....	508
Questions and Answers Module 4 from GARP -Responsible and Ethical AI	509
Module 5 – Data and AI governance	515
1.0 Introduction.....	516

Artificial Intelligence Risk Certificate

2.0 Data Governance.....	517
2.1 Data strategy: Developing Vision, and Setting goals and priorities.....	517
2.2 Data Quality: Accuracy, Consistency and Integrity.	518
2.3 Data Provenance: verifying the legal right to use data, especially alternative data	519
2.4 Data Classification: Structure and Confidentiality	520
2.5 Metadata Management: Collection, Documentation and management	521
2.6 Data protection, security and compliance: Regulatory Aspects and Overview.	521
2.7 Data Access: Permission and Secure and Effective Data Sharing	522
2.8 Compliance: Ensuring Compliance with legal and regulatory requirements when handling data	523
2.9 Roles and Responsibilities.	524
3.0 Model Governance	525
3.1 Model Development and testing	527
3.1.1 who is responsible for testing QRMs?	529
3.1.2 How are White Box and Black Box tests conducted?.....	531
3.1.3 What is the use test?	532
3.2.1 Model Validation: What is model Validation and when it should be performed?	534
3.2.2 Should there be limits on Model Use?	538

Artificial Intelligence Risk Certificate

3.3.1 Model Governance Policies	540
3.3.2 Model Documentation	541
3.3.3 Model Inventory and Model Landscape.....	541
3.4 Roles and Responsibilities.....	547
3.4.1 Communication and Interaction	551
3.5 Model Review and Model Changes.....	551
3.6 Recommendations for Establishing Model Risk Governance Framework for ML/AI	557
4. Model Validation	558
4.1 Design: Bad choice and Misspecification, Parameter uncertainty	560
4.1.1 Motivation	560
4.1.2 Model Developers	561
4.2 Modeling: Numerical and Statistical Issues	563
4.2.1 Discretization	563
4.2.2 Approximations	564
4.2.3 Numerical Evaluation	565
4.2.4 Random Numbers	566
4.3 Implementation: Software Engineering Data	567
4.3.1 Model Implementation Tasks.....	567
4.3.2 Model Adaptation Tasks.....	568
4.4 Processes and Misinterpretation of Results	572
4.4.1 Running a QRM	572

Artificial Intelligence Risk Certificate

4.4.2 Misinterpreting the results of QRM	573
4.5 Final Thoughts.....	575
Questions and Answers Module 5 from GARP -Data and AI Governance	577

Module 1 - AI and Risk – Introduction and Overview

Key topics Being covered on this chapter include:

How to explain some central principles around classic AI, including search methodologies and recursion.

Describe, at a high level, how reinforcement learning works.

Describe, at a high level, how neural networks work, and how they differ from classical AI systems.

Articulate the potential and limitations of deep learning.

Identify key breakthroughs leading to advances in AI and ML.

Compare and contrast reinforcement, supervised, and unsupervised learning, and identify practical applications for each technique.

Discuss Risks associated with inscrutability in AI and ML.

Discuss Risks associated with over-reliance on AI systems

Discuss ways in which AI exposes individuals, organizations and society to risk.

Artificial Intelligence Risk Certificate

1.0 Classical AI

Intelligence or reason is a quality that distinguishes humans from other species. There has been always the speculation about the possibility of duplicating the power of human reason with artificial technology. This concept began to seem feasible after the industrial revolution, where Charles Babbage conceived a mechanical, gear-driven “difference engine” that could replace error-prone human calculations in the efficient production of mathematical tables.

Later, the analytics engine is widely seen as anticipating the digital computer, through bringing it to successful completion would be a huge feat of engineering , even with moder technology.

The Crucial breakthrough occurred when the digital computer emerged as the brainchild of Alan Turing in 1936, although his “computing machine” was a theoretical invention.

It was designed to prove fundamental results about mathematics and what we now call the theory of computation. Such machine would be able to follow any given recipe for a calculation that we were able to devise, and in that sense would be a universal programmable computer.

Turing presented his visionary ideas and thoughts about the tremendous potential power of intelligent machines on his paper, “Computing machinery and intelligence” dated 1950. This Centers on an idea of an imitation

Artificial Intelligence Risk Certificate

game, a.k.a Turing Test, in which a computer proves itself to be intelligent if it can generate textual conversations of a quality indistinguishable from that of an intelligent human, raging over any topic chose by the Human “Interrogator”.

This is highly controversial philosophically but has become more salient recently due to chatbots based on Large Language Models (such as ChatGPT). These ones have provided, for the first time, examples of computer programs that are relatively plausible contenders to achieve such high levels off intelligence.

1.1 Specific versus General AI

Media discussion and futuristic speculation around AI, is often concerned with the question of whether an artificial system could surpass human general intelligence. This Concern originates from the legacy of Turing Test, but it would probably reflect a natural concern about challenges to our supremacy in reason and understanding, which are abilities that are so bounded to our species identity.

Such concepts of “fear” associated, often linked to the existential risks arising from Artificial General Intelligence (AGI). It is far less threatening to consider that AI can surpass human reasoning in specific tasks, such as calculating mathematical tables or generating complex statistics. But, it is this type of Machine, the one

Artificial Intelligence Risk Certificate

designed to be intelligent at solving specific problems, that will mostly concern us in what follows.

Some will count a system as intelligent, if this one is provided of the general ability to be intelligent. It has become commonplace to talk of intelligent or “Smart” systems that are relatively limited. Even Turing’s legacy does not point very heavily towards a purely general interpretation of AI.

In 1950, on his paper, the idea was to demonstrate the possibility of such a computer system that could not be reasonably denied as being intelligent. In this sense, it makes clear sense to consider a system that was based on conversational principles, that would allow to compare in a versatile way against the reasoning and skills of a Human, as a sort of conceptual existence proof of that possibility.

But once we have been convinced that there is no objection in principal to artificial intelligence, it seems odd to deny that a system can be intelligent in addressing specific tasks, without having more general competence.

1.2 Good Old-Fashioned AI (GOFAI) and Classical AI

This includes techniques that try to automate the sort of reasoning we do.

The term artificial intelligence was coined in 1956 by John McCarthy of Dartmouth College. Echoing Turing's notion of universality, this has been the conjecture that every aspect of learning or any feature of intelligence, can in principle be so precisely described that a machine can simulate it. The paradigm of intelligence was accordingly explicit, precise processing, as typified by symbolic logic or other formal rule-governed system.

In 1936, Turing had emphasized their ability to mimic logical proofs, constructed step by step from a given set of "Axioms" and systematically applying such rules. Following in this tradition, in 1955 Allen Newell, Herbert Simon and Cliff Shaw (Rand Corporation) started developing their Logic Theorist Program, designed to generate proofs of propositional logic, the most simple and fundamental form of Modern Symbolic Logic. Their work was used to prove the Logistic Thesis, that all of mathematics could be deducted from suitable axioms by pure logic, so it was an excellent choice of high-profile application of machine intelligence.

Another paradigm aspiration of AI was to create a program capable of playing "intellectual" chess at a high level.

Artificial Intelligence Risk Certificate

Alan Turing have tried to implement this, but his ideas never left the paper due to their engineering limitations at the time. Later, in 1997 Deep Blue own against Garry Kasparov due to more powerful machines and sophisticated techniques.

Quite apart from their intrinsic interest and their promotional value in raising public awareness of the possibility of AI, such intellectual games provided an attractive testbed for the development of AI for at least three reasons.

First, they are relatively Simple, compared to the complexities of day-to-day life, and hence could plausibly be tackled even when computing resources would have been severely stretched in even recording the details of a physical environment, let alone taking on the immensely complicated challenge of reasoning about it.

Secondly, such games are straightforwardly rule-governed, with clear and explicit regulations, regarding the setup of the game, how the pieces are permitted to move, the effect of such moves on the board position and criteria for winning, losing or drawing.

Thirdly, they are competitive, and we already have the idea that players can be rated according to their relative skill level.

Hence, such games, lend themselves to very naturally to the assessment of that skill, to that it is straightforward to judge researchers progress as they endeavor to

Artificial Intelligence Risk Certificate

produce a program capable of competing at the higher levels.

1.3 Simple Reinforced Learning

Classic Ai works in a way that we can understand well, using logical reasoning, clear representation of data, and explicit calculation. This differs quite well from neural networks; it is vital to be aware of the differences to understand the majority of risks that contemporary AI involves

Reinforcement Learning, plays a huge role in Contemporary AI. Reinforcement learning is where an agent, that can be a human or AI, that by applying different actions, that generate an outcome, and by observation experience tries to understand which works better.

Successful actions are positively reinforced, which are more likely to occur in the future, and vice versa. The likelihood of such action is contained in some type of numerical expression.

In modern machine learning the role of these weights are hard to understand.

So, from the game example, if a computer that is based on supervisory learning plays a random opponent, will eventually become much better playing, but his capacity of learning will be biased by the fact that the player sometimes loses games it should win, meaning that

Artificial Intelligence Risk Certificate

the computer is getting positively reinforced on incorrect answers.

Now, if the player is also a reinforcement learning agent, this means that will always try to achieve the Successful action, and that his actions, since is being successful, are being constantly reinforced. By this, the computer, playing against a perfect player, can in a much faster pace, improve its capacity to play, since it is being positively reinforced on Successful Actions.

The later, means that the computer has achieved a perfect capacity of prediction , while in the random scenario, the computer chances of winning are x times higher, but since learns from mistakes, it is not perfect.

1.4 Lookahead

The reinforcement learning was playing noticeable but not on an intelligent way. Searching amid all possibilities, and search/look ahead for opportunities.

Binary search, the possibility range is divided by 2. (Decision Three).

Breadth first search explores all the routes in parallel, eliminating as we go, those that can be the sub-optimal.

Depth-Search we explore each route in turn to the end (while avoiding loops- these are the options we crossed out)

Artificial Intelligence Risk Certificate

A* uses a heuristic – here an estimate of time still to travel, based on crow-flies distance allows us to assess how far a station is from oxford, in this case choosing the one closest to Oxford .

Lookahead become essential when our aim is to form a multistep plan to achieve some goal, within a context where there is a very large number of possible situations overall, but with a relatively constrained and predictable range of options in any particular situation.

We follow a sequence of implications with the idea that eventually we will either find a contradiction, which lead us to rewind/backtrack the path chose, or we will generate a complete solution. Focus on the promising of using a strategic planning approach, whereby we try to identify what short-term outcomes would usefully contribute to the completion of our task.

1.5 Search

Working through various options, we are conducting a search through the possible outcomes to find one that will best achieve our aims. Generating sets of possible outcomes and searching through them is one of the most fundamental techniques in AI.

In a binary search, the space is progressively divided in two until the desired target has been found. Whenever a guess is unsuccessful, the “Search Space” is reduced by at least at half, removing the part that has been ruled out

Artificial Intelligence Risk Certificate

by the user's response. Assuming that the answers are consistent the x number of times needed will equal to:

$$x = \text{roundUP}(\log_2(\text{universe space}))$$

A more complicated way is attempting to find the optimal route from one city to another by taking an appropriate sequence of journeys, with no obvious best combination to reach our destination.

Disadvantages of using breadth-first search is that it is inefficient and expensive in memory, whereas using depth-first search may waste time focusing on poor options initially while neglecting short and straightforward paths that the breadth- first search would have found far more quickly.

Better than the above is to employ methods of heuristic search, where we take advantage of some additional information, to help us decide which options are worth exploring first. The best-known example is the Technique A* algorithm, in which the locations are kept ordered according to the sum of the distance travelled so far and an optimistic estimate of the distance still to go, and exploration of ongoing routes always starts from the location that currently has the shortest sum.

Heuristics techniques such as the one presented, can indeed help in facilitating many search problems, but unfortunately in more complex or demanding cases the search space grows so enormously as the relevant number of possible inputs increases.

Artificial Intelligence Risk Certificate

Leading to a combinatorial explosion as the relevant number of possible choices get multiplied together, making the certainty of a solution unfeasible.

1.6 Recursion

Fundamental to computational science and base for classical Artificial Intelligence.

Divides the complex task into a simple task... Its convenient to be the same task category. A recursive function is a function that calls itself. Reducing the complicated task by reducing them to simple examples of themselves

Tower of Hanoi is one of the most famous and elegant illustrations of the recursive technique, whereby a complex problem is solved by being progressively reduced to simpler instances of the same kind of problem.

This simple line of reasoning provides a complete solution to the tower of Hanoi, dividing every problem of moving a pile of n disks into two smaller problems of moving pile of n-1 disks (plus one move of disk n) so that in the end everything is reduced down to single disk movements. Moving the entire pile of n disks by this method is given by:

$$2^n - 1$$

1.7 Recursive Adversarial Tree Search

Recursion can also be applied to solve the relatively complex problem of playing an adversarial game, by combining lookahead and search.

On TIC TAC TOE, is crucial to devise a function that can assess the value of any position to the player whose turn it is to move in that position. Red, should always play the move that gives Blue the lowest possible value in the remaining n-1 positions.

Adversarial search is where red is supposed to look for whichever move is least advantageous for her opponent.

The above is only true if we have an oracle, infallible method. In the real world, what we should do is then repeat the same sort of reasoning at the next level, as to consider a possible, suboptimal, continuation from the previous position.

This strategy should be carried out until where Blue has an immediate winning move.

Consequences of this line of thinking:

If we can evaluate any n-1 position, then we can evaluate a n position.

This pattern continues, with the upshot that if we can evaluate any 0-position (i.e. position with no remaining moves to be played), then we can evaluate any 1-position, and so on until 9- position

Artificial Intelligence Risk Certificate

Apart from this, we need to be able to recognize when a game has ended owing to completion of a line (yielding a position of -1 to the player whose turn it would otherwise be).

1.8 Complexity, heuristics and Refinement by Reinforcing Learning

The previous example, tic tac toe, that is an assessment algorithm that is capable of exhaustively analyzing all the possible lines of play in a game, implicating creating a “search tree”, and keeping track of those positions value can be implemented so simply (less than 25 lines of code).

There is a catch to this, that is the curse of exponential complexity that leads to a combinatorial explosion. For more complex games, the number of possibilities per Move is just so large that will take forever with today’s computational capacity.

As Arthur Samuel IBM, a practicable checkers program could not possibly work by exhaustive search through all possible combinations of moves.

At any point in the game position, it should limit its exhaustive search to n moves and then to assess the positions reached through this analysis, and then select which path is best to follow) by using heuristics. In other words, calculable criteria or rules of thumb that can usefully guide the choice of position to aim for, even

Artificial Intelligence Risk Certificate

though they cannot be guaranteed to reach an optimal solution.

By assuming a specific number of criteria, then we can give them some weights, and at each point, to compare one position with another overall, we should calculate the value for each position by the following formula

$$\begin{aligned} \text{value} = & C_1 * W_1 + C_2 * W_2 + C_3 * W_3 + C_4 * W_4 \\ & + C_5 * W_5 + C_6 * W_6 \end{aligned}$$

Choosing a move in a particular position now becomes a matter of examining the positions that would arise from all the analyzed possible sequences of n moves in that position.

If two moves provide the same value, we should choose randomly between them, hence yield the twin advantage of making the program less predictable and providing a more-varied basis for learning (the program).

Human judgment, provided by practical experience with the game, is involved here in identifying plausible criteria for assessing positions, which indeed might seem fairly straightforward for an expert player. But, in this case, how many relative weight should we provide to each criteria?

The question above is far more difficult to assess, but there is no need to rely on Human judgement when selecting the crucial weights that feed into our move-selection algorithm, for this is where reinforcement learning can play a role.

Artificial Intelligence Risk Certificate

By continuously play interactively, the weights in our current strategy are progressively refined (by simulated evolution) and we hope to ultimately achieve something close to an optimal set of weight, and thus the best possible strategy of this kind.

Depending on the game and suitable choice of criteria, this sort of machine learning technique can enable a program to learn to play better than its designer.

1.9 Limits of Classical Artificial Intelligence

Based mainly on the simple context of familiar rule-governed games.

In less well-disciplined contexts, however, classical AI techniques struggled to fulfil their apparent early promise. One serious difficulty was how to represent the characteristics and relationships of things in the world, and especially the rules that govern their behavior.

Even human experts find it very hard do elucidate the implicit background assumptions that guide their judgements, and spelling out our ordinary common-sense understanding of things proves to be extraordinary difficult.

A related issues was the so called “frame problem”, keeping track of which aspects of a situation change when some action is performed, but also which aspects stay the same.

Artificial Intelligence Risk Certificate

Codifying these things in detail added yet more fuel to the sort of combinatorial explosion that we have already mentioned, which quickly, resulted when attempts were made to apply general purpose search mechanisms to classical AI problems of any serious complexity.

This becomes even worse if the data used is ambiguous or uncertain, requiring yet more complex calculation if probabilities were to be considered.

Real-time data about physical things was inevitably error-prone, with computer vision systems relying on tailor mad algorithms, first to identify such things as edges, shapes, textures and colors, and then to synthesize these features into the representation of similar objects.

In response to such limitations the focus of much classical AI research moved during 1980s toward so called expert systems, which were designed to capture knowledge within specific domains, often using logic programming as general purpose reasoning mechanism.

These systems, despite being very successful, their functionalities could not easily be generalized beyond those specific domains without hitting the same old problems.

AI were notoriously brittle, failing in unexpected ways when applied beyond their familiar narrow boundaries, or to situations that had not been explicitly foreseen by their designers.

Artificial Intelligence Risk Certificate

With this, progress continued everything that could be feasibly assessed yielding impressive results, while the dream of Alan Turing's of general artificial intelligence, was likely to remain unfulfilled, at least for the foreseeable future.

2.0 Neural Networks

Alan Turing had discussed AI around 1941, but what is often considered the first published contribution to modern AI development appeared soon after in the Bulletin of Mathematical Biology of 1943 by Warren McCulloch and Walter Pitts.

This paper intended as a contribution to the theoretical neurophysiology , describing the activity of neurons in the brain, and proposed that the behavior of these neurons could be analyzed in terms of propositional Logic.

Ironically in view of the late contrast that would be drawn between logical and neural approaches the paper task of neural nets as equivalents as Turing's machines, and hence as limited by Turing Computability.

Donald Hebb's 1949, suggested a way in which associative learning could take place within networks of neurons, based on "neurophysiological postulate", in which states that One cell by proximity of another, or by firing another, will suffer such a change that will increase its efficiency.

Artificial Intelligence Risk Certificate

Frank Rosenblatt later took this further, proposing a specific computational model for how the brain learns and stores information in which, rather than treating memories as encoded representations of experiences, he instead favored the Hebbian approach.

Rejecting any symbolic or algorithmic approach, he accordingly formulated his new model, the Perceptron, in terms of probability theory rather than symbolic logic.

2.1 Artificial Neurons

Rosenblatt's Perceptron led to a standard model of an Artificial "neuron", which takes several numerical inputs (i_1, i_2, i_n) and outputs a single value V . Each input is multiplied by some weight (w_1, w_2, w_n), and these are added together with some constant bias term as w_0 to make the net input. This is later fed into an activation function to generate the output value, such as:

$$V = f(w_0 + w_1 i_1 + w_2 i_2 + \dots + w_n i_n)$$

The activation function is chosen to be non-linear, and often approximates to a step function that gets triggered when the net input is greater than some threshold, making the neuron's output an "All or nothing". This non-linearity is crucial if a network is to be able to learn non-linear behavior.

In recent years, artificial neurons of this general kind have been powerfully linked together in "deep"

Artificial Intelligence Risk Certificate

networks that have multiple additional layers between the input and the output layer.

From figure 1.6, the greens are the input layers, the blues are the hidden layers, and the red ones are the output layers.

Each neuron in the input layer is given a specific numeric level of activation, in such a way that the overall pattern of activations represents the input data.

For instance, each neuron's activation might store the color value of one of the pixels in an image. In this example, every neuron in the input layer is connected to every neuron in the first hidden layer (a-layer), so that the level of activation of the a-neurons will depend on their activation function f_a as applied to the net input that they receive from the input layer.

In the same way, every neuron on the second hidden layer (b-layer) will depend on their activation function f_b as applied to the net input that they receive from the a-layer, and so on.

Typically, the activation function will be consistent across the neurons, but the weights of the individual connections will vary, thus influencing the propagation of activity through the network from the input layer, through the hidden layers until the output layer.

It is the changing in weights that represents the learning of a network. In a successful trained network, the weights will have evolved in such a way that the

Artificial Intelligence Risk Certificate

activation of the output neurons does indeed provide the desired response to the relevant input.

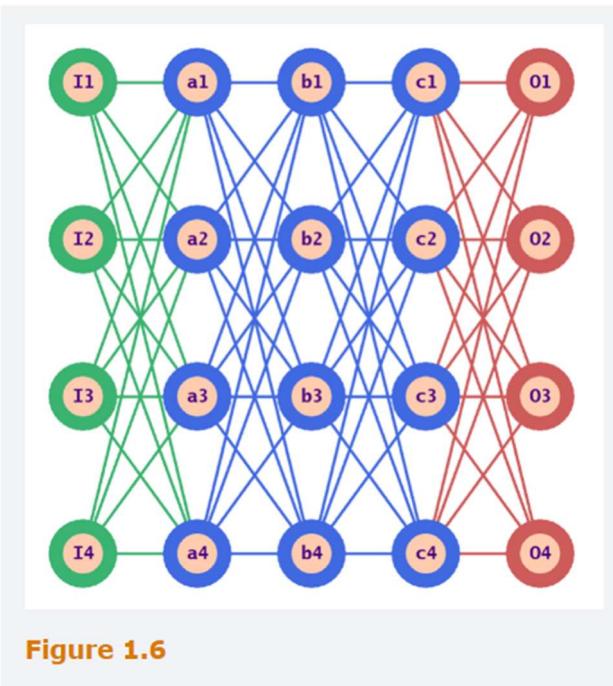


Figure 1.6

2.2 Connectionism and its Early Challenges

Rosenblatt's work provided widespread interest in artificial neural networks and an appreciation of some of their particular strengths as learning mechanisms.

In 1969, Marvin Minsky and Seymour Papert published a fierce critique in their book Perceptron's, which undermined the enthusiasm.

Perceptron's were incapable of learning some simple logical functions, thus apparently wrecking any prospect that they might provide a route toward "intelligent" information processing of any complexity.

The above is through for neural networks containing a single layer, and that more complex functions could be learned by arranging artificial neurons into multiple layers, hence denominated Deep Neural Networks.

Adding additional layers complicated the process of learning until the Back Propagation on the 1980s

The Output should match the input when the activation of the function is closest to 1 (maximal activation) and vice versa, when close to 0 minimal activation. However, if the actual pattern of activation is quite different then learning is required, so we should calculate how much a small change in the weights of the final output layer would contribute to improving the match.

Back propagation is then the process of working back through the layers, doing a similar calculation for all the other weights in the network.

Artificial Intelligence Risk Certificate

Once this has been done, the weights are now adjusted by a small amount, in the direction that would bring improvement to the model capability of forecasting.

This process should be repeated thousands or millions of times so that the network learns by iteratively adjusting its weights.

The Discovery and success of back propagation led to a resurgence of interest in what became known as “connectionism”, in 1986

This show that layers of simple interacting neurons, could achieve learning of complex, cognitively relevant functions.

There were 4 main reasons for the comeback being that:

1st is that this is biologically inspired and could provide insights on humans think.

2nd it seemed to learn in the same we do, through association and feedback in response to success or failure.

3rd this was really general and could cover a bunch of domains.

4th the storage of the learned information, rather than being explicitly represented, was obscurely distributed through the network of weights, which made the learning process more robust in response to “noisy” of ambiguous data, better able to generalize and less liable to break down entirely as the quality of data declines.

Artificial Intelligence Risk Certificate

In the 1990s the enthusiasm declined due to the impossibility of replicating in practical terms the theoretical concepts. Such concepts would come later as the computational capacity increased.

2.3 Deep Learning Proves its Potential

In 1998 Yann LeCun, publish work on a deep network with several “convolutional” layers, each of which has the effect of applying a local filter, Kernel, repeatedly to points across the grid, which in this case is a grid of pixel values in a greyscale digitized image.

The filter consists of a small matrix of numerical weights, and each of these weights is multiplied by the pixel activation value at its position, with the sum of these products providing the corresponding activation value in the convolutional layer (at the filters central point).

The weights at the beginning are set as random, and then learn in the same sort of way as other weights on the network. The virtue of such a convolutional layer is to enable efficient identification of local features in an image, which can then feed into the remainder of the network.

The model contained 8 layers, starting with the input layer (containing the pixels from the image), then two sequences of convolutional layers, followed by a “pooling” layer (which reduces the dimension of “feature map” by dividing it into 2X2 squares and

Artificial Intelligence Risk Certificate

averaging, then two further layers, and finally the output layer that would classify the image.

In 2012 AlexNet own the ImageNet Large Scale Visual recognition challenge with only 15% error margin. The Network was nearly 100 times bigger than the Yann LeCun one, holding around 900.000 neurons.

The computational power required had been obtained due to the high demand for high quality animated computer games, that required powerful graphic processing units, that turned out to be great for huge matrix operations that are required by neural networks.

2.4 Deep Learning Beats Symbolic AI at its Own Game

Another breakthrough arrived in 2013, when the London Company Deep Mind announce its success in programming a deep convolutional network to learn to play vintage Atari video games from the 1970s, which outperforms vastly better than an expert human.

The major breakthrough was that the system was not provided with any information regarding the game's goal, the information on the screen or the effects of user action (e.g. pressing buttons). It had to learn to do entirely on the basis of knowing that a certain number of distinct actions were available and trying these out in response to pixel information about the changing images and the game score. Everything else was done by deep

Artificial Intelligence Risk Certificate

reinforcement learning based on the score feedback, so in a sense, the system was teaching itself how to play from scratch.

In 2014 DeepMind was acquired by google and soon after created AlphaGo, a system capable of beating Go world champions, that is a game that until the moments was considered to subtle and complicated for computer algorithms to master in the foreseeable future.

Later, Alpha zero was able to teach it self other games, with residual user inputs, such as how to identify a win, how pieces can actually move, general rules of thumb and had efficient recursive adversarial tree searching built into it.

In contrast to Deep Blue and other traditional AI programs, AlphaZero learn how to play entirely on its own without any input from human experts or game databases. A few hours of self-training allow it to defeat the Stockfish champion program.

In this way, AlphaZero was altogether a impressive system, teaching itself new skills to an even higher level than human ingenuity had been able to achieve, and thus representing massive and potentially frightening progress towards artificial General intelligence.

2.5 The inscrutability of Deep Learning

Part of the promise and threat of deep learning lies precisely in its ability to represent all kinds of information in ways that it works out for itself in response to training data and feedback. This information is stored implicitly within the weights of a neural network that consists of layers of artificial neurons.

The input layer of the network is setup to reflect the specific problem case. The output layer is set up to signal the corresponding solution. But for a deep, multilayer system, the activation of each neuron depends on the input received from the neurons to which is connected on the immediate previous layer ((possible involving convolutions or other kinds of local processing)).

These inputs depend both on the level of activation of those previous neurons, but also on the weights given to the relevant connections. These weights are adjusted during the reinforcement learning process (by back propagation or Refinements thereof), which typically involves going iteratively through the training data, assessing the results outputs, and gradually refining the relevant weights until a sufficient match between inputs and outputs has been matched.

Neither the weights nor the roles of the neurons in the intermediate layers are predetermined when the learning process starts.

By the end, the immensely complex pattern of weights implicitly represents what the network has learned, but

Artificial Intelligence Risk Certificate

in a way that an unaided human will find impossible to interpret, and whose behavior they will be able to predict only by experience.

Humans should not be misled to suppose that the machine themselves “understand” what they are doing in any reflective way, and not only because they are completely non-sentient, and hence, have no awareness or conscious understanding of anything).

In conspicuous contrast to Classical AI techniques, their way of working is far more closely analogous to our own unconscious pattern-recognition than it is to how we think when explicitly calculating or reasoning about something.

A proper analogy is to a chess grandmaster, that without knowing empirically how to fundament his decision on a given position, will take it. The same with the system. The grandmaster also needs to be able to understand when a tactical combination is imminent. But as mentioned before, AlphaZero, can perform explicit calculation, exploring options down the “game tree” of branching possibilities.

Thus, within its carefully specified and rule-governed context, it was able to generate its own training data and feedback by playing a lot of different games against itself, learning by experience which patterns were most conducive to success.

Artificial Intelligence Risk Certificate

2.6 The Dawn of Artificial General Intelligence?

2022 gave us the more significant and widespread shock of AI, with Open Chatbots such as ChatGPT. These ones have been developed using techniques of deep learning, but this time with colossal human textual input in the form of around 300 billion words from various sources on the internet.

And unexpectedly, we are not facing a Technology capable of overcoming the Alan Turing's' tests in its full generality, to converse plausibly, flexibly, coherently and informatively about a vast range of topics, and without relying on pre- prepared outputs.

How to assess ChatGPT in theoretical Terms. The systems were developed by applying statistical analysis on those 300 billion words of textual data to create a large language model (LLM) that records the probability that any given sequence of words, will be continued in different ways.

Accordingly, ChatGPT's primary method of working is to predict which individual words are most likely to follow any particular textual context, and then to choose one of these words. It's not always the same word since an element of randomness allows to respond differently every time prompted). That word is then added to the text, and the process repeats itself.

Alongside the purely automated learning that generated the trillion or so parameters (i.e. stored probabilities) in the LLM (reflected in the deep's network weights, as

Artificial Intelligence Risk Certificate

explained earlier), human users improved the system responses through supervised fine tuning, in which it was given a wider range of typical prompts, together with a suitable human-crafted responses.

Then, a stage a stage of reinforcement learning from human feedback was applied, whereby the system generated a range of responses to each given prompt, and humans assessed the relative suitability of such responses (like or not).

This feedback was then statistically analyzed to generate a “reward model”, which could it turn be used to rate responses in general, and thus assist in selecting the most appropriate. This type of working method is based on imitation with variation/randomness, through the sort of response that it found in the massive textual resources that was used to train it.

The second point regarding this line of working, is that the information it stores implicitly within its trillion or so network weights has been tuned to reflect the characteristics of the textual data, rather than the characteristics of whatever domain the text might concern.

Despite not being able to play games, or any other of difficult game/action because has no mechanisms analysis, lookahead, it is a good representation of general intelligence, about a vast range of topics where it has not been augmented with specific assistance.

Artificial Intelligence Risk Certificate

Unless the language model in some domain can generate indirectly a reliable model of the reality (which looks more plausible in domains that are primarily conceptual or expressive), It will have nothing like such internal understanding, and can quite appropriately be described as a stochastic parrot.

3.0 Machine Learning and its Risks

AI rose to prominence, in the beginning as a relatively conventional branch of computer science (with well understood data structures and algorithmic methods), but then took a radically different trajectory with the dramatic, accelerating, and previously unexpected rise of contemporary machine learning over the last quarter-century.

Understanding the contrast, is key and beneficial to appreciate why AI is now seen as posing quite distinctive and novel risks in a way that previous computer systems did not.

3.1 Four Types of Machine Learning

1st is Reinforcement Learning, in which the machine through trial and error , positive and negative feedback learns how to assess the target variable.

2nd is Supervised Learning, in which the aim is to learn from a set of labelled example (Training data) either how

Artificial Intelligence Risk Certificate

to categorize further examples of the same kind of things, or to predict some characteristic. This is long part of the repertoire of classical computation, in the forms of methods such as linear regression, decision trees, and support vector classification.

The problem we are facing today, is that the kind of supervise learning, that is facing a lot of risks, is derived from Deep Learning, particularly because of its inscrutability and hidden basis.

This is ubiquitous in applications to recognize handwriting, faces, and road signs, to analysis of medical scans and visual scenes.

Learned predictions (often called regressions when the relationships involved are straightforward) operates differently, in that now the “labels” are typically numerical values rather than categories, and the aim of the learning process is to be able to predict the corresponding value for new instances (forecasting). (Credit scoring, weather predictions)

3rd Unsupervised Learning, where the aim is to identify patterns within the given data, but without the help of human input in the form of specified labels. The two most prominent forms of such learning focus on Cluster analysis and Dimensionality Reduction.

Cluster analysis is a learning technique in which groups of similar objects are automatically identified without needing a training set.

Artificial Intelligence Risk Certificate

Dimensionality reduction, on the other hand, is a technique in which the main “dimensions” of variation among the objects are identified, enabling their range and relationships to be more easily grasped.

This is yet again a form of learning that can be done using classical methods as well as using deep networks.

Principal Component analysis is a notable example of both Cluster analysis and Dimensionality reduction.

4th is Semi-Supervised Learning, which is a combination of supervised and semi-supervised learning. It is used in cases where only part of the available dataset is labeled. The unlabeled data is used to determine patterns within the explanatory variables, or it is fitted with “pseudo-labels” determined by estimating a model on the labeled part of the data.

3.2 Examples of Unsupervised Learning (Principal Component Analysis)

First example is regarding a physical example. Suppose that we had to plot the position of houses in a village, starting from coordinates in terms of latitude, longitude and height.

PCA will then transform those variables into a different framework, so the various dimensions, instead of lining up with latitude, longitude and height, will instead line

Artificial Intelligence Risk Certificate

up with whatever directions best discriminate between the data items.

Thus, if the village has grown up on a long fairly straight road on mainly flat land, then the dimension that gives most discriminating information, i.e. along which the houses are most spaced out, will be horizontal and in the direction parallel to the road, our first vector.

To provide a proper coordinate frame, the second vector must be orthogonal, i.e. at right-angle to the first, and to preserve most information, we choose another horizontal vector (because houses will vary more in Distance than in height).

Our third vector must be orthogonal to the others and hence vertical.

Other, easier example is by assessing a bunch of authors from century 18 and count how many times they use the most common 50 words, around 16 different texts.

This yields 50 fractional values for each of the 16 texts, representing the relative frequency of each word within it, i.e., if a word occurs 61 times on a 10.000 words text, then the text value is 0.0061. We then imagine these 16 texts placed within a 50-dimensional space, one equally for each word, in which the 50 coordinates for each text are proportional to these relative frequencies.

Thus, each of the 16 texts is represented by a point in space, in such a way that two texts whose points are close together have broadly similar patterns of word

Artificial Intelligence Risk Certificate

occurrence. This allows to see the whichever direction that provides more information, since this whichever direction shows the texts maximally spaced out in a two-dimensional projection.

What is striking about this method is that it clusters the authors in such a way as to identify patterns of similarities between them, even though it has been done entirely automatically.

It is really unclear to how to understand what similarities and differences are, since there is no simple explaining for Vector 1 and Vector 2 significance, since they do not represent simple words, but hence complicated linear functions of 50 different word frequencies.

To sum up, the PCA provides a powerful method of dimensionality reduction, whereby a large array of points in multidimensional space can be reduced to a graph in two dimensions, while effectively representing the most significant measures of closeness and distance within that multidimensional space.

Despite being hard to pin down the PCA vectors, does not reach the extreme level of inscrutability provided by deep learning networks.

3.3 Risk of Inscrutability

1st when classification or prediction is based on supervised learning with deep networks, it is likely that any bias that exists in the labelling of the training data will be implicitly learned by the model and perpetuated, but in such way that its presence is hidden and hard to eradicate.

For instance, considering a model that allows to select candidates, in a company where the majority of the employees are males, even by not asking directly the gender, the model will use other proxies and would equate into a biased result.

Another related risk is to privacy, because this deep implicit entanglement of complex information within the learned network can easily hold clues to personal characteristics that we would prefer to keep secret.

This also leads to risk of manipulation, in both political and commercial sphere.

All these problems arising from the inscrutability of deep networks models are rendered more intractable because their complex incomprehensibility also makes it very hard to identify, with a view to eradication, the source of any such hidden clues or biases.

Unlike a traditional expert system, the deep network model can provide no explanation of how it reaches its decisions and predictions.

Artificial Intelligence Risk Certificate

The true explanation is hidden behind the vast web of weights, possibly billions in number, which cannot possibly be rendered humanly comprehension.

Companies that employ deep learning systems without addressing these issues carry huge reputational risk when things go wrong.

3.4 Risk of Over Reliance

Inscrutability can also engender over-reliance, as we come to depend on a system whose mode of operation seems completely mysterious to us, and yet appears to be impressively expert at what it does. This can undermine our autonomy as responsible individuals and lead to serious dangers as we neglect to develop or apply our own judgement and fail to realize when the system is getting things wrong.

Neural networks have long been considered more robust than classical AI alternatives in the context of noisy, ambiguous, or incomplete data.

But to the contrary, it has recently become clear that deep learning models are commonly vulnerable to deliberately designed “adversarial” examples, whereby two images that differ imperceptibly may be categorized by the system as quite different, and even very confidently.

In such type of models, if the task/data involves any nuances, that are less common or entirely novel, then

Artificial Intelligence Risk Certificate

there is a serious risk that the user will end up with a code that sounds quite plausible, but that is incorrect, and it is hard to identify why.

The use of generative AI to generate code doesn't necessarily guarantee quality code or time savings and could expose a firm to financial risk if managed poorly.

3.5 Risk To individuals, Organizations and Society

Risk poses to individuals vary widely and may originate with organizations or institutions with whom they interact or through their own behavior. If AI systems contributing to decision making are poorly designed or trained, using inappropriate data, then one could suffer because of faulty, biased, or unfair decision making.

Due to the form of who individuals interact with technology, and the tendency of humans to over-rely on automated systems, can lead to complacency and reduced vigilance, which can potentially impact an individual in the form of reduced decision-making autonomy or a risk to safety and well-being.

AI that uses a lot of individual data can be used to manipulate. Risk poses to organizations originates due to over reliance in AI, false sense of confidence, bad decision making, costumer dissatisfaction, and commercial loss. Depending on the type of over reliance,

Artificial Intelligence Risk Certificate

the company may face reputational, regulatory and legal risks.

Advancements in AI create the potential for risks at the society level, mainly job losses, a widening wealth gap between those who have the skills to work alongside AI and the exacerbation of global inequality between countries well positioned to take advantage of AI.

AI Can also be used for actors of bad Ill, such as Deep fakes to mislead and misguide the audience.

In the last few years, the discussions around AI have created the fear of “existential” risk to humanity, specifically in relation about superintelligence and the possibility that AI can use autonomous weapons, making decisions against humanity interest and so on.

It is crucial to be noted, that there is currently no consensus regarding the nature or extent of existential risk posed by AI.

Overreliance on AI systems is also called automation bias.

Artificial Intelligence Risk Certificate

Questions and Answers Module 1 from GARP

1. What are the Four most Basic Forms of Machine Learning ?

Reinforcement Learning, Supervised Learning, Unsupervised Learning and Semi-Supervised Learning.

2. What are some reasons why the inscrutability of Deep Network Machine Learning is Problematic

The **presence of bias** can be hidden and hard to eradicate. The **Risk to Privacy**, because the deep implicit entanglement of complex information within the learned network can easily hold clues to Personal characteristics that we would prefer to keep secret. **Risks of Manipulation**, in both commercial and Political Sphere, because those who can identify our personal characteristics and foibles may also be able to play on them through the now familiar phenomenon of targeted Advertising.

3. What is the risk associated with Overreliance on AI Systems?

Overreliance on AI Systems, also called automation **Bias**, can lead to **complacency** and **reduced vigilance**, which can undermine our autonomy as responsible individuals. This can create the potential for a dynamic in which one neglects to develop or apply one's own judgment and fails to realize when the system is getting things wrong

Artificial Intelligence Risk Certificate

4. True or False: The Fact that AI-Specific regulations are still emerging results in organizations having little or not AI-Related regulatory risk

False – Although AI-specific regulations are still emerging, businesses practices that rely on AI are still Subject to existing privacy laws and model governance regulations, so in cases where such practices hurt individuals or groups, organizations are potentially subject to regulatory risk and legal risk.

5. In the Context of Reinforcement Learning, why might Lookahead be applied?

Lookahead becomes essential when our aim is to form a multistep plan to achieve some goal, within a context where there is a very large number of possible situations overall, but with a relatively constrained and predictable range of options in any particular situation.

6. Differentiate between Cluster Analysis and Dimension Reduction

Both are Unsupervised Learning Techniques. Cluster Analysis is a Learning technique in which groups of similar objects are automatically identified without needing a training set. Dimensionality reduction, on the other hand, is a technique in which the main “dimensions” of variation among the objects are identified, enabling their range and relationships (e.g. comparative closeness) to be more easily grasped.

Module 2 – Tools and Techniques

Learning Objectives

Differentiate between machine learning and classical econometrics.

Differentiate among unsupervised, supervised, semi-supervised and reinforcement learning Models.

Distinguish between different data types.

Describe how to encode categorical variables.

Describe how to clean data and the benefits of cleaning.

Describe data preparation techniques and their benefits.

Apply transformations to a set of data.

Discuss how principal Component analysis (PCA) is used to reduce the dimensionality of a data set.

Explain the difference between the training, validation and test data sub-samples, and how each is used.

1.0 Machine Learning

Machine learning is a set of tools for data analysis and modeling. Is an aspect of Artificial Intelligence. Covers a range of techniques in which the model is trained to recognize patterns in data. Includes prediction and classification. Gained large popularity on the last decade due to the advances on computer science.

Machine learning offers advantages over traditional econometrics methods in such areas as handling big data, handling non-linearity, reducing dimensionality and handling missing data.

Handling big data: ranging from clustering algorithms to neural networks, there are ML approaches that can handle very large amounts of data more effectively than traditional econometric methods. These techniques are highly useful for making use of available data that is growing exponentially in volume, as well as an increase in dimensionality with the rise in digitalization of the global economy.

Handling non-linearity: There are many non-linear relationships and patterns in data that traditional econometric techniques might miss, which machine learning tools, such as decision trees, random forests and neural networks can help identify and model.

Reducing dimensionality: Machine learning tools such as principal component analysis (PCA) and feature selection can be particularly helpful for prediction and classification tasks when the number of variables is large

Artificial Intelligence Risk Certificate

and when the number of variables is greater than the number of observations, a situation in which standard econometric methods tend to have great difficulty.

Handling missing data: There are ML techniques such as K nearest neighbors that can be used to handle missing values in large datasets in a flexible way.

Machine Learning bypasses the need for a theoretical reason and goes straight for data processing and is a set of techniques that can work the relationship by themselves, being optimized for predictions.

Machine learning is not on inference, but on the ability to produce out-of-sample predictions. Machine Learning tends to deviate from that of classical statistics.

1.1.1 ML, Classical Statistics and Econometrics

Econometrics is often letdown by bad theorizing at the beginning of the process.

Classical Statistics and Econometrics usually hypothesized that the data generating process can be approximated based on some economic or financial theory. The analyst decides on the model and the variables to include, and the computer algorithm's role is generally limited to estimating the parameters and testing whether they are significant. Based on the results, the analyst decides whether the data supports the pre-specified theory.

Artificial Intelligence Risk Certificate

In contrast, Machine learning treats the data generating process as unknown and uses techniques such as regularization to select the relevant predictors.

Model Selection is at the heart of the empirical design of ML applications, and tuning, the process of searching through many models to identify the top performers, is a common characteristic of all ML methods. In contrast, to traditional statistics the focus is not on inference, but on the ability to produce reliable predictions out-of-sample.

Therefore, tools such as measures of out-of-sample predictions accuracy and an understanding of the bias-variance trade off play more important roles than traditional statistics such as R-squared, t-values and p-values

Table 1.1 Differences in the terminology between machine learning and traditional statistics / econometrics.

Statistics	Machine Learning
Data point	Example, instance
Dependent variable, explained variable, predicted variable, regressand	Output, outcome, label, target, response variable, ground truth
Independent variable, explanatory variable, predictor, regressor	Feature, signal, input, attribute
Estimation	Training, learning
Estimator	Learner (classifier), algorithm

1.2 Four Types of Machine Learning

There are 4 types of machine learning methodologies, reinforcement learning, unsupervised, supervised and semi supervised.

Unsupervised Learning: Concerned with recognizing patterns in data. For each observation we have a vector of features, but no corresponding output value to predict. Involves clustering the data or finding a small number of factors that explain the data.

Is not used to generate predictions and at first glance might not appear to be very worthwhile. However, for instances, unsupervised learning could be used by a bank to assess a set of transactions, and check for anomalies that might be suspicious and worthy of further investigation. Dimensionality reduction and PCA are also components of unsupervised learning

Supervised learning: Is concerned with prediction and classification. For each observation in the data set, we have a vector of attributes and an associated output or label. The algorithm learns from the “labeled” data with the aim of producing accurate predictions of the target value for new, unseen, and unlabeled instances.

In the financial realm, supervised machine learning found early application, in algorithmic trading and high frequency trade execution. A successful example of classification is in credit decisions.

Artificial Intelligence Risk Certificate

Semi-supervised learning: the objective is to make predictions. But only part of the available data is labeled. The remaining is used to determine pattern within the explanatory variables, or if it is fitted with “Pseudo-labels” determined by estimating a model on the labeled part of the sample.

Reinforcement learning: Focus on making a series of decision to reach a goal. The learning environment might be static, or dynamic, with changes occurring while the process involves. There are no explicit labels. Feedback is provided in the form of reward during the learning process, which encourages a desired behavior, but without giving explicit instructions to the learner. Uses trial and error approach where the desired behavior is rewarded. This is quite useful when decisions need to be made repeatedly so that the algorithm can learn based on the rewards or sanctions received in previous rounds. The output from the reinforcement learning application is a recommended action given the circumstances rather than a prediction, classification or cluster. Is used for instance to find the optimal way to buy or sell large number of shares.

Parametric vs. Nonparametric: Machine learning can also be divided into two approaches. **Parametric methods** require the modeler to make an assumption about the functional form of the relationship between the features and the label. This map can be a linear function or a nonlinear, highly complex function. The parameters that describe this map are then estimated or learned using the available data. **Nonparametric**

Artificial Intelligence Risk Certificate

methods do not make any explicit assumption about the functional form of the map or relationship between the features and the output.

Parametric methods carry the risk that the chosen functional form is wrong, and therefore the resulting model will not fit the data well. On the contrary, **non-parametric methods** are very flexible and capture very complex data patterns. However, they require many observations to obtain an accurate estimation of the map.

1.3 Exploratory Data Analysis

Is a crucial step before building a machine learning model. Is the process of:

Collecting Data

Cleaning the Data

Visualizing the Cleaned Data

Analyzing the Final Data

Data is collected, and then transformed into a usable format and then treated by cleaning duplicates, removing or filling N/A. The Data is then analyzed and visualized to understand any relationship contained in it. Visualization and basic statistical analysis will be useful in selecting features for building a robust ML model.

1.3.1 Data Collection and Preparation

Data analyst often spend 80% of their time cleaning the data, as good data cleaning can make all the difference between successful and unsuccessful ML Project. Data analysis can only be as good as the underlying data, a concept often summarized by the acronym GIGO, which stands for Garbage in, Garbage Out.

Structured data are organized in rows and columns, where typically, each column represents one attribute, and the rows contain different r observations for the attributes. Census data or the database containing the characteristics of the borrowers of a bank, are examples of structured data.

Unstructured Data are not arranged according to a preset to a data model. Examples are sensor data, image data, web logs, network traffic, texts... Unstructured data are inherently more difficult to analyze, as they must be represented first in a format that is readable by a machine.

Semi Structured data refers to data that is partly structured and partly unstructured, such as geo-satellite images or textual.

Numerical and Categorical Data: Structured data sets may contain two types of attributes. Attributes such as age and income are numerical and have a natural ordering, being called continuous numeric or quantitative. Attributes like marital status are said to be categorical and can take discrete values. A special case

Artificial Intelligence Risk Certificate

of categorical variables is binary data. Categorical variables may have or not have a natural ordering.

Longitudinal variables and cross-sectional variables (data):

The characteristics of a pool of mortgages loans is an example of *cross-sectional data*, where each loan observation is independent of the other observations. On the other hand, *longitudinal data* are related to each other temporarily, spatially and through network connections (time series and so on). Observations cannot be understood independently from one another.

Textual and Other data: Natural Language processing (NLP) is a technique that is applied to the process of processing textual data.

Table 1.2 Classification of Different Data Types

Classification Basis	Data Type	Sub Type	Example
Primary data Types	Numerical data	Continuous data	Income, age, temperature
		Discrete data	Number of daily emails received Number of pages in books
	Categorical data	Nominal data: Data with no natural ordering implied.	Marital status, gender
		Ordinal data: Data with implied natural ordering	Educational level, credit ratings, Likert scale responses (strongly agree, agree, neutral, disagree, strongly disagree)

Artificial Intelligence Risk Certificate

Organization of data	Structured: Data that can be organized in rows and columns	Census data, mortgage loan origination, and performance data
	Unstructured: Unorganized data. Most alternative data belong to this category	Textual data: Prospectus, books, news releases, transcripts of interviews and earnings calls, etc. Audiovisual: Podcast recordings, voice messages, webinars recordings, maps, photographs, paintings
	Semi-structured: Data has some structure, but it is not a fixed format.	Email, CSV files, HTML files etc.
Longitudinal vs. Cross-sectional data	Longitudinal data: Contains data spanning several time periods, or connected spatially or through network relationships	Stock prices and other financial data series, Economic data series
	Cross-sectional data: Data is for a single point in time or time period. It provides a snapshot of a specific moment or time period.	2024 annual income of different individuals Product sales in May 2024 for different companies

Nominal Scales: No order, such as eye color.

Ordinal Scales: meaningful order but not having clear interval between variables. Customer satisfaction ratings.

Interval Data: meaningful order in which a consistent interval between values. Does not have a true zero point, implying that a value of zero does not represent absence of the quantity being measured(Temperature, calendars, longitude and latitude).

Artificial Intelligence Risk Certificate

Ratio Data: Implies a natural ordering with a clear interval between values and it has a true zero point. (height and distance traveled).

1.3.2 Data Cleaning

Is crucial due to the errors often originated on the collection process.

Inconsistent recording: For data to be read correctly, it is important that all data are recorded in the same way.

Unwanted observations: observations not relevant to the task at hand should be removed.

Duplicate Observation: These should be removed to avoid biased

Outliers: Are observations on a feature that are significantly different from the remaining data, such that suspicion arises that they were generated by a different underlying process. Should be checked carefully as they have big impact on the output. Not all predictive models are sensitive to outliers, such as tree-based classification models and support vector machines are deemed to be robust in the presence of outliers. If a model is sensitive to outliers, data scaling can often minimize the problem.

Missing Data: Most common problem encountered during the data-preparation stage. Can be structurally missing or just unavailable. Is crucial to understand why the values are missing, and if the pattern of missing data

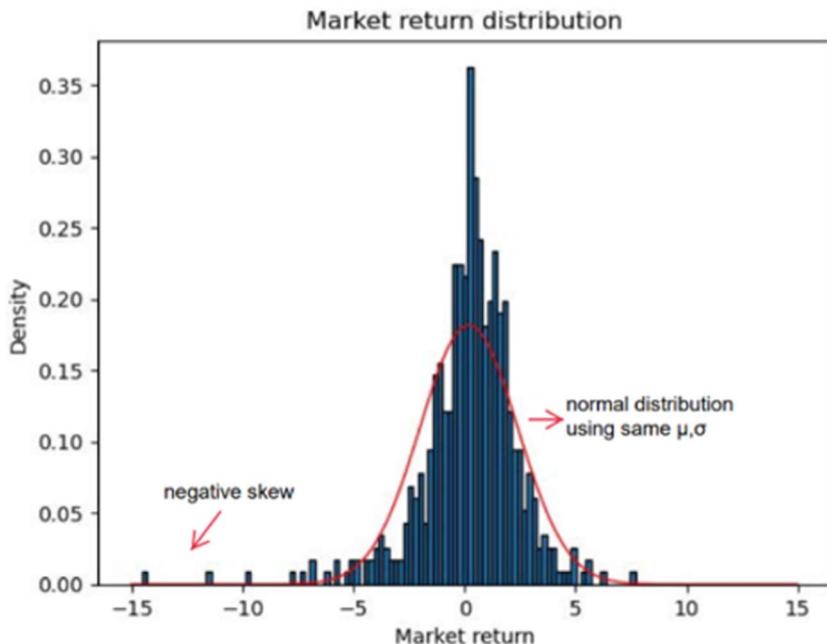
Artificial Intelligence Risk Certificate

is associated with the outcome. The latest is called informative missingness and can induce significant bias in the model. If missingness is not informative, the removal of a small number of observations with missing data from a large sample is not a problem. Otherwise, one approach is to replace missing observation on a feature with the mean or median of the observations of the same feature. This technique is called imputation.

1.3.3 Data Visualization

Great to gather patterns and identify potential problems, such as outliers. Are often used to achieve a reasonable understanding of the shape of the distribution of one or more variables. This allows the analyst to detect skewness levels on the data, needs for data transformation.

Figure Below shows a histogram of 521 weekly observations of US Market Returns collected between April 2010 and March 2020.



Artificial Intelligence Risk Certificate

A statistical summary is a useful tool to capture quickly the primary statistics of the data at hand. A summary of the returns data is presented in [Table 1.2](#).

Table 1.2 Statistical Summary of Weekly Market Returns (%) from 2010 to 2020

Item	Value
Mean	0.19
Median	0.33
St. Deviation	2.19
Skewness	-1.35
Minimum	-14.54
Maximum	7.68

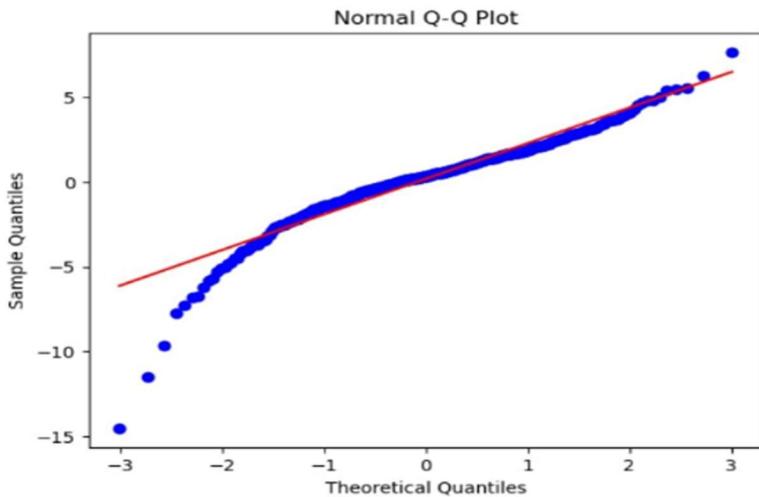


Figure 1.2 A normal Q-Q plot of US weekly market returns for a sample spanning from 2010 to 2020

Artificial Intelligence Risk Certificate

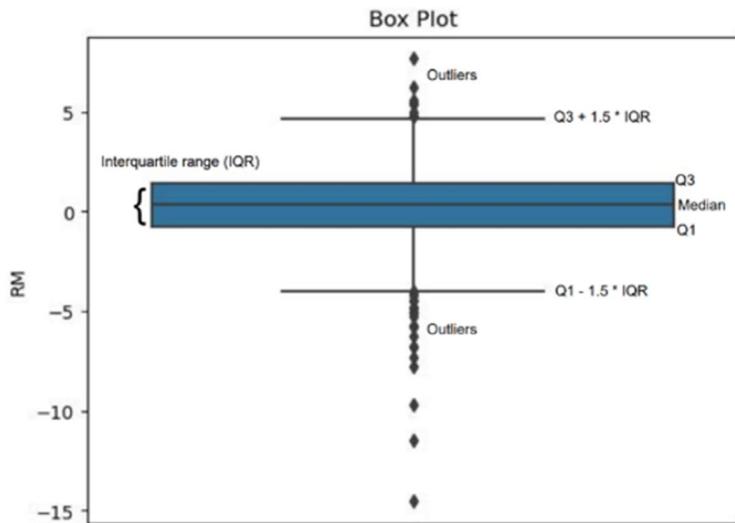


Figure 1.3 A box-and-whiskers plot of US weekly market returns for a sample spanning from 2010 to 2020.

Table 1.3 Correlation Matrix Showing the Relationships Between Daily Changes in Different Parts of the Treasury Yield Curve⁶

Treasury Maturity	1Y	2Y	3Y	5Y	7Y	10Y	20Y	30Y
1Y	1.00	0.85	0.80	0.72	0.65	0.58	0.45	0.40
2Y	0.85	1.00	0.96	0.90	0.83	0.76	0.61	0.54
3Y	0.80	0.96	1.00	0.96	0.91	0.84	0.70	0.63
5Y	0.72	0.90	0.96	1.00	0.97	0.93	0.81	0.74
7Y	0.65	0.83	0.91	0.97	1.00	0.98	0.89	0.83
10Y	0.58	0.76	0.84	0.93	0.98	1.00	0.95	0.90
20Y	0.45	0.61	0.70	0.81	0.89	0.95	1.00	0.95
30Y	0.40	0.54	0.63	0.74	0.83	0.90	0.95	1.00

1.3.4 Feature Extracting

Although quantitative data can be directly inputted into a model, qualitative data needs to be transformed in a way that is suitable for statistical analysis.

The process of transforming non-numeric information into numbers is sometimes termed **Encoding**.

Nominal data, attribute a dummy variable in format of 0,1,2,3 and so on. This could equate to problems since one move from 0-1 could be assumed by the model to have the same impact as a move from 1-2 which would probably not be the cases.

In cases such as marital status, (in the case of the marital status example, this could be 0 for “single,” 1 for “married,” 2 for “divorced,” 3 for “other categories,” etc.). Instead of having different categories and dummy’s, we should apply binarization, i.e. for each marital status a 0 or a 1, and the remaining marital status will be accordingly attributed a value between 0 and 1.

As a further example of a categorical variable, suppose we were developing a model to determine whether applications for credit cards should be accepted, and a piece of information we wish to include in the model relates to the applicant’s region of residence in the US. Suppose further that we have five categories: Pacific, Rocky Mountain, Midwest, Northeast, and South. It might be tempting to think that we could set up a single variable taking values such as Pacific = 0; Rocky Mountain = 1; Midwest = 2; Northeast = 3; and so on.

Artificial Intelligence Risk Certificate

However, because the information has no natural ordering, it would be inappropriate to code it as if it did.

Again, the correct approach is to set up a separate 0–1 dummy variable for each category. Then, for each individual applicant, the dummy variables corresponding to the four categories that do not apply would take the value 0, whereas the one that applies would take the value 1.

However, including dummy variables for all categories in regression-based models can lead to introduction of **multicollinearity, a phenomenon known as the dummy variable trap.**

This originates when two dummy variables are perfectly correlated, resulting in inaccurate calculations of regression coefficients and standard errors.

In order to solve this, it is necessary to impose constraints on the parameters of regression coefficients. Two commonly used constraints are **omitting one of the Dummy variables from the equation or setting constant term (bias) of the equation to zero.**

A slightly different situation is when there is a natural ordering for the categorical data (ordinal variable).

On some occasions, is interesting to convert numeric attributes into categorical ones. **Discretization** is the process of transferring continuous functions, models, variables and equations into discrete counterparts.

Artificial Intelligence Risk Certificate

Table 1.4. Examples of Different Data Conversions

Data Type Converted	Example	Mapped to
Categorical - Nominal	Marital Status	Single (0/1), Married (0/1), Divorced (0/1) etc.
Categorical - Ordinal	Education Level Credit Ratings	No High School Diploma (0), High School Diploma (1) College (2), Post Graduate (3), Doctorate (4), Professional (5) AAA (6), AA(5), A(4), BBB (3), BB (2), B (1), Unrated (0)
Numerical	Firm size (market capitalization)	0, Micro-Cap (<\$250MM) 1, Small (\$250MM -\$2B) 2, Medium (\$2B - \$10B) 3, Large (\$10B-\$200B) 4, Mega Cap (\$200B and above)

1.3.5 Data Scaling

Machine learning required that all variables are measured on the same scale. Otherwise, the techniques will not be able to determine the parameters appropriately and the results will be dominated by the feature with the **largest magnitude**.

Standardization involves subtracting the sample mean of each variable from all observations on that variable and dividing it by the standard deviation.

Mathematically, the j^{th} observation on the i^{th} variable, x_{ij} , would be changed to:

$$\bar{x}_{ij} = \frac{x_{ij} - \hat{\mu}_i}{\hat{\sigma}_i}$$

Where $\hat{\mu}_i$ and $\hat{\sigma}_i$ are the estimated Mean and Standard Deviation, respectively, of the Sample observations on variable i. This process creates a new variable that can take on any value but has a mean of Zero and a variance of one.

Normalization, sometimes called the **min-max transformation** takes a slightly different tack, creating a variable that is bounded between zero and one, but that will usually not have a mean of zero or unit variance.

$$\bar{x}_{ij} = \frac{x_{ij} - x_{i,min}}{x_{i,max} - x_{i,min}}$$

Where $x_{i,min}$ and $x_{i,max}$ are the minimum and maximum of the observations on variable i.

Artificial Intelligence Risk Certificate

Standardization is better when we have Outliers and do not want to exclude them

The three main reasonings for these processes are:

Numerical Stability of the learning algorithm. The difference in the scale of the variables could cause Overflow/underflow easily.

Ease of interpretation of the model parameter estimations. If the scales of the variables differ significantly, so do those parameters estimates and it would make evaluation of the importance of each estimate difficult.

Determining whether the out-of-sample prediction is within the range of the training data. If the out-of-sample data corresponds to a normalized value greater than 1 or smaller than 0, that means that the prediction is an extrapolation that may correspond to a large prediction error.

Artificial Intelligence Risk Certificate

1.3.6 Data Transformation

When data is highly skewed it is a good practice to transform it. As a rule of thumb, If the ratio of the highest value to the lowest value is larger than 10, we consider the variable to be highly skewed. Is common practice to use the Natural Log. Although this not usually produces a symmetric distribution, the date are better behaved. Also, can be applied the squared root of the inverse transformation.

Although scaling does not change the shape of the underlying distributions and correlations, transformation of data will result in changes to the distributions and correlations.

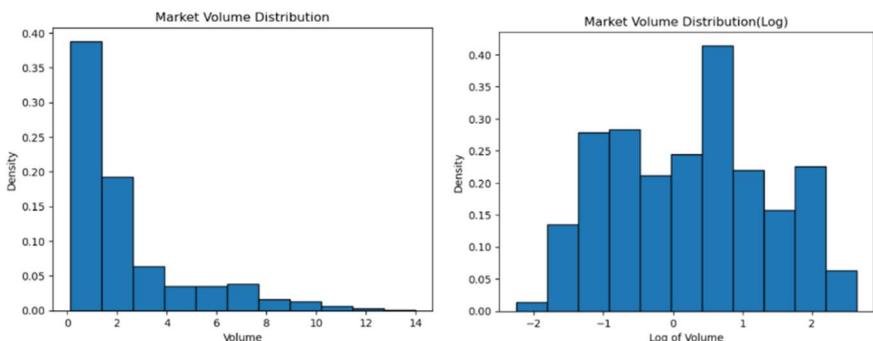


Figure 1.5 The distribution of the trading volume of the S&P500 from 1990 to 2010 (on the left) and of its log (on the right).

1.4 Dimensionality Reduction Techniques

Large datasets can often be represented more compactly which makes the application sophisticated and computationally intensive algorithms easier. Techniques such as PCA use the correlations in the data to represent it in a smaller number of dimensions. These techniques, which also belong to the realm of unsupervised learning, are often used to obtain a smaller number of uncorrelated features, from a larger number of correlated variables.

1.4.1 Principal Component Analysis

Is the most used dimensionality reduction technique. The idea behind this method is to find linear combinations of the original predictors that summarize most of the variability in the data. These linear combinations are called Principal Components. The first PC is the linear combination of the original predictors that captures the most variability in the data among all possible linear combinations.

The Second PC is the linear combinations of the predictors that capture the most variability that is not explained on the First PC. This one is constructed to be orthogonal to the first PC, that is, the two are uncorrelated. The process continues the same way to the Third and subsequent PCs, until the variability of the data has been explained, which will provide the same

Artificial Intelligence Risk Certificate

number of PCs as features. Components weights, i.e., the weight given to each PC helps to understand which features are the most important to explain the variability of the data.

The j^{th} principal Component can be represented as follows:

$$PC_j = b_{1j}P_1 + b_{2j}P_2 + \cdots + b_{mj}P_m$$

Where P_i denotes each of the original predictors (after standardization or transformation), m is the number of original predictors and b_{ij} is the weight assigned to P_i in PC_j , which captures the importance of predictor i for the Principal Component j .

An PCA example is on the yield Curves from US Treasuries with different maturities. From Daily changes ranging from November 2018 to October 2023, using PCA, an analyst can find a small number of uncorrelated variables that describe the Treasury yield Curve.

The Table below shows the Principal Component weights obtained from daily changes for the 8 different maturities. The daily data was scaled before running the principal component analysis. To explain the movements fully, all eight components are necessary. However, when the actual movements are expressed as a linear combination of the components, the first component explains more than 82% of the variation, and the first 3 components explain more than 98% of the variation. This is due to a high degree of correlation between the yield movements, and the bulk of

Artificial Intelligence Risk Certificate

information contained in them can be captured by a small number of PCs, which are then used as explanatory variables in subsequent regression models, rather than the Yields themselves.

Table 1.5 Principal component weights for 8 US Treasury yield series: one-year, two-year, three-year, five-year, seven-year, ten-year, twenty-year, and thirty-year

Treasury Maturity	Principal Component							
	1	2	3	4	5	6	7	8
1	0.29	-0.54	0.73	-0.29	0.06	0.02	0.02	0.00
2	0.35	-0.40	-0.15	0.60	-0.35	-0.46	-0.03	-0.03
3	0.37	-0.27	-0.30	0.19	0.15	0.76	-0.27	-0.04
5	0.38	-0.07	-0.33	-0.22	0.24	-0.07	0.67	0.43
7	0.38	0.08	-0.23	-0.37	0.14	-0.24	-0.04	-0.76
10	0.38	0.23	-0.07	-0.30	-0.04	-0.25	-0.64	0.49
20	0.34	0.42	0.20	-0.06	-0.71	0.30	0.25	-0.05
30	0.32	0.48	0.38	0.50	0.51	-0.05	0.05	-0.03

Looking at the Weights above, the first PC loads positively and almost equally on all interest rate variables, which would imply that all the eight yields tend to move on the same direction. So, the first PC can be interpreted as movements on the yield curve level. The second PC Loads negatively, on short term yields and positively on long term yields, implying that as the shorter-term yields rise or fall, the long-term yields tend to move the other way around. Interpreted as movements on the curve slope. The Third PC loads positively on one-year, negatively from two-to ten years

Artificial Intelligence Risk Certificate

and positively on long term yields. We can interpret as a twist in the yield curve.

PCA is extremely useful when the original predictors are highly correlated with each other. In fact, besides reducing the dimensionality of the predictor set, PCA helps solve problems associated with multicollinearity and improves numerical stability of models that require low correlation among the predictors. In fact, PC are built to be uncorrelated with each other and this appealing feature has contribute to their popularity.

It must be understood that PCA seeks to find linear combinations to explain the variability in the predictors without any understanding of the predictor's measurement scale or their distribution, i.e. if they are skewed. Data should be always scaled, due to PCA attributing more weight to the largest magnitude variables.

PCA is often used in high dimensional problems to choose a lower number of predictors before applying another method. The number of PCs to retain should be provided by the Researcher.

A heuristic way that is often used to support the decision is to construct a plot. The plot usually shows that the amount of explained variability decreases as the number of components increases (Scree Plot).

1.5 Training Validation and Testing

When a dataset is used for prediction, the analyst wants the model to be able to generalize well to the data that have not yet been used to estimate it.

For this purpose, in conventional econometrics, it is common, although not universal, to retain a part of a data sample for testing the fitted model and determining how well it can predict observations on the dependent variable that it has not seen. This leads to a distinction between in-sample and out of sample parts of the data.

The use of **out of sample (test)** is even more crucial in machine learning, as there is typically little economic or financial intuition behind the modelling assumption and the risk of choosing a complex model that accurately fits the dataset at hand but does not generalize well to unseen data is high.

The estimation of a model that is too complex and captures the noise in the dataset at hand rather than the true nature of the relationship between the features and the output is generally known as **overfitting**.

On the other hand, **underfitting** occurs when significant patterns in data are not captured by the model.

The **Validation set** is used to select between competing models. We are comparing alternative models to determine which one generalizes best to new data-.

Once this model selection has been undertaken, the validation set has already been “contaminated” and is no

Artificial Intelligence Risk Certificate

longer available for genuinely independent test of the model's performance.

The **test set is used** to determine the retained model effectiveness.

A good model will be able to generalize, which means that it will fit almost as well to the test sample as to the training sample because the machine has learned the crucial elements of the relationship without fitting to unimportant aspects (noise) that would likely repeat in the test set.

1.5.1 Sample Splitting and Preparation

One rule of thumb is that roughly two-thirds of the sample is used for training, and the remaining one third used equally split between validation and test.

If the training sample is too small, this can introduce **biases in the parameter estimation**, whereas if the validation sample is too small, **model evaluation can be inaccurate**, so that is hard to identify the best specification.

If the output data in the sample have no natural ordering (**i.e. they are cross-sectional**) then the three samples should be drawn randomly from the total dataset.

On the other hand, if the data are timeseries, it is common for the training data to be the first part of the sample, then the validation and finally the test. This

Artificial Intelligence Risk Certificate

sample split has the advantage of allowing the model to be tested on the most recent data.

Cross-validation involves combining the training and validation data into a single sample, with only the test data held back. Then the combined data are split into equally sized sub-samples, with the estimations being performed repeatedly and one of the sub samples left out each time.

The Technique known as **k-fold cross-validation**, splits the combined training and validation data available, n, into k samples, with the test data excluded from the combined sample.

1.6 Software for Machine Learning

Programming languages are probably among the most important tools in the data analyst' toolbox.

The best approach for someone who is interested in learning a new programming language is to select one that is flexible with a rich ecosystem of third-party open-source libraries. Such R studio and Python, which are both open-source scripting languages that run on Windows, macOS and Linux platforms and are commonly employed by data analysts for machine learning tasks. Both R and Python can perform virtually every data analyst task, have an easy-to-read syntax and are relatively easy to learn. On the one hand, R tends to be preferred by statisticians as it is great for data

Artificial Intelligence Risk Certificate

visualization and has a rich environment of statistical packages. On the other hand, Python is a general-purpose language, and it is better at non-statistical tasks such as web scraping and textual analysis.

The most important python toolboxes for a data analyst are NumPy, SciPy, Panas, Scikit-learn, TensorFlow and Keras.

NumPy is the building block for many other Python's toolboxes as it provides the framework to perform operations with multidimensional arrays and to support fundamental linear algebra functions.

SciPy is a collection of numerical algorithms that is used, for instance, to solve optimization problems, a crucial task in machine learning applications.

SciKit-Learn is a machine learning library built on NumPy and SciPy, and it offers tools to perform preprocessing tasks, model training, selection and testing several machine learning methods.

Pandas is a toolbox for data manipulation providing high performance functions for merging, aggregating, and reshaping the data as well as ways to deal with missing values.

A popular package for building and evaluating machine learning models in R is CARET(Classification and regression training). Contains tools for data splitting, preprocessing, feature selection, model tuning and variable importance estimation.

Questions and Answers Tools and Techniques – Module 2 GARP

- 1. For each of the following terms used in Classical Statistics, provide the equivalent term in Machine Learning parlance:**

Intercept = Bias

Slope = Weight

Explanatory Variable = feature

Dependent Variable = Output or Label

In-sample Period = Training Data

Out-of-sample Period = Test Data

1.2

A- What are the Main differences between Machine Learning and more Conventional Econometric Techniques

Under Conventional Econometrics approaches, the researcher selects a particular model or hypothesis and tests whether it is consistent with the available data. The Emphasis is on inference and the main tools are t-statistics, p-values and R-Squares.

Under Machine Learning approaches, the emphasis is on letting the data decide the features to include in the model, with very few assumptions or theory. Inference is less

Artificial Intelligence Risk Certificate

important while the focus is on the model's prediction or classification accuracy out-of-sample.

B- For what kinds of Problems would machine learning likely be more suitable than conventional econometric modeling?

Machine learning techniques have advantages when applied to problems where there is little theory regarding the nature of a relationship, or which features are relevant.

It is used when the number of data points and the number of features are large (big data or wide data, as opposed to tall data where the number of predictors is strictly smaller than the number of observations).

Machine Learning might also be preferable when the relationship between features are nonlinear.

1.3 What are the main differences between Supervised and Unsupervised Machine Learning methods?

In Unsupervised learning problems, for each observation we have a vector of features, but no corresponding output value to predict.

On the contrary, in Supervisory Learning problems, a set of labeled examples is provided. In other words, there are instances for which values of the predictors and the outcome variables are available.

The Goal is to predict the Outcome for new, unseen, unlabeled instances.

Artificial Intelligence Risk Certificate

1.4 An example of

A- A Classification Problem

In a classification problem, the label is of categorical nature. An example is to discriminate among credit applicants who will default and those who will not.

The predictors could be the characteristics of the borrowers while the output is a label indicating whether they have defaulted.

B- A Prediction Problem

A prediction problem concerns the prediction of a numerical value. An example is the prediction of the sale of the price of a house.

The predictors could be characteristics of the house while the target variable is the sale price.

1.5 What is an Outlier and How should be treated?

Outliers are observations on a feature that are significantly different from the remaining data, such as being several standard deviations from the mean, such that suspicious arise that they were generated by a different mechanism.

Their treatment depends on the problem at hand. Sometimes they are the fruit of errors in collecting or transcribing the data.

Other times they convey useful information about the data, such as in Fraud Detection.

Artificial Intelligence Risk Certificate

Remedies include using algorithms that are robust to the presence of outliers or data Transformation.

1.6 What are the benefits of using PCA?

Principal Component Analysis involves projecting a feature dataset onto a smaller number of components. For instance, if the dataset involves ten features, the first five components might be used, which would then reduce the number of input variables by half.

This is particularly useful in situations where the features are highly correlated or very numerous and estimating a model containing them could be challenging.

By construction, the Principal Components are uncorrelated. The technique is straightforward to implement, no matter how many features or data points there are, because the components are simply linear combinations of the features.

1.7 Standardization and Normalizations.

Applicant ID	Income (\$)
1	50,000
2	30,000
3	27,000
4	88,000
5	36,000
6	47,000
7	18,000

Artificial Intelligence Risk Certificate

Applicant ID	Income (\$)	Standardized Income	Normalized Income
1	50,000	0.33	0.46
2	30,000	-0.53	0.17
3	27,000	-0.66	0.13
4	88,000	1.98	1.00
5	36,000	-0.27	0.26
6	47,000	0.20	0.41
7	18,000	-1.05	0.00

A. Standardize each data point.

[View Answer](#)

The first step in standardization is to compute the sample mean and standard deviation of the variable. In this case, the sample mean is \$42,286, and the sample standard deviation is \$23,041. Then each data point is standardized by subtracting the sample mean and dividing by the sample standard deviation, that is:

$$\bar{x}_{ij} = \frac{x_{ij} - 42,286}{23,041}.$$

B. Normalize each data point.

[View Answer](#)

The first step in normalization is to identify the minimum and maximum values among the sample data, which are 18,000 and 88,000 respectively. Then each data point is normalized by subtracting the minimum and dividing by the difference between the maximum and the minimum:

$$\bar{x}_{ij} = \frac{x_{ij} - 18,000}{88,000 - 18,000}.$$

2.0 Unsupervised Learning

Unsupervised Learning is associated with a model's use of unlabeled data to develop insights or pattern recognition with no specific guidance or rules.

A common application of unsupervised learning is clustering analysis, also known as segmentation, which aims to separate data points into groups based on the closeness of their features.

Clustering places points that are similar into the same group and points that are dissimilar into different groups. There are many applications of clustering analysis including:

Organizing customer data into groups to determine the characteristics that separate their purchasing behaviors. Investigating whether subsets of banks accounts can be clustered into groups likely and unlikely to involve fraudulent transactions or money laundering, which is an example of anomaly detection.

Creating cluster of documents and newswires with similar content. Cluster is also helpful for identifying the structure of a set of features prior to conducting a classification or prediction task.

In other words, even where we have labelled data, we might choose to deliberately ignore the labels initially to focus first on better understanding the characteristics of the features.

Artificial Intelligence Risk Certificate

All Clustering applications are based on measurements of distance, as explained below, and consequently it is essential that the data are normalized or standardized before analysis.

Otherwise, if one of the features has a larger scale than the others, it will end up dominating the measures so that other features are rendered irrelevant.

There are different types of clustering. One set of techniques is hierarchical clustering, where we begin with either one cluster containing all points and successively separate them into sub-clusters, or alternatively we begin with each data point in a separate cluster and successively combine them together.

The former approach is known as divisive clustering, while the latter approach is called agglomerative clustering.

A dendrogram is an example of a hierarchical clustering technique. A second set of clustering techniques is partitional clustering which separates the dataset by locating observations based on their centroids, which are center points of the cluster.

The K-Means clustering algorithm is an example of such techniques. A final set of clustering techniques is based on density points in feature space, of which DBSCAN (density-based clustering non-parametric algorithm) and SNN (shared nearest neighbor) are examples of.

Artificial Intelligence Risk Certificate

Inertia is an example of a general approach to understanding the number of clusters, because the better the model fits, the closer the data points will be, collectively, to their respective centroids.

Calculate the Euclidian distance measure between a data point and the centroid to which it has been allocated. With this, Inertia or Within-Cluster sum of Squares (WCSS) is:

$$WCSS_k = \sum_{j=1}^{n_k} d_j^2$$

Where n_k is the number of datapoints within the K^{th} Cluster. The Total WCSS for all the clusters is given by:

$$WCSS = \sum_{k=1}^K WCSS_k$$

The lower the Inertia, the better each cluster fits the data, the more separation between clusters.

The Inertia measures within cluster distances.

The Silhouette Scores compare within cluster distances and the distance between cluster, and the ideal cluster is a set of clusters in which the points within the cluster are close to the centroid and the clusters are quite far apart from one another.

The silhouette scores compares the average distance of each observation from other points in its own cluster (known as cluster cohesion, denoted a_i , whit its average

Artificial Intelligence Risk Certificate

distance from points in the closest other cluster, denoted b_i .

$$s_i = \frac{b_i - a_i}{\max [b_i, a_i]}$$

For each observation s_i will lie within the range [-1,1].

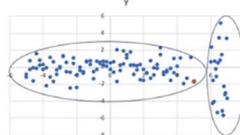
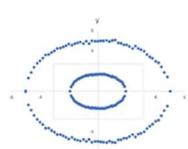
By averaging the silhouette scores for all the data points within a cluster, we can calculate one silhouette score, U_j , for each of the K clusters. It is then the common average of these scores:

$$s = \frac{1}{K} \sum_{j=1}^K U_j$$

A better value of K is the one that gives a higher silhouette score s, which implies that the points within a cluster are closest together but furthest away from other clusters.

Problems for K means are the data and the methods, which assumes that data points should be spherically aligned across a centroid, which is not always the case.

- Examples of cases where K-means would fail to find the clusters:



Artificial Intelligence Risk Certificate

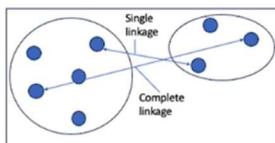
Other big issue is the presence of outliers which can cause distortion of how our data looks like.

K cluster handles big data quite well but not many features, being a.k.a. the curse of dimensionality.

Hierarchical clustering does not require a priori specification of the number of clusters as K means does. Instead, it utilizes every possible number from 1 to Number of features.

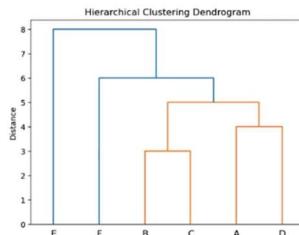
To measure effectiveness instead of distance measure uses a linkage criterion, such as single linkage which calculates the distance between clusters based on the distance between two data points from those clusters that are closest to one another, while complete linkage calculates the distance between data points in the clusters that are the furthest apart from one another.

- Single linkage: calculates the distance between clusters based on the distance between two data points from those clusters that are closest to one another.
- Complete linkage: calculates the distance between the points in the clusters that are furthest apart from one another.



Artificial Intelligence Risk Certificate

- The hierarchical clustering process can be represented with a dendrogram where the height of each line represents the distance between clusters.



2.1 K-Means Clustering Algorithm

Is a straightforward, unsupervised, algorithm to separate N observations into clusters. The number of required clusters, K, is determined at the outset by the analyst.

Often analyst, try several different Values of K and them aim to choose the most appropriate from among them.

The algorithm sometimes also known as Lloyd's algorithm proceeds as follows:

First, randomly, choose initial values for the centroids, $u_{ij} = 1, \dots, K$ which are the centers of the Cluster.

Secondly, allocate each data point, $X_i, i = 1, \dots, N$ to its nearest centroid.

Thirdly, recalculate the centroids to be at the centers of all the data points assigned to them.

Artificial Intelligence Risk Certificate

Fourthly, repeat steps 2 and 3 until the centroids no longer change.

Instead of centroids, the cluster can be determined with reference to medoids, which are the most frequently occurred points, if the features are categorical rather than taking continuous values.

Steps 2 and 3 require a definition of distance of each observation to the centroids. There are two commonly used measures. The first is the Euclidean ("as the crow flies") distance, and the second is the Manhattan distance measure.

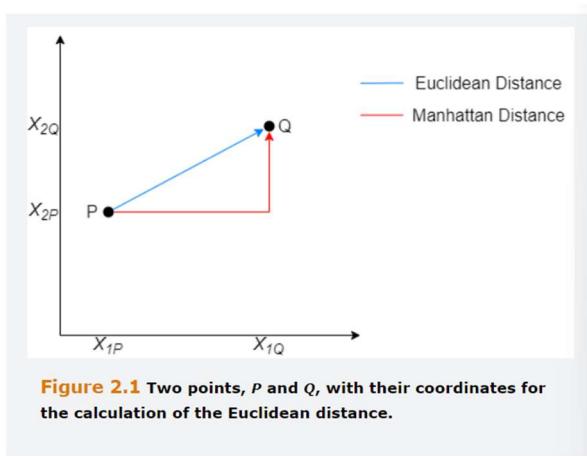
For illustration example, supposed that we have two features, X₁ and X₂, and two observations on each of them, represented by the points, P and Q, which have coordinates, (X_{1P}, X_{2P}) and (X_{1Q}, X_{2Q}). The Euclidean distance, d_E , between the two points would be calculated as the square root of the sum of the squares distance in each dimensions.

$$d_E = \sqrt{(x_{1Q} - x_{1P})^2 + (x_{2Q} - x_{2P})^2}$$

The measurement would be constructed in the same fashion if there were more than two dimensions. If there were m features for two points P and Q, the distance would be the square root of the sum of the square's distances

$$d_E = \sqrt{\sum_{j=1}^m (x_{jQ} - x_{jP})^2}$$

Artificial Intelligence Risk Certificate



Note that for simplicity the formula above describes the distance between one point P and another point Q. However, the purpose of K-means is not to minimize the distance between points, but rather to minimize the distance between each point and its centroid.

To see how K-means works in practice, consider ten fictional data points reported in [Table 2.1](#). Assume that $K = 2$.

Table 2.1 Ten fictional data points with two features, x_1 and x_2 .

i	x_1	x_2
1	6.8	12.6
2	3.3	11.6
3	3.8	9.6
4	4.4	7.8
5	7.6	17.4
6	6.5	19.9
7	4.5	2.7
8	8.4	6.9
9	6.6	7.8
10	4.5	7.3

Artificial Intelligence Risk Certificate

Although the K-means algorithm should be applied on data that underwent a rescaling of some from, for the sake of this example we skip that step, given that the two features are measured on a similar scale.

The Euclidean Distance between i=1 and the first centroid (i=2)

$$d_E = \sqrt{(6.8 - 3.3)^2 + (12.6 - 11.6)^2} = 3.64$$

The Euclidean Distance between i=1 and the second centroid (i=9)

$$d_E = \sqrt{(6.8 - 6.6)^2 + (12.6 - 7.8)^2} = 4.80$$

Table 2.2 reports the Euclidean distance between each datapoint and each of the two centroids. Each point is assigned to the closest centroid.

Artificial Intelligence Risk Certificate

Table 2.2 Ten fictional data points with two features, x_1 and x_2 and their Euclidean distance from random centroids

i	x_1	x_2	Distance to k_1	Distance to k_2	j
1	6.8	12.6	3.6	4.8	1
2	3.3	11.6	0.0	5.0	1
3	3.8	9.6	2.1	3.3	1
4	4.4	7.8	4.0	2.2	2
5	7.6	17.4	7.2	9.7	1
6	6.5	19.9	8.9	12.1	1
7	4.5	2.7	9.0	5.5	2
8	8.4	6.9	6.9	2.0	2
9	6.6	7.8	5.0	0.0	2
10	4.5	7.3	4.5	2.2	2

Obviously, the points from where we started are no longer centroids of these two newly formed clusters. Therefore, we should now compute the new centroids based on our assignment in the first step. The centroids are computed as the averages of the features of the points allocated to that cluster

For the first Cluster the average of X1 is:

$$\mu_1 = \frac{6.8 + 3.3 + 3.8 + 7.6 + 6.5}{5} = 5.6$$

For the first Cluster the average of X2 is:

$$\mu_2 = \frac{12.6 + 11.6 + 9.6 + 17.4 + 19.9}{5} = 14.2$$

The new centroid of cluster 1 is [5.6,14.2]. For cluster two would be [5.7,6.5]. We can now compute the

Artificial Intelligence Risk Certificate

distances of each data point to the new centroids and assign each observation to the cluster with the closest centroid. The algorithm stops when the clusters no longer change. In this case, in the following iteration in the point [3.8, 9.6] mover from cluster one to cluster two. After this iteration the cluster no longer changes, so the algo stops.

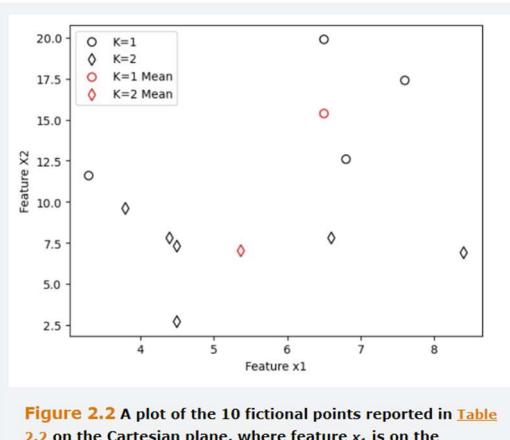


Figure 2.2 A plot of the 10 fictional points reported in Table 2.2 on the Cartesian plane, where feature x_1 is on the horizontal axis and feature x_2 is on the vertical axis.

2.1.1 Performance Measurement for K-Means

The first option is to rely on a measurement called inertia, i.e. WCSS, in which the better the model fit's the closer the data points will be, collectively, to their respective centroid.

The Between Cluster Sum of Squares (BCSS) which is the sum of squares of the distances of each centroid to the centroid of all data points weighted by the number of data points in each cluster. This measure increases as the number of Clusters grow.

A slightly more sophisticated approach to performance measurement is the to use the Variance Ratio Criterion (VRC), which is a.k.a the Calinski-Harabasz index:

$$VRC = \frac{BCSS(N - K)}{WCSS(K - 1)}$$

Where N is the total number of data points. A higher VRC implies a model with better grouping of clusters.

2.1.2 Selecting the Starting Positions of the Centroids

Choose Randomly. If we choose a different set of initial allocations, it is likely that we get a different set of solutions, particularly for large values of K.

To mitigate the impact of the initial selection, is common practice to run the K-means several times with different

Artificial Intelligence Risk Certificate

random initialization, and then select the one with lowest inertia.

A further alternative is to use K-means ++ which involves establishing the initial position of the centroids far from each other in the feature space, so that the position of only one of the centroids is chosen randomly. This can lead to better outcomes and faster convergence to the optimal solution.

Actually, randomly selected Centroids could mean that points are close to each other and make it harder to distinct between Cluster.

2.1.3 Selection of K

In the same way that R squared will never fall when more explanatory variables are added to a regression model, the inertia will never rise as the number of centroids increase.

In the limit, as K tends to N, each data point will have its own cluster and the inertia, WCSS, will fall to zero.

A straightforward rule of thumb is to set K equal to the integer closest to the square root of half the number of data points,

$$K = \sqrt{N/2}$$

However, this is rather arbitrary and will not vary depending on the characteristics of the data set.

Artificial Intelligence Risk Certificate

There are two better alternatives such as Scree Plots and the Silhouette method.

On Silhouette score, which ranges from [-1;1], if the score is equal to 1 would imply that all the points allocated to each cluster were exactly on the centroid of their respective cluster, and $s=0$ implies that the cluster are significantly overlapping. $S<0$ is empirically unlikely but would correspond to points being mis-clustered.

2.1.4 Selection of K Example

We now apply the K -means clustering algorithm to annual value-weighted stock index returns (all stocks on the NYSE, Amex, and NASDAQ) and Treasury bill yields³ from 1927 to 2021.

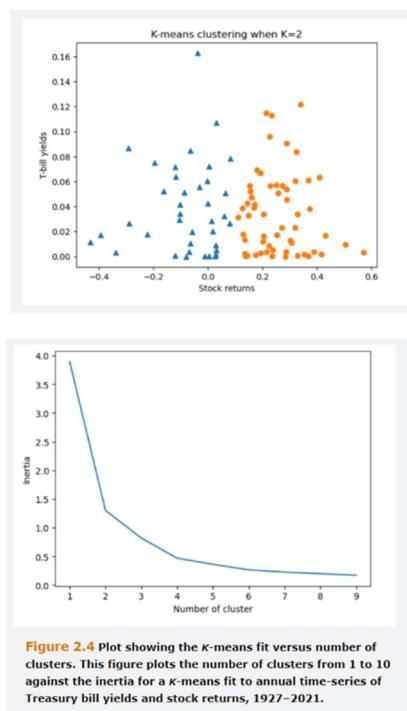


Figure 2.4 Plot showing the k -means fit versus number of clusters. This figure plots the number of clusters from 1 to 10 against the inertia for a k -means fit to annual time-series of Treasury bill yields and stock returns, 1927–2021.

2.1.5 Advantages of K Means

Is widely applied in finance and many other areas. Compared with other clustering techniques, its advantages include:

Firstly, makes Intuitive sense and is easy to visualize if there are up to three features.

Secondly, is quite fast and scales well to large number of data points

Thirdly, will always converge to a solution, even though it might not always be the most appropriate solution

2.1.6 Problems with K Means

Non-spherical Clusters: Since it is based on distance from the centroid, tends to form spherical clusters. 1st problem is if a cluster is within other cluster, because the algorithm will be unable to differentiate, the 2nd problem is that sometimes points within a cluster are closer to the other centroid than from the one from its own cluster.

A solution for this is to either apply a non-linear transformation to the input data, known as Kernel Function, or to switch to a different technique that can cope better with a wider variety of data patterns.

Presence of Outliers: Outliers might generate their own cluster or create distortion on the cluster centroids, especially if K is a high value. Standardizing the data

Artificial Intelligence Risk Certificate

using min-max will mitigate this issue to some extent, although from this perspective it may be preferable to remove outlying data points entirely from the sample prior to beginning the analysis.

Curse of Dimensionality: Tends to perform poorly as the number of features increases. In order to solve this problem, some features can be removed, the features can be transformed to a smaller subset using Principal Component analysis or similar, employ the cosine distance, which is a measure that does not always increase with respect to the number of features that can be employed instead of the Euclidean Distance.

2.1.7 Fuzzy K Means

All the above involves hard clustering, where data points are allocated uniquely to each centroid, meaning that a particular data point is either in cluster j or it is not. Each point is allocated to one and only one cluster, which is known as Winner takes all clustering regime.

An alternative is soft clustering, also known as Fuzzy clustering or Fuzzy C means, in which data point can belong to one or more clusters, with a probability assigned to each cluster.

By applying this is necessary to change the inertia equation, because of calculating the direct summation of d_j^2 , the equation incorporates a probability raised to a power f, in which f is the fuzziness Coefficient. The large

F the greater the extent to which the clusters will overlap., i.e., they will be fuzzier.

2.2 Hierarchical Clustering

The limitation of K-means clustering is that necessitates the number of clusters being specified a priori.

Hierarchical clustering does not require this, instead utilizes every possible value of K (from 1 to N) as an Integral part of the process.

Divisive: This begins with just one cluster and them sequentially partitions the data by adding another cluster until every data point has its own cluster.

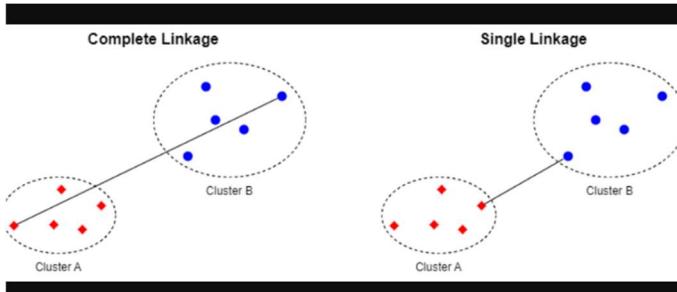
Agglomerative: This begins with each point having its own cluster, and then the Closest two cluster in feature space are merged. The process ends when all points have been merged into a single cluster.

In these two processes, we will need to capture the distance between one cluster and a single point on that cluster.

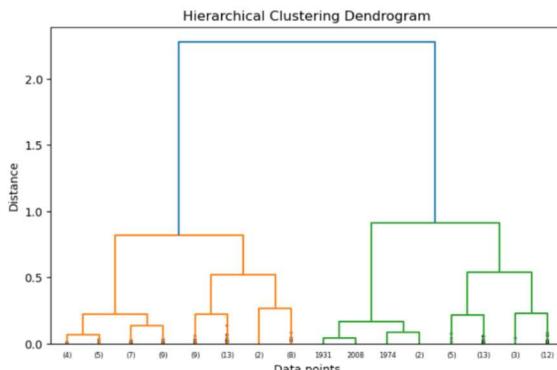
Instead of a distance measurement, a linkage criterion is used, of which there are several variants. The most straightforward of these is a single linkage, which simply calculates the distance between clusters based on the distance between two data points from those clusters that are closest to one another.

Artificial Intelligence Risk Certificate

Another possibility is to use a complete linkage, which calculates the distance between the points in the cluster that are furthest apart from one another.



When the process of division or agglomeration is completed, the output is presented in a dendrogram, which has the data points on the X axis and distances on the Y axis.



The example above is related to the stocks and bond example from the K-means topic.

Artificial Intelligence Risk Certificate

If we specify a cutoff so that only three levels from the original split are shown to avoid proliferation of tiny clusters culminating in data points at the bottom.

The value in parentheses shows the total number of points that are within a cluster, whereas the value without parentheses corresponds to the individual data for a specific year.

The distance is calculated using a slightly more sophisticated method known as Ward linkage, which minimizes the variance of the clusters being grouped.

To select the optimal number of clusters, we examine the heights of each vertical line before a split occurs, which shows how much the distance drops as an additional cluster is introduced.

In this case, if splitting the data into two clusters provides a large drop in distance relative to others, it would suggest that a division of cluster is preferred. But if the line is small, would suggest the incremental benefit in terms of reduced distances is not worthwhile. Like a Scree Plot, a dendrogram will not be able to show a uniquely correct optimal number of clusters, rather the optimal number will be a matter of interpretation.

2.3 Density Based Clustering

Another approach to clustering is the family of techniques including DBSCAN (Density based Spatial Clustering of applications with noise) and SNN (Share Nearest neighbors) that are based on density points. Here, a region in feature space constitutes a cluster if the density of points exceeds a pre specified threshold. DBSCAN distinguishes between three types of data points:

First, an Observation is a core point if at least a pre-specified number of other points are within a threshold distance of it.

Second, an observation is a border point if it is within the threshold distance from a core point, but it has fewer than the pre-specified number of other points close to it.

Third, an observation is considered a noise point if it is neither a core nor a border point.

Each core point constitutes a cluster, and the border points are allocated to their nearest core point cluster. Noise points are ignored and remain un-clustered. The threshold and number of points required to assign an observation as core are two hyperparameters to be tuned.

Although more complex – in terms of both intuition and the optimization method – density-based clustering has two important advantages. First, it can handle non

Artificial Intelligence Risk Certificate

spherical distributions of points in feature space. Second, it is considerably more robust with respect to outliers, and indeed observations identified as noise points are automatically excluded from the clusters.

Extra 2.A Different Distance Measurement

Manhattan distance, also known as L-Norm, between Point P and Q

$$d_M = |x_{1Q} - x_{1P}| + |x_{2Q} - x_{2P}|$$

Extending to m dimensions:

$$d_M = \sum_{j=1}^m |x_{1Q} - x_{1P}|$$

The **Euclidean Distance** is the Direct Route, whereas the Manhattan measure gives an approximation to the distance between two buildings that might be required when driving a car.

Minkowski Distance consists of nesting both the Euclidean and Manhattan distances in a broader framework.

$$d = \left[\sum_{j=1}^m |x_{1Q} - x_{1P}|^L \right]^{1/L}$$

If L=1 the Manhattan measure is calculated, if L=2, the Euclidean measure and if finally, L=∞, it is calculated the

Artificial Intelligence Risk Certificate

Chebyshev Distance, this is the maximum of the absolute distances along each dimension.

Cosine Similarity, is the cosine of angle between two vectors, P and Q

$$S_C(P, Q) = \cos \theta = P \cdot \frac{Q}{||P|| ||Q||}$$

Expressed in Scalar form

$$S_C(P, Q) = \frac{\sum_{j=1}^m x_{jQ} x_{jP}}{\sqrt{\sum_{j=1}^m x_{jQ}^2 \sum_{j=1}^m x_{jP}^2}}$$

Which is bounded to lie in the [-1;1] interval with all notations as above. In machine learning this measure is used to analyze the similarities between documents.

The **Cosine Distance** is defined as the complement of S_C

$$d = 1 - \frac{\sum_{j=1}^m x_{jQ} x_{jP}}{\sqrt{\sum_{j=1}^m x_{jQ}^2 \sum_{j=1}^m x_{jP}^2}}$$

Which is bounded to lie in the [0,2] interval.

Mahalanobis Distance is the distance between a data point and a distribution. Is the multivariate equivalent of the Euclidean Distance. It can be also viewed as a multivariate version of the Z-Score.

$$D^2 = (x - \mu)^T \Sigma^{-1} (x - \mu)$$

Artificial Intelligence Risk Certificate

Where x is the vector of observations, μ is the vector of means, and Σ is the variance covariance matrix. The distance measure is used for cluster analysis and classification as well for detecting outliers.

Extra 2.B Silhouette Method

Figure 2B.1 produces silhouette plots for $K = 2$ up to 9 starting from the top left to the bottom right plot. In each of these plots, the horizontal axis represents the silhouette score and each of the colored areas represent the silhouette scores within a cluster. For example, in the case of two clusters, cluster 1 and cluster 2's constituent scores are plotted in the top left chart. The vertical dashed lines are the average silhouette scores for the value of K , which are 0.57 for two clusters, 0.47 for three clusters, and 0.48 for four clusters, with lower numbers for the remainder. Ideally, the sizes of each silhouette would be the same for a given value of K , which is not the case here for $K \geq 3$. We would also like to see every silhouette having a right-hand tip bigger than the average score line, which holds for all values of K except 9. When K becomes large, the thicknesses of the silhouettes become increasingly variable, with some thin lines, indicating small clusters, and some negative values, indicating incorrectly clustered points. Overall, we would conclude that $K = 2$ or 3 would be ideal here.

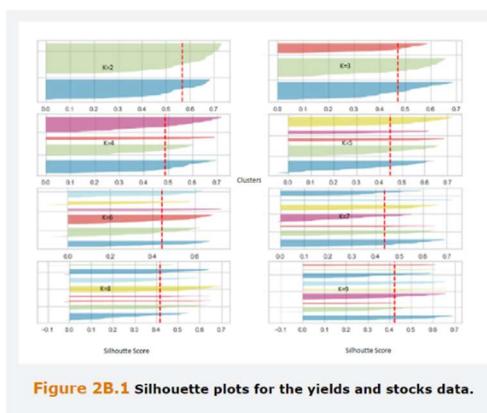


Figure 2B.1 Silhouette plots for the yields and stocks data.

Artificial Intelligence Risk Certificate

Extra 2.C K-Means Clustering Example

A retail bank is interested in examining the Characteristics of its customers who take a loan to buy a car. It identifies two features, client age and loan amount in US Dollars, for six costumers. Data on the Table below is raw and Standardized.

Use this data to form two clusters using the K-means algorithm, integrating twice with initial centroid positions, assuming they were randomly selected, of (0.5,0.5) and (1,1) in standardized feature space.

Data point number	Age	Loan amount	Standardized age (SA)	Standardized loan amount (SLA)
1	23	15000	-0.96	-0.75
2	19	8000	-1.19	-1.48
3	46	22000	0.38	-0.02
4	55	25000	0.90	0.30
5	61	35000	1.25	1.34
6	33	28000	-0.38	0.61

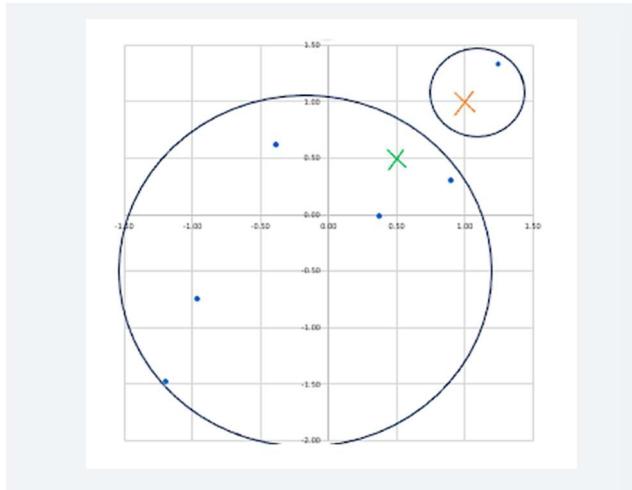
The first stage is to work out the distances from each point individually to each of the centroids. Then a point is allocated to the Centroid to which it is closest. We work with the Standardized features and assume Euclidean Distance measure is used.

The calculations are straightforward (Just use the Euclidean distance for each data Point).

Artificial Intelligence Risk Certificate

Once calculated the distances, attribute this point to the respective Cluster.

The initial centroid positions (denoted by crosses) and clusters formed are shown in the following figure:



Having Completed the allocations to first stage Cluster, we then reposition the Centroids at the Centers of those clusters by computing the average of each co-ordinate over the data points allocated to that cluster.

Because there is only one data point allocated to that cluster (1,1)the position of the centroid will shift to the coordinates of that data point, namely (1.25,1.34).

We take an average of each of the features across the other five data points to get (-0.25,-0.27)

Next, we repeat the process by determining which of the two new centroids each point is closest to.

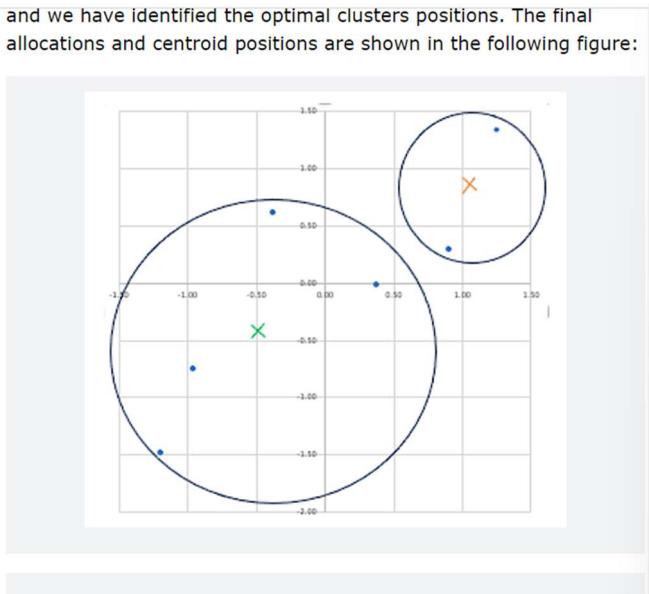
Artificial Intelligence Risk Certificate

Point 4 now switches to the other Cluster, but the other four points are still allocated to the same cluster as in the previous stage.

We then recalculate the positions of the centroids, which will differ due the reallocation of Point 4.

It turns out that when we calculate the distance of the points to each of these new centroids, none of them switch clusters and therefore that is the end of the process, and we have identified the optimal clusters position.

and we have identified the optimal clusters positions. The final allocations and centroid positions are shown in the following figure:



Artificial Intelligence Risk Certificate

Table 2.C.1 First stage distance calculations and allocations to clusters

Data point number	Age	Loan amount	Standardized age (SA)	Standardized loan amount (SLA)	Distance to (0.5,0.5) centroid
1	23	15000	-0.96	-0.75	1.92
2	19	8000	-1.19	-1.48	2.6
3	46	22000	0.38	-0.02	0.53
4	55	25000	0.9	0.3	0.45
5	61	35000	1.25	1.34	1.13
6	33	28000	-0.38	0.61	0.88

Table 2.C.2 Second stage distance calculations and allocations to clusters

Data point number	Age	Loan amount	Standardized age (SA)	Standardized loan amount (SLA)	Distance to (-0.25, -0.27) centroid
1	23	15000	-0.96	-0.75	0.86
2	19	8000	-1.19	-1.48	1.53
3	46	22000	0.38	-0.02	0.68
4	55	25000	0.9	0.3	1.28
5	61	35000	1.25	1.34	2.2
6	33	28000	-0.38	0.61	0.89

Table 2.C.3 Third stage distance calculations and allocations to clusters

Data point number	Age	Loan amount	Standardized age (SA)	Standardized loan amount (SLA)	Distance to (-0.54, -0.41) centroid
1	23	15000	-0.96	-0.75	0.54
2	19	8000	-1.19	-1.48	1.25
3	46	22000	0.38	-0.02	1.00
4	55	25000	0.9	0.3	1.60
5	61	35000	1.25	1.34	2.50
6	33	28000	-0.38	0.61	0.89

Questions and Answers Module 2 Chapter 2 from GARP

2.1 How would you choose the number of Clusters when using unsupervised Learning?

The more clusters are used when fitting an unsupervised Learning model, the better the fit of the algorithm to the data, but as the number of clusters increases, the usefulness of the models starts to diminish.

Determining the most appropriate number of clusters for a particular dataset could involve constructing a “scree plot”, which charts the Inertia (Sum squared distances of each point to its centroid), against the number of clusters. We would then search for the number of clusters beyond which the inertia only declines very slowly. Silhouette Scores, which compare the distance of each point (a) to points in its own cluster and (b) to points in the Closest other Cluster, can also be used.

2.2

A- Explain the steps in using the K-means clustering algorithm

First, specify the number of centroids, K and choose a distance Measure, such as the Euclidean or Manhattan Distance.

Artificial Intelligence Risk Certificate

Secondly, scale the features using either standardization or normalization.

Thirdly, select K points at random from the Training data to be the centroids.

Fourthly, allocate, each data point to its nearest centroid.

Fifthly, Given the points allocated to each centroid, recalculate the appropriate location of the Centroids.

Sixthly, if the positions of the centroids have changed from those in the previous iteration, then repeat step 4. If the positions of the centroids have not change (and the clusters are not changed) then stop.

B- In practice the K-means clustering algorithm is often carried out with several different initial values for the centroids. How would you choose between clusters that result from different initial choices of centroids?

You could select the centroids where the total inertia was the lowest, as this would represent the choice of centroid positions that best fitted the feature data.

C- How do you use the Scree Plot for choosing the number of K-means Cluster.

A Scree Plot allows you to check the “Elbow” in the plot, where its gradient changes from steep to almost flat and that is the optimal point of K.

Artificial Intelligence Risk Certificate

2.3 What are the two types of hierarchical clustering? What are the advantages of hierarchical Clustering?

Hierarchical Clustering starts with all points in one cluster and then sequentially splits them into separate clusters until an optimal allocation is reached (divisive Hierarchical Clustering) or starts with each data point in its own cluster and sequentially combines them until an optimal allocation is reached (agglomerative hierarchical clustering).

The Advantages of hierarchical clustering are firstly, that it does not require a pre-defining number of clusters and secondly, it can uncover hierarchical relationships within the data, which can reveal nested clusters within larger groups. Thirdly, the dendrogram produced are really straightforward to interpret.

2.4 State if True or false

A- K-means with Euclidean distance can only be used when the clusters are approximately spherical.

True, because K-means is based on Linear Euclidean distances, it runs into problems when the cluster is not approximately spherically shaped. When the Manhattan distance measure is used with K-means, the clusters are approximately rhombus-shaped (losango). This can result in poorly defined clusters, or some points even being allocated to the wrong clusters.

Artificial Intelligence Risk Certificate

B- The WCSS will never rise when the number of clusters is increased in a K-means application

True, the situation is analogous to what happens to the residual sum of squares when more features are added to a linear regression. When additional clusters are added (i.e., the value of K increases) the fit of the model to the data cannot get worse. Therefore, the WCSS must fall as the new cluster will capture one or more of the data points better than the cluster to which it was previously allocated.

2.5 What is a Dendrogram and how would we interpret one?

A dendrogram is a pictorial representation of the steps in a hierarchical clustering application, which shows how the clusters are split or combined. Each bifurcation (for divisive clustering) or combination (for agglomerative clustering) of the lines shows a cluster being formed or removed, respectively. The heights of the vertical lines show the impact of the marginal cluster on the distance of the points affected by it to their nearest cluster. If a particular vertical line is long, this would suggest that the additional cluster has a considerable effect on the model fit and therefore is worth incorporating.

2.6 Determine the Centroid of a cluster comprising Banks A,B and C using the Raw (unscaled) data.

Artificial Intelligence Risk Certificate

Features	Bank A	Bank B	Bank C
Number of customers (millions)	1.2	6.0	0.5
Total size of loan book (USD bn)	5	25	7
Number of branches	80	400	50

The centroid is simply the average of the feature values across the three banks. So, if we define the coordinates space as three dimensional point, with the raw data in their original units, the centroid is given by

$$\left(\frac{1.2 + 6.0 + 0.5}{3} + \frac{5 + 25.0 + 7}{3} + \frac{80 + 400 + 50}{3} \right)$$

Which is (2.567, 12.333, 176.667)

3.0 Supervised Learning for Numerical Data

Learning objectives:

Identify uses and limitations of single and multi-variable linear and non-linear regression.

Interpret the results of single and multi-variable regression and non-linear regression.

Identify problems that may occur with linear regression models and possible remedies for them.

Describe how logistic regression models can be applied to classification problems.

Describe the use of linear discriminant analysis for classification problems.

This chapter covers models that used for supervised learning arising from econometrics, techniques originating from computer science and more commonly associated with machine learning.

Single Variable and Multiple Variable Linear Regression

The simplest linear regression model is one that has only a single feature or input, variable x along with the output or target variable y .

Using multiple variable linear regression, it is possible to explore interactions between variables by incorporating interaction terms. It is also possible to incorporate power terms in a multiple variable regression.

Artificial Intelligence Risk Certificate

The Main methods for estimating the parameters of the regression model are:

Least squares

Maximum Likelihood

The method of moments

A specific case of the first technique, least squares, it known as the Ordinary Least Squares (OLS), which is the most straightforward and commonly used approach for linear regression models.

Simple Linear regression is given by:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

While Multiple linear regression

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_m x_{mi} + u_i$$

Ordinary least squares, works to estimate the correct linear relation as trying to minimize the error of using x to predict y.

Problems with features and functional form

The model omits relevant features such as a lack of Data or lack of awareness of their relevance, and the parameters being estimated

Artificial Intelligence Risk Certificate

The model includes irrelevant features, then estimate precisely and is hard for the model to generalize from the specific training sample to the test sample.

The model incorporates features in the wrong way, known as incorrect functional form,

Multicollinearity

Perfect Multicollinearity is when two features have an exactly linear relationship such as , $x_2=10x_3$

Near Multicollinearity is when two or more of the variables are closely, but not perfectly, correlated with one and another.

Techniques for addressing near multicollinearity include:

Removal of one or more highly correlated variables

Turning the highly correlated variables into a ratio or difference

Use regularization

Outliers

There is no widely accepted formal definition of an Outlier, and in broad terms refers to an anomalous data point that lies a long way from the others. They can have a considerable effect on the estimated parameters.

Artificial Intelligence Risk Certificate

In order to detect outliers we can use a Residual Plot, in which we examine the plot of the residuals (the difference between the actual data points and the corresponding values fitted from the regression line), and noting any point that lie further from the line than others.

We can also use the Cooks Distance, which measures the influence of each individual data point on the parameter estimates. Achieves by removing each data point separately from the regression and determining the difference in model fit for all the remaining points. The bigger the Cooks distance the more influence the data point has on parameter estimation.

Heteroskedasticity

Refers to non-constant variance of the error term and can lead to inefficient parameter estimation and errors in the determination of the statistical importance features.

A plot of residuals against fitted values can be helpful in identifying heteroskedasticity.

In a good Model, a model that is Homoscedastic, we will see errors all clustering around zero.

If the model is highly heteroskedastic, it's probably because we lack using a relevant variable.

Artificial Intelligence Risk Certificate

Classification Problems

There are many instances where a model's output (dependent variable) is categorical.

Predicting a qualitative outcome is defined to be a classification problem and assigning an observation to one class rather than another is referred to as classifying the observation.

A specific case of categorical data is where the output is binary, that is, it only has two outcomes.

We might be interested in modeling the probability of one of the outcomes occurring.

One outcome (referred to as Positive outcome) is assigned a value of 1 and the other (referred as negative outcome) is assigned a value of 0.

A Standard linear model is inappropriate in such case because there would be nothing in the models design to ensure that the estimated probabilities lie between zero and one, and we could obtain nonsensical predictions.

Logistic Regression

A Logistic regression uses a cumulative logistic function transformation, resulting in the output being bounded between zero and one.

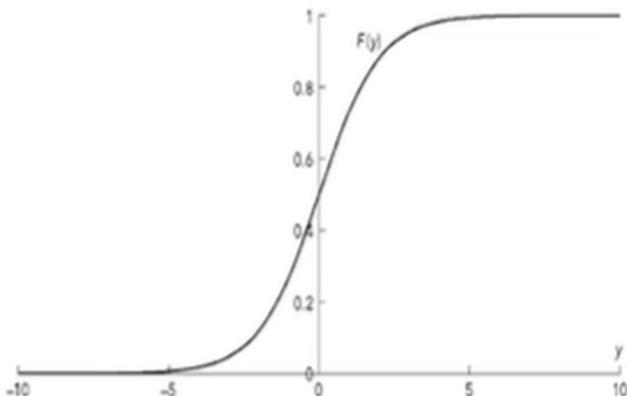
The logistic function has a sigmoid shape and is written by

Artificial Intelligence Risk Certificate

$$F(y_i) = \frac{1}{1 + e^{-y_i}}$$

When there are m features, the functional form y_i is estimated as,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_m x_{mi} + u_i$$



Linear Discriminant Analysis

A Logistic regression works really well for binary classification problems.

When there are multiple, well separated classes, logistic regression estimates can be very unstable. In this case, an alternative is offered by LDA.

Artificial Intelligence Risk Certificate

Like with logistic regression, the idea is to assign each instance to the class with the highest conditional probability.

A discriminant function is calculated for each of the classes which gives the probability that a new data point belongs to that class.

New data points are classified based on which class has the highest probability.

3.1.1. Simple Linear Regression

Also known as bivariate regression because there are only two variables.

Simple Linear regression is given by:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

The model postulates that y varies due to changes in x , here y is a linear function of x and an unobservable error term, u , with mean zero and constant variance.

x and y are observable variables (y is the target and x is the feature, in machine learning parlance), where as β_0, β_1 are the parameters to be estimated.

Using econometric terminology β_0 is the intercept parameter, and it is interpreted as the value that y would take if x equals to zero. β_1 is the slope, and measures the impact on y of a unit change of x .

Artificial Intelligence Risk Certificate

In machine learning, the intercept is known as the bias and the slope is the weight.

There are three main methods for estimating the parameters of a regression model:

Firstly, **Least Squares**

Secondly, the **Maximum Likelihood**

Thirdly, the **Method of Moments**

The Ordinary least Squares is the most straightforward approach and hence it is most used for linear regression models.

A linear regression model embodies a linear relationship that can be represented by a straight line. Being slightly more specific, the model is both linear in the parameters and linear in the variables.

In order to use OLS, the model must be linear in the parameters, although it does not necessarily have to be linear in the features.

3.1.2. Multiple Linear Regression

In the majority of cases, having a model with an unique feature is not sufficiently flexible to capture all the variability in the target variable, and we can build a much better model by considering multi variables.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_m x_{mi} + u_i$$

In the multiple linear regression, there will be $m+1$ parameters ($m \geq 1$) to estimate. One for the intercept, and one for each of the m slope parameters.

Once we have data on y and $x_{1i} \dots x_{mi}$, again, OLS can be used to estimate the Parameters.

In the multiple linear regression model, each parameter measures the partial effect of the attached variable after controlling for the effects of all the other features included in the regression.

Even within this straightforward framework, we can nonetheless incorporate a wide range of specifications. For instance, it is common to apply the logarithmic transformation to some or all the feature variables and/or the output variable.

Such a transformation would imply a different interpretation of the parameter estimates but OLS could still be used as the model would remain linear in the parameters.

Artificial Intelligence Risk Certificate

Alternatively, we could incorporate interaction terms, i.e. features multiplied together, or power term of features:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}x_{2i} + u_i$$

Where, y depends not only on the levels of x_{1i} and x_{2i} , but also on how they work together. Hence, β_3 will capture any complementarity between them, where changes in both variables are required to have an impact on y rather than in isolation.

A commonly employed example is how the amounts of water and fertilizer affect crop yields, because an abundance of one and none of the other will not lead to high yields. Hence, the amount of one of them influences the effectiveness of the other, implying a need to model them jointly.

A model including power term could be as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{2i}^2 + u_i$$

Here, only a squared term on x_{2i} is included on the model, and it is common to stop there, allowing for a quadratic relationship between x_{2i} and y .

But, in principle, would be feasible to include cubed terms, fourth-order powers, and so on. However, when adding further terms, one must be mindful not to overfit the data.

To use OLS for model estimation, the output variable y must be continuous, but the features could be

Artificial Intelligence Risk Certificate

continuous or discrete, being that the discrete ones should be encoded by dummies.

Box 3.1: The effect of experience and having a degree on wage rates

Suppose that we have the data given in [Table 3.1](#), which show the hourly salaries in US dollars of 14 notional employees at a US retail bank alongside the number of years of experience that each has, the square of the number of years of experience, and a dummy variable capturing whether they have a college degree.

Table 3.1: Salary, experience, and degree dummy variable for notional sample of bank employees

Experience	Experience ²	Degree	Salary
1	1	0	15.46
3	9	1	22.40
7	49	1	27.47
12	144	1	34.31
9	81	0	33.08
3	9	0	18.70
20	400	1	35.06
22	484	0	35.78
0	0	1	14.81
4	16	1	14.84
6	36	0	25.78
8	64	0	28.17
4	16	1	26.36
2	4	0	11.12

In this example, we believe that salary should be driven by experience. So, to begin with, we run a simple linear regression of salary (y) on experience (x_1), and we would obtain the following:

Artificial Intelligence Risk Certificate

$$\widehat{\text{salary}}_{y_i} = 16.88 + 1.06 \text{experience}_i$$

We use $\widehat{}$ above the output variable to denote the fitted equation from the regression line. Here, the intercept estimate is 16.88, meaning that someone just joining the bank with no experience could expect to earn 16.88 per hour on average, and each additional year of experience would lead to an average salary increase of 1.06.

After a while.

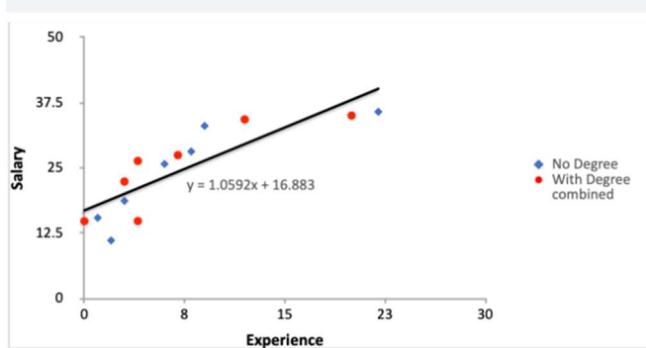


Figure 3.1: Experience versus salary and fitted straight line

The plot suggests that the line fits the data well, although it seems that additional years of experience lead wages to increase significantly when experience is low but the impact tails off after a while.

It appears from the diagram that each additional year of experience leads to a lower incremental increase in wages.

Artificial Intelligence Risk Certificate

We cannot capture this with a linear model, because it embodies a fixed gradient of the fitted line and therefore a fixed relationship between x and y whatever the value of x .

In order to capture the potential non-linearity in the relationship between ages and experience, we can fit a quadratic model, where the squares of experience is included in the model as a second feature.

The fitted model is now:

$$\widehat{\text{salary}}_{y_t} = 11.82 + 2.77\text{experience}_i - 0.08\text{experience}_i^2$$

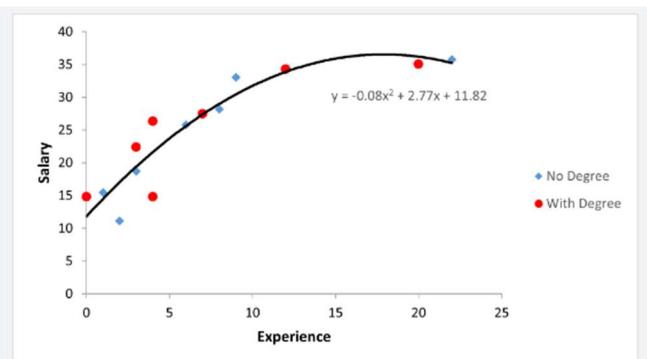


Figure 3.2: Experience versus salary and fitted quadratic equation

Note that when the square of experience is added, the remaining parameters estimation change compared to their values when only the level of experience was included on the model.

Artificial Intelligence Risk Certificate

Know that we have two features relating to experience, determining the relationship between experience and salary is slightly trickier.

$$\frac{\widehat{\delta \text{salary}_{y_i}}}{\delta \text{experience}_i} = 2.77 - 2 * 0.08 \text{experience}_i$$

It is clear that when the square of experience is included, the relationship between salary and experience is no longer constant with respect to experience.

Setting the expression to zero and rearranging it to make experience the subject of the formula will show the value of the feature that maximizes expected salary, which is 29.6 years.

The regression model now includes two variables, experience and its squared value, but the degree feature can affect the salary as well. This is captured by a dummy variable taking the value 1 if the person has a degree and zero if they do not.

$$\widehat{\text{salary}_{y_i}} = 11.38 + 2.76 \text{experience}_i - 0.08 \text{experience}_i^2 + 0.92 \text{degrees}_i$$

The 0.92 should be interpreted as an employee with a degree could expect to earn an additional 92 cents per hour on average compared with someone having identical experience but no degree.

If we had more than one dummy variable in the model to capture other qualitative information, we would interpret the associated parameters the same way.

3.1.3. Potential problems with Regressions

In some cases, even if we can estimate values for the parameters, these ones might no longer be optimal or reliable.

Problem 1 is the use of wrong features or wrong functional form.

On the previous example, we assumed that the most appropriate features to model variation in wage rates for bank employees were based on their experience and degree level qualifications.

Nevertheless, it might be that the model did not include the most relevant features, and broadly there are three ways that the model can be wrong.

Firstly, the model omits some relevant features, which could occur if the true relationship describing the output includes some extra features that the researcher has not included in the model, due to lack of data or unawareness of their relevance. This could be a serious misspecification that could lead the parameter estimates to be biased and not become more accurate as the sample size increases.

Secondly, the model includes some irrelevant features, this is less serious than the first misspecification, but can result in inefficiency where the parameters are not estimated precisely. The model will find out to be hard to generalize from the specific training sample to the test data.

Artificial Intelligence Risk Certificate

Thirdly, the model includes the correct features, but they are incorporated in the wrong way. This is known as an incorrect functional form. It could occur, for instance, is the true relationship between the features and the output is non-linear but a linear regression model is used.

These three problems are hard to resolve in practice than they appear, because the researcher never knows the true relationship between features.

The remedy for this is Strong theoretical knowledge of the problem at hand and the wider context can be valuable in guiding the model development, rather than a purely data driven approach.

Problem 2 is multicollinearity which occurs when the features are highly related to one another.

Perfect multicollinearity occurs when two or more of the features have an exactly linear relationship that holds for every data point. The solution is to remove one or more of these perfect correlated features from the model.

Near multicollinearity occurs when two or more features are almost perfectly correlated. A common consequence is that the parameter estimates become highly unstable, changing wildly when a feature is added or removed from the model. The remedies include removing one or more of these features or turning them into a ratio/difference rather than including them individually.

Artificial Intelligence Risk Certificate

Problem 3 are Outliers, which are anomalous data points that lies a long way from the others. These ones can even a high impact on the parameter's estimations.

The least square technique used to estimate the parameters in a regression model takes the sum of square of distances from the points to the fitted line and the process of squaring these distances means that points that are a considerable distance from the others will exert a disproportionate effect on the estimates.

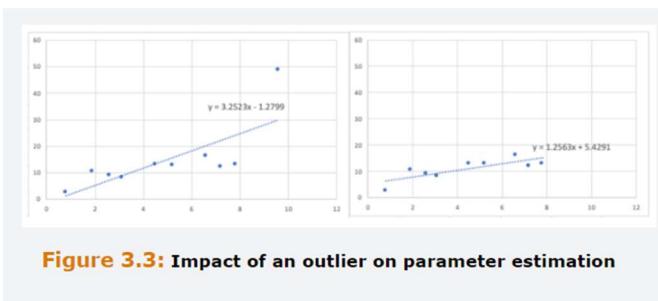


Figure 3.3: Impact of an outlier on parameter estimation

Outliers can be detected by examining a plot of the residuals, the difference between the actual data points and the corresponding fitted values from the regression line, and noting any points that lie further from the line than others.

Note that if both the input and output values are further from the other datapoints but the point nonetheless lies near the regression line, this would not be classified as an outlier.

Artificial Intelligence Risk Certificate

A more sophisticated method is to calculate the Cook's distance, which measures the influence of each individual data point on the parameter estimates.

This is achieved by removing each data point separately from the regression and determining the difference in model fit for all the remaining data points. If a particular data point is not very influential for parameter estimation (hence is not an outlier), the model fit will not be changed by a lot and the Cook's distance will be small.

Problem 4 is heteroskedasticity, meaning that the variance is not constant, and occurs frequently in time series data.

Can lead to several issues with regression estimation, most notably that it becomes inefficient and that it is hard to accurately evaluate the empirical importance of each future for determining the output.

As for outlier detection, a residual plot can sometimes be useful in detecting heteroskedasticity, where we would be looking for whether the spread of the residuals around their mean, usually of zero, is constant or systematically changing.

There are also some statistical tests for Heteroskedasticity, such as Goldfeld-Quandt test, which splits the sample into two parts and statistically compares the residual variances between the two.

Alternatively, the White's test involves obtaining the fitted values, \hat{y}_i , from a regression and conducting a

Artificial Intelligence Risk Certificate

second, auxiliary, regression of the squares of the fitted values on the squares of the features and interactions between them. If there is no heteroskedasticity, the parameter estimates from this auxiliary regression will not be statistically significant.

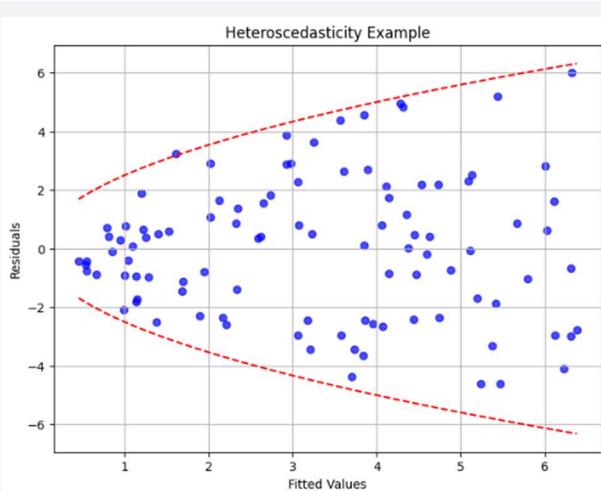


Figure 3.4 Plot of fitted values against residuals to demonstrate heteroskedasticity

Several remedies for Heteroskedasticity are to weight the observations to account for the changing error variance using a technique known as weighted least squares (WLS) instead of OLS. Alternatively, making a logarithmic transformation of the variables and using them in place of the raw variables.

3.1.4. Stepwise Regression Procedures

As discussed above, the presence of non-informative or redundant features in linear regression model can add uncertainty to the predictions and reduce the effectiveness of the model, especially in the presence of highly correlated features.

This introduces the need to remove non informative predictors.

Stepwise regression, like LASSO regularization technique is a method for feature selection. It belongs to the category of wrapper methods, which add or remove predictors to a regression with the aim of finding the combination that maximizes the model performance.

The inclusion or exclusion of features is based on criterion that measures the predictive accuracy of alternative set of predictors. A popular approach is to choose the model that minimizes the Akaike information criteria (AIC) which is a measure of prediction errors adjusted to account for the number of features in the model.

Unlike R squared, AIC penalizes large models and therefore it can either increase or decrease when an additional feature is added to the model.

There are various stepwise procedures, but the most straightforward are the unidirectional forward and backward stepwise selection methods.

Artificial Intelligence Risk Certificate

The forward starts with a model with no features and includes additional features into the model one-at-the time starting from those that produce the largest drop in the AIC. The procedure stops when the addition of any new predictor fails to decrease the AIC.

The backwards begins with the full model and removes the predictors one-by-one starting with the least important, until any further elimination fails to decrease the AIC.

	Forward Selection	Backward Selection
Step 0	No predictors (AIC: 3202.69)	All predictors: (AIC: 3122.82)
Step 1	Age (AIC: 3193.36) *	All – Age ⁴ (AIC: 3120.91) *
	Age ⁵ (AIC: 3202.2)	All – Age ³ (AIC: 3120.98)
	Age ² (AIC: 3203.33)	All – Age ⁵ (AIC: 3121.20)
	Age ⁴ (AIC: 3203.63)	All – Age ² (AIC: 3123.96)
	Age ³ (AIC: 3204.66)	
Step 2	Age + Age ³ (AIC: 3123.20) *	All – Age ⁴ – Age ⁵ (AIC: 3131.17)
	Age + Age ² (AIC: 3125.5)	All – Age ⁴ – Age ³ (AIC: 3153.06)
	Age + Age ⁴ (AIC: 3126.69)	All – Age ⁴ – Age ² (AIC: 3169.77)
	Age + Age ⁵ (AIC: 3132.87)	No variables to be removed
Step 3	Age + Age ³ + Age ⁵ (AIC: 3124.50)	
	Age + Age ³ + Age ⁴ (AIC: 3124.77)	
	Age + Age ³ + Age ² (AIC: 3125.19)	
	No variable to be included	
Selected Model	Age + Age ³	All – Age ⁴

As it emerges from the above example, the two procedures will not necessarily select the same model. In fact, more generally, there is no guarantee that either of the two methods will select the optimal model, as only a subset of the 2^m models, in which m is the number of predictors, is considered.

Artificial Intelligence Risk Certificate

Although in principle it is possible to estimate all the possible models and compare them, this is very impractical when the set of candidate predictors is large. Therefore, when the set of predictors is large, both forward and backwards stepwise regression offers a valid alternative.

Both procedures have their pros and cons and the choice between the two depends on the problem at hand. Backwards selection tends to be more computationally efficient when the set of candidate predictors is large. However, as the full model is the first to be estimated, the number of predictors is required to be strictly smaller than the number of observations. In contrast, forward selection can also be applied when the number of predictors is larger than the number of observations.

3.2. Classification Problems

In finance, a lot of times, the model output, dependent variable, is categorical and where the output for each observation can only be one of a small number of categories.

In the Machine learning jargon, the problem of prediction a qualitative outcome is defined to be a classification problem and assigning an observation to one class rather than another is referred to as classifying the observation.

In binary situations, it is interesting to model the probability of one of the outcomes occurring. One outcome, referred to as the positive outcome, is assigned a value of one, and the other, the negative one, a value of zero.

A Standard linear model would be inappropriate to apply in such cases because there would be nothing in the model's design to ensure that the estimated probabilities lie between zero and one, and we could obtain nonsensical predictions.

Artificial Intelligence Risk Certificate

3.2.1. Logistic Regression

This specification uses a cumulative logistic function transformation resulting in the output being bounded between zero and one.

The logistic function has a sigmoid shape and is written by

$$F(y_i) = \frac{1}{1 + e^{-y_i}}$$

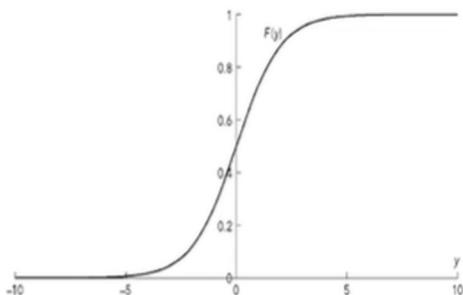
When there are m features, the functional form y_i is estimated as,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_m x_{mi} + u_i$$

And the probability that $y_i = 1$ is given by:

$$P_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_m x_{mi} + u_i)}}$$

And the probability that $y_i = 0$ is $(1-P_i)$



Artificial Intelligence Risk Certificate

The parameter estimates from a logit model cannot be interpreted in the usual fashion due to the presence of the logistic transformation, which is nonlinear. Nevertheless, their signs and levels of statistical significance can still be examined.

From an example of credit lending, in which 0 is not to default and 1 is to purely default, the borrowers with longer loan terms and those paying higher interest rates have a significantly higher probability of default, whereas those with a mortgage have a significantly lower probability of default.

Table 3.3 Parameter estimates from a logistic regression to model loan outcomes.

Parameter	Definition	Estimate	Standard error
Bias term	The intercept	-5.3041***	1.051
Amount	Total sum borrowed	-0.0001	0.000
Term	Length of the loan (months)	0.0768**	0.034
Interest rate	APR charged (%)	0.1147**	0.045
Installment	Monthly installment	0.0025	0.004
Employment history	Length of borrower's employment history (years)	0.0428	0.059
Homeowner	1 = owns home; 0 otherwise	0.1149	0.409
Mortgage	1 = has a mortgage; 0 no mortgage	-0.9410**	0.435
Income	Annual income (USD)	-0.0001	0.000
Delinquent	Number of times borrower has been more than a month behind with payments in the past two years	0.0985	0.113
Bankruptcies	Number of publicly recorded bankruptcies	-0.1825	0.361

3.2.2. Other Types of limited Dependent Variable Models

Discrete choice models are based on models that try to predict a categorical, multiple choice, dependent variable.

The Classical example is where a commute chooses between different transport types and we are interested in modelling the probability that a particular individual will travel by each mode, which would be the output.

As per the Binary output It would not be appropriate to use linear regression model, we should use an extension of the logit regression describe on the previous chapter, known as **Multinomial Logit Models**.

We estimate models for all categories but one, which will serve as a baseline category. As per the above example, we could model the probability that a commuter will choose to travel by car, and the probability of traveling by bike. Then the probability of traveling by bus is simply one minus the sum of the two calculated probabilities.

Ordinary variables models, distinct from the above because here the categorical variables have implicating ordering, such as credit ratings, risk scales from 1-10 and so on.

For modelling ordinal variables where there are more than two outcomes, ordered logit models would be used. The estimation principles are the same as for the

Artificial Intelligence Risk Certificate

binary case, but the values of cutoff parameters between categories must also be estimated.

In some circumstances, the values of the output variable in a model that we observe are not sampled randomly from the population, resulting in a biased sample.

3.3 Linear Discriminant Analysis

When there are multiple, well-separated classes, the estimates from a logistic regression turn out to be very unstable. In this case, an alternative to the logistic regression is offered by the LDA.

LDA assumes that the joint distributions of features is multivariate normal with a common variance-covariance matrix, but with different mean vectors. Like logit, the idea is to assign each instance to the class with the highest conditional probability.

A discriminant function is calculated for each of the classes. It is the probability that a new data point belongs to that class. New data points are classified based on which class has the highest probability. It is possible to show that the discriminant function $\hat{\delta}_j(x)$ are linear functions of x for each class j , for $j=1,\dots,g$.

$$\hat{\delta}_j(x) = x^T \hat{\Sigma}^{-1} \hat{\mu}_j - \frac{1}{2} \hat{\mu}_j^T \hat{\Sigma}^{-1} \hat{\mu}_j + \log(\hat{\pi}_j)$$

Where x is the feature vector, $\hat{\mu}_j$ is the vector containing the means of the predictors of each g classes, estimated

Artificial Intelligence Risk Certificate

using the training sample data, $\hat{\Sigma}^{-1}$ is the inverse of the data covariance matrix, $\hat{\pi}_j$ is the prior probability of class j, and T denotes the transpose operator.

The prior probabilities could be assumed to be equal for all classes or estimated using the frequency of class j in the training data. A new data point is assigned to the class for which $\hat{\delta}_j(x)$ is the largest. LDA has proven to work greatly in practice, even when the assumptions are not met.

As an example, consider a sample of 10 borrowers to be classified as defaulted or not default. Because LDA, like PCA, requires data scaling, the data provided in the table have been already standardized.

Default	Balance	Income
0	-0.55	-0.75
0	-0.23	0.19
0	-0.76	0.53
0	-0.08	-0.73
0	-0.72	1.32
0	-0.79	2.00
0	-0.80	-0.34
1	0.46	-0.71
1	1.95	-0.96
1	1.51	-0.54

A value of 1 means that it defaulted, 0 otherwise.

Artificial Intelligence Risk Certificate

A first stage is to calculate the class means for each predictor from the data by averaging the observations that belong to each class.

$$\hat{\mu}_0 = \begin{pmatrix} -0.56 \\ 0.32 \end{pmatrix}, \hat{\mu}_1 = \begin{pmatrix} 1.31 \\ -0.74 \end{pmatrix}$$

The covariance matrix can be estimated as,

$$\hat{\Sigma} = \begin{pmatrix} 1 & -0.58 \\ -0.58 & 1 \end{pmatrix}$$

Where the variance of both predictors is equal to 1 as the predictors have been standardized. The inverse of this matrix is,

$$\hat{\Sigma}^{-1} = \begin{pmatrix} 1.52 & 0.88 \\ 0.88 & 1.52 \end{pmatrix}$$

Finally, we can set the prior probabilities to 0.7 and 0.3, based on the frequencies of each class in the training sample.

Suppose now that we have to classify an applicant with a standardize balance equaling to -1.42 and a standardize income equaling to -0.2. This means that we can write $x^T(-1.42 \quad -0.20)$

For Default we would have,

$$\delta_0 = (-1.42 \quad -0.20) \begin{pmatrix} 1.52 & 0.88 \\ 0.88 & 1.52 \end{pmatrix} \begin{pmatrix} -0.56 \\ 0.32 \end{pmatrix} - \frac{1}{2}(-0.56 \quad 0.32) \begin{pmatrix} 1.52 & 0.88 \\ 0.88 & 1.52 \end{pmatrix} \begin{pmatrix} -0.56 \\ 0.32 \end{pmatrix} + \log(0.7) = 0.30$$

And nondefault:

$$\delta_1 = (-1.42 \quad -0.20) \begin{pmatrix} 1.52 & 0.88 \\ 0.88 & 1.52 \end{pmatrix} \begin{pmatrix} 1.31 \\ -0.74 \end{pmatrix} - \frac{1}{2}(1.31 \quad -0.74) \begin{pmatrix} 1.52 & 0.88 \\ 0.88 & 1.52 \end{pmatrix} \begin{pmatrix} 1.31 \\ -0.74 \end{pmatrix} + \log(0.3) = -3.95$$

Because nondefault is higher the applicant is nondefault.

Appendix 3.A The Heckman 2 stage Procedure

The values of the output variable in a model that we observe may not be sampled randomly from the population, which leads to a biased sample.

For instance, the willingness to respond to a survey is often correlated with the variables we are trying to measure. So if we asked bank costumers to complete the survey indicating their satisfaction with the level of service that they have received, those who are the least happy may be those most likely to complete the survey.

If we were then interested in modelling the factors that affect costumer satisfaction levels using these survey results, our parameter estimates would be biased.

Another situation where such an issue would occur is in the context of modelling the value of share repurchases. Most listed firms do not make share repurchases, and therefore the output variable would have some problematic characteristics: no observations would be negative, and the bulk of observations would have a value of 0, with the remainder having a distribution a long way from 0.

The Heckman approach deals with these situations by separating them into two stages.

Firstly, model the probability that a bank costumer will complete a survey or firm will choose to make share repurchases using a binary logit function.

Artificial Intelligence Risk Certificate

Secondly, model the determinants of costumer satisfaction among those who have elected to complete the survey or model the size of the repurchase among firms that have chosen to make them.

Appendix 3.B Fisher Discriminant Analysis

Is an alternative to the Linear discriminant analysis (LDA). The idea is to find a vector on which to project the data such that the maximum between group variance of the projection relative to withing-group variance is obtained.

In this respect, LDA can be related to dimensionality reduction techniques. We find a projection on the data on a lower dimensional space that is optimal for classification of the data. Once this vector have been found, new data to be classified are projected onto this vector and assigned to the class whose mean they are closer to.

In other words, we aim to find a way to separate the data into g distinct classes such that the distance between the means of different classes is maximized while the variation within each class is minimized. Technically, we find the vector b that maximizes the signal-to-noise ratio, which is given by the ratio of the between and within group variances.

$$\frac{b^T B b}{b^T W b}$$

Artificial Intelligence Risk Certificate

Where B is the between Group covariance matrix and W is the Within group covariance matrix.

To make it concrete, consider a simple case where only two classes are available. In this case, solving the maximization problem gives:

$$b = \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$$

Where $\hat{\Sigma}$, $\hat{\mu}_2$ and $\hat{\mu}_1$ are the covariance matrix, and the class means respectively. The discriminant vector is perpendicular to b.

Therefore, the discriminant score of a new data instance x is obtained by projecting x onto vector b:

$$x^T \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$$

We classify x as belonging to class 2 if $x^T \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > c$, in which c is a threshold to be decided by the researcher. If we are ready to assume that the two classes display approximately the same distribution, then the optimal threshold is given by:

$$c = \frac{1}{2} \hat{\mu}_2^T \hat{\Sigma}^{-1} \hat{\mu}_2 - \frac{1}{2} \hat{\mu}_1^T \hat{\Sigma}^{-1} \hat{\mu}_1$$

In this case, the Fisher formulation of the problem is equivalent to the probabilistic formulation that was presented earlier for LDA with equal prior probabilities.

Artificial Intelligence Risk Certificate

Appendix 3.C Linear Discriminant Analysis Example

Data

Subscribed	Balance (Dollars)	Age
yes	2343	59
yes	45	56
yes	1270	41
yes	381	37
yes	40	35
yes	22	31
no	390	29
no	6	53
no	71	58
no	0	33

Subscribed is the outcome of the campaign, Balance is the outstanding balance of the Client at the bank in US Dollars and Age is the age of the client in Years.

Perform the following Tasks:

A. Standardize the features.

Artificial Intelligence Risk Certificate

- B. Estimate the within-group means for the clients who subscribed (class 1) and those who did not subscribe (class 2).
- C. Compute the data covariance matrix
- D. How would you classify a client who is 51 years old and has an outstanding balance of 10,635 Dollars?

A.

To standardize the features, we first must compute their sample mean and standard deviation. The sample mean of the first feature is $\widehat{balance} = 456.8$ and its standard deviation is $\widehat{\sigma}_b = 769$. The sample mean of the second feature is $\widehat{Age} = 43.2$ and its standard deviation is $\widehat{\sigma}_a = 11.99$.

The Standardized features are obtained from subtracting the mean and divided by the standard deviation.

Subscribed	Balance	Age
yes	2.45	1.32
yes	-0.54	1.07
yes	1.06	-0.18
yes	-0.10	-0.52
yes	-0.54	-0.68
yes	-0.57	-1.02
No	-0.09	-1.18
No	-0.59	0.82
No	-0.50	1.23
No	-0.59	-0.85

Artificial Intelligence Risk Certificate

B. The within groups mean are simply the means of the features for each class. For Class 1, the vector of the features means would be

$$\mu_1 = \begin{pmatrix} 0.29 \\ 0.00 \end{pmatrix}$$

Where the first is obtained by averaging the Class “yes” standardized values of Balance and the second the standardized values of age.

Same thing for second class to obtain

$$\mu_2 = \begin{pmatrix} -0.44 \\ 0.00 \end{pmatrix}$$

C. As the two features have been standardized, their variance are both equal to one. Therefore, we only need to compute the covariance between the two features, which is given by:

$$cov(x_1, x_2) = \frac{\sum_i^N (x_1 - \hat{x}_1)(x_2 - \hat{x}_2)}{N - 1}$$

Hence, the covariance Matrix is:

$$\begin{pmatrix} 1 & 0.33 \\ 0.33 & 1 \end{pmatrix}$$

D. The only further piece of information needed to determine is the probability of each class. The probability of yes is 0.6 and 0.4 is the probability of no. The inverse covariance matrix can be easily obtained by hand or using the function MINVERSE in Excel.

Artificial Intelligence Risk Certificate

The standardize features of the new data point are:

$$age_{sd} = \frac{51 - 43}{12} = 0.65$$

$$Balance_{sd} = \frac{10635 - 457}{769} = 13.24$$

Therefore we get:

$$\begin{aligned}\hat{\delta}_1 &= (13.24 \quad 0.65) \begin{pmatrix} 1.12 & -0.37 \\ -0.37 & 1.12 \end{pmatrix} \begin{pmatrix} 0.29 \\ 0.00 \end{pmatrix} - \frac{1}{2}(0.29 \quad 0.00) \begin{pmatrix} 1.12 & -0.37 \\ -0.37 & 1.12 \end{pmatrix} \begin{pmatrix} 0.29 \\ 0.00 \end{pmatrix} \\ &\quad + \log(0.6) = 3.76\end{aligned}$$

For Class 2

$$\begin{aligned}\hat{\delta}_2 &= (13.24 \quad 0.65) \begin{pmatrix} 1.12 & -0.37 \\ -0.37 & 1.12 \end{pmatrix} \begin{pmatrix} -0.44 \\ 0.00 \end{pmatrix} - \frac{1}{2}(-0.44 \quad 0.00) \begin{pmatrix} 1.12 & -0.37 \\ -0.37 & 1.12 \end{pmatrix} \begin{pmatrix} -0.44 \\ 0.00 \end{pmatrix} \\ &\quad + \log(0.4) = -7.51\end{aligned}$$

Since the Class 1 is higher, we predict to be class one. In fact, from the large value of standardized balance, 13.24, it is quite clear that the new client should be a subscriber.

Questions and Answers Module 2 Chapter 3 from GARP

3.1

A. What is an Outlier in the Context of regression?

An outlier is a data point that demonstrably does not fit with the pattern of the others so that in a regression context, the fitted and actual values would be a long way apart, leading to a residual of larger magnitude than the others.

B. How Can outliers be detected?

There are various methods available to detect outliers. A good first step is to examine the residuals from the purposed model to see whether any are significantly larger in absolute value than the others. More formally a measure known as Cook's distance can be calculated for each point. This evaluates how much each parameter would change if a given data point were excluded from the sample. Large Values of Cook's distance indicate a point that would be more likely considered an outlier.

Artificial Intelligence Risk Certificate

3.2 What is a stepwise regression and how does it work.

Stepwise regression is a technique for feature selection. Beginning with a list of candidate features that could be included in the model, the analyst selects an approach: Either beginning with a model containing no features (forward selection) or containing all the features (backwards selection). With forward selection, the analyst adds the feature that would have the most additional explanatory power until a further addition does not decrease the AIC. With Backwards selection, all features are initially included in the model, and they remove one-by-one starting with the feature having the least explanatory power until removing a further variable fails to decrease AIC.

3.3 Explain why linear Regression cannot be used when the dependent variable in a regression model can only take the values 0 or 1.

Linear regression is used to fit models that use numerical data that is continuous. But here the target variable is a binary value. When linear regression is used, there is nothing in the estimation process that would ensure the fitted values from the regression model would lie between 0 and 1. Truncating the fitted values to 0 and 1 would be inadvisable as the result would be too many values at these extreme points.

Artificial Intelligence Risk Certificate

3.4 Explain the main assumptions underlying LDA.

The Standard approach to Linear discriminant Analysis assumes that the data arise from g multivariate normal distributions with different mean vectors but common covariance Matrix.

3.5 Exercise

Suppose that we have the following results for a model estimated using ordinary least squares for the relationship between a firm's return on Assets (ROA) and its SIZE.

$$\widehat{ROA}_i = 2.6 + 4.22SIZE_i - 1.32SIZE_i^2$$

Where ROA is measured in percent and SIZE is measured in \$M.

A . According to this model, is the relationship between SIZE and ROA Linear?

No, because the equation includes a squared term in SIZE, the relationship between SIZE and ROA is non-linear, therefore the relationship between the two variables will depend on the value of SIZE.

B . Comparing two firms with market caps of 100M and 110M, what would be the difference in ROA?

Best way is just to replace SIZE with the values, calculated ROA and the Differences between both SIZE values. It would be 0.14% higher for the 110M one.

Artificial Intelligence Risk Certificate

C . What would be the optimal size of firm for an investor to choose, if they wanted a firm with maximal ROA, and what ROA would this generate.

$$\frac{\delta \widehat{ROA}_i}{\delta SIZE} = 4.22 - 2.64SIZE_i$$

We need to set the derivative to zero, and rearrange for SIZE to give the value of SIZE that Maximizes ROA:

$$4.22 - 2.64SIZE_i = 0 \Leftrightarrow SIZE_i = 1.6$$

Therefore, a firm with market cap of 160M\$ would have the highest possible value of ROA. To calculate the latter, we simply plug 1.6 into the original fitted equation given:

$$\widehat{ROA}_i = 2.6 + 4.22 * 1.6 - 1.32 * 1.6^2 = 6.03\%$$

4.0 Supervised Learning part 2: Machine Learning Techniques.

Learning Objectives:

Machine Learning techniques for classification and prediction problems, such as decision trees, k nearest neighbor and Support vector Machines. An overview of Neural networks.

Firstly, differentiate between the two types of decision tree and illustrate how each is constructed and interpreted.

Explain how pruning and ensemble techniques can be used to enhance the performance of decision trees.

Apply the K- nearest neighbor method for Classifications

Illustrate how support vector machines are used to clarify the data

Describe how neural networks are constructed and discuss associated challenges.

Discuss advanced neural networks structures.

Describe how autoencoders are used for dimensionality reduction and differentiate between autoencoders and PCA.

Artificial Intelligence Risk Certificate

4.1 Decision Trees

Are supervised machine learning techniques that examine input features sequentially. At each node is a question, which branches the observation into another node or a leaf (terminal node).

Although these ones are particularly popular for classification problems, they can also be employed to estimate the value of a continuous variable and so are sometimes known as classification and regression trees (CARTs).

CARTs are popular due to their interpretability, white box model, in contrast to other machine learning techniques such as neural networks, a black box model.

They tend to perform less well than other, black box, more sophisticated machine learning models.

To improve their performance, trees are often combined using ensemble techniques such as random forests, bagging and boosting.

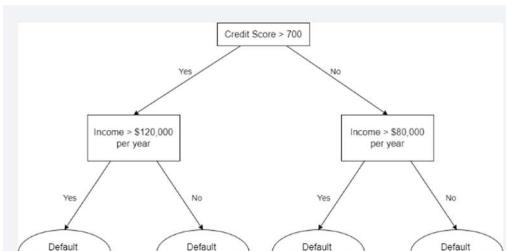


Figure 4.1 Illustration of a simple decision tree for assessing creditworthiness.

4.1.1 Regression Trees

The goal is to split the feature space into regions such that we minimize the residual sum of squares (RSS), given by:

$$RSS = \sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

Where y_i is an observation in the training set, \hat{y}_{R_j} is the average outcome of the observations in region j, and J is the total number of regions.

Unfortunately, it is computationally infeasible to check all possible partitions of the feature space to find the one that minimizes RSS.

Therefore, we employ a top-down recursive binary splitting search. In this approach, we start with all the observations in one region and search for the split that produces the maximum reduction of the RSS. Then, for each of the two regions obtained in this way, we look for a further best split, and we proceed recursively until a given stopping criterion is reached.

For instance, suppose that we want to predict house price, based on age of the house and distance to the closest metro station.

A regression tree is a set of rules that tells us how we can optimally segment the training sample into regions of the feature space. If we estimate a tree for this dataset, we find that the first step is to split the sample between

Artificial Intelligence Risk Certificate

the houses that are less than 826.83 meters from the closest metro station.

Then we split this subsample between those that are smaller than 11.7 years, or greater than.

Region R1 from the below figure, contains all the houses that are below 11.7 years and closer than 826.83 m from the metro.

The prediction of the price per unit area for the houses in R1 is 52.25, which is the average price of houses belonging to that region of the feature space.

We continue to split the space into non overlapping regions until further splits fail to improve the prediction accuracy or some stopping criterion is reached.

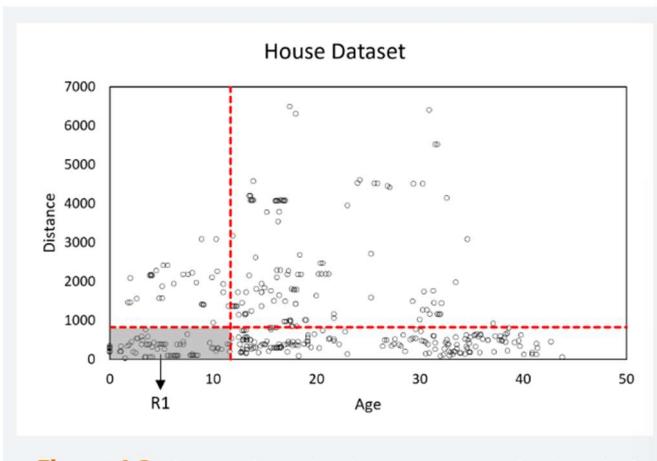


Figure 4.2 Distance from the closest metro station (vertical axis) and age (horizontal axis) for a sample of 414 houses.

Artificial Intelligence Risk Certificate

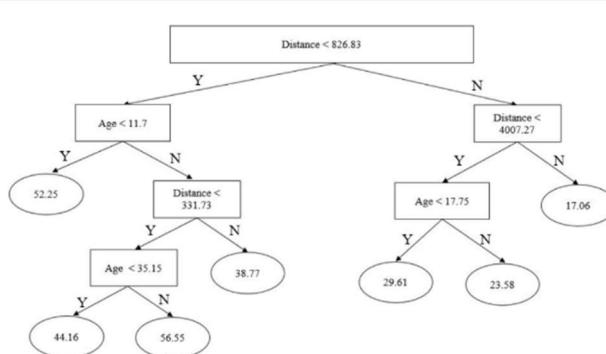


Figure 4.3 Completed regression tree for the prediction of house price per unit of area. The squares denote decision nodes and the circles denote leaf nodes.

4.1.2 Classification Trees

These ones follow a very similar logic to regression trees except that the outcome variable is categorical.

The objective is to split the data into groups that are as pure as possible, i.e., they contain the largest proportion of one class as possible. However, in classification problems, the RSS cannot be used as a criterion to determine the splits and we need to find a measure of purity. These ones are considered to be the Entropy and Gini Coefficient.

Entropy is a measure of disorder in a system.

$$\text{Entropy} = - \sum_{j=1}^J p_j \log_2(p_j)$$

Artificial Intelligence Risk Certificate

Where J is the total number of possible outcomes, and p_j is the probability of the j^{th} outcome for $j=1,\dots,J$. Note that the formula includes the logarithm to base two rather than the LN.

The Gini coefficient is a measure of the impurity of a node and can be calculated as:

$$Gini = 1 - \sum_{j=1}^J p_j^2$$

A small value of the Gini index indicates that a node mostly contains instances from the same class. Gini and entropy usually lead to very similar decision trees.

4.1.3 Classification Trees Example

Suppose that a risk manager at an equity fund is concerned that firms held within portfolio will stop paying dividends next year and so whishes to build a model to predict whether a firm i will pay ($y_i=1$) or will not pay a dividend ($y_i=0$).

A model to classify this output is based on the variables from the table below. It's a mixed sample of 20 non paying/paying dividend firms. For the outcome and binary values, the value 1 means a positive outcome.

The % of retail investors is a continuous variable that could take any real value from 0 to 100%. In a decision tree, we need to select a threshold value for each

Artificial Intelligence Risk Certificate

continuous variable that maximizes the information gain at a particular node.

Note that the optimal threshold will vary depending on the node at which splits occur.

Ideally, a particular question will provide a perfect split between categories, i.e. each terminal node will be a pure set. For instance, if it had been the case that no technology stocks paid a dividend, this would be highly beneficial information and the node containing technology stocks would be pure, only containing nonpaying dividend stocks.

On the other hand, the worst possible scenario would be exactly half of the tech stocks paid dividend, which would make this variable not as useful to be a discriminant factor.

Table 4.1 Data for dividend payment decision tree example

Data point	Dividend	Earnings_drop	Large_cap	Retail_investor	Tech
1	1	0	1	40	1
2	1	1	1	30	0
3	1	1	1	20	0
4	0	0	0	80	1
5	1	0	1	20	0
6	0	1	0	30	1
7	0	1	0	40	0
8	1	0	1	60	0
9	1	1	1	20	1
10	0	1	1	40	0
11	0	0	0	20	1
12	0	0	1	70	0
13	1	1	0	30	1
14	1	0	1	70	0
15	1	0	1	50	1
16	1	0	1	60	1
17	1	1	1	30	0
18	0	1	0	30	1
19	0	0	0	40	0
20	1	1	1	50	0

Artificial Intelligence Risk Certificate

Looking at the output variable, 12 out of 20 firms in the sample paid a dividend.

Although it is possible to construct the tree using entropy, we will use the Gini coefficient as the calculations are slightly simpler.

We measure the Gini coefficient before knowing anything about the features/independent variables.

$$Gini = 1 - \sum_{j=1}^J p_{j^2}$$

$$Gini = 1 - \left(\left(\frac{8}{20} \right)^2 + \left(\frac{12}{20} \right)^2 \right) = 0.480$$

This provides a base level with which we can compare the fall in the Gini Coefficient, which represents an information gain.

The first step is to select the feature that will go at the root node. This choice is made by selecting the one that would cause the Gini to drop the most, Large_Cap.

First, examining the earning drops variable, among firms with an earnings drop =1, six paid dividends and four did not.

$$Gini = 1 - \left(\left(\frac{6}{10} \right)^2 + \left(\frac{4}{10} \right)^2 \right) = 0.480$$

Similarly, among firms with no earnings drop, six paid dividend and four did not, having the same Gini

Artificial Intelligence Risk Certificate

Coefficient. We next calculate the average Gini coefficient for splitting according to this feature, which is calculated from weighting the coefficients for firms with earnings drop and for those without earnings drop according to the proportion of firms in each of those two categories.

$$Weighted\ Gini = \frac{10}{20} * 0.480 + \frac{10}{20} * 0.480 = 0.480$$

We can calculate the information gained by:

$$Information\ Gained = 0.480 - 0.480 = 0$$

This adds zero value, since the number of firms paying a dividend, both for firms that experienced an earnings drop and those that did not.

Repeat the process for the other 3 features.

Large_Cap feature is 1

$$Gini = 1 - \left(\left(\frac{11}{13} \right)^2 + \left(\frac{2}{13} \right)^2 \right) = 0.260$$

Large_Cap feature is 0

$$Gini = 1 - \left(\left(\frac{1}{7} \right)^2 + \left(\frac{6}{7} \right)^2 \right) = 0.245$$

$$Weighted\ Gini = \frac{13}{20} * 0.260 + \frac{7}{20} * 0.245 = 0.255$$

$$Information\ Gained = 0.480 - 0.255 = 0.255$$

Artificial Intelligence Risk Certificate

Repeating the above for tech, information gained would be 0.003.

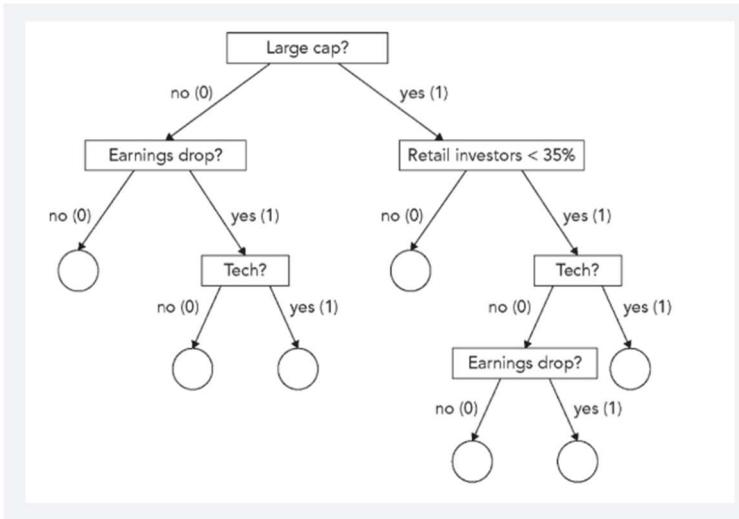
For the % of retail investors, because the variable is continuous, it is necessary to use an iterative procedure to determine the threshold that maximizes the information gain. It turns out that the information gained is less than 0.255 so Large_Cap is the best choice for root node. Once this is done, the three branches out separately for the large_cap firms (13 firms) and for those that are not (7 firms).

At subsequent nodes, features are chosen in a similar way to minimize Gini, and in this way to maximize the Information gained. Threes do not need to be simetrical.

The three is completed when either a leaf is reached that is a pure set or all the features have already been used so that the data cannot be split further.

Creating a perfect classification tree is impossible in this example, so although some branches end up in a pure set others not.

Artificial Intelligence Risk Certificate



4.1.4 Pruning

Small trees offer several advantages over large trees:

Interpretability, fewer irrelevant features, and especially, avoidance of overfitting. As well as employing a separate testing sub sample, overfitting can be prevented by using stopping rules specified a priori, noted as pre or online pruning, pr by pruning the tree after it has been grown (post pruning).

A simple example of stopping rule is when a x number of breaches is reached, no further splitting is allowed. Another example is the termination of the splitting of a node if the number of observations under that node is smaller than a certain number.

Artificial Intelligence Risk Certificate

Whereas Pre pruning prevents a tree from growing to much, post pruning consists of growing the tree fully and the identifying weak links' ex post. In other words, it consists of replacing some subtrees with leaves whose label is the class of most of the instances that reach the subtree in the original classifier.

There are several pruning algorithms that can be distinguished between top-down and bottom-up approaches depending on whether they start at the leaves or at the root of the tree.

One of the simplest forms of pruning is reduced error pruning, which is a bottom-up approach. Starting ate the bottom, this algorithm replaces a node with its most popular class at any time that the resulting pruned tree does not perform worse than the original tree in the validation sample.

Another method in used to prune regression trees is cost complexity pruning. It consists of adding a penalty term to the RSS such that a trade off between accuracy over the training sample and the number of terminal nodes is established. The extend of the trade-off is determined by a tuning parameter alpha, which is chosen with cross validation.

4.1.5 Ensemble Techniques

Trees tend to conduct worse performances than other qualifiers. A way to improve their performance is to ensembled a bunch of trees into a metamodel.

This has 2 objectives, being the first that through the “wisdom of crowds” and a result somewhat like the law of large numbers, model fit can be improved by making many predictions and averaging them. Second, the techniques can build in protection against overfitting.

Bootstrap Aggregation

Also know as bagging, involves bootstrapping from among the training sample to create multiple decision trees. The resulting predictions or classifications are aggregated to construct a new prediction or classification. A basic bagging Algorithm involves the following:

Firstly, sample a subset of the complete training set.

Secondly construct a decision tree on the usual fashion.

Thirdly, repeat steps 1 and 2 many times, sampling with replacement, so that an observation in one subsample can also be in another subsample.

Fourthly, if the problem is a regression, average across forecast of the several trees to obtain the final prediction. If the problem is a classification, record the class predicted by each of the trees and take a majority

Artificial Intelligence Risk Certificate

vote: the class predicted by most the threes is the overall prediction.

Because the data are sampled with replacement, some observations will not appear at all. The observations that were not selected (called out of bag data) will not have been used for estimation in that replication can be used to evaluate model performance.

Pasting is an approach identical to bagging, except the sampling takes place without replacement, so that each data point is only used once.

Random Forests

Aggregating several forecasts work particularly well when the different learners exhibit low correlation. Random forests provide an improvement over bagging by reducing the correlation across trees.

To achieve this, each time that a tree is split, only a random subset of all the features is considered. Although it may appear counterintuitive to purposefully exclude some features as each tree taken alone may result in a suboptimal result, it is a very strong predictor.

The logic is that, if, for instance, a feature is a very strong predictor whereas the rest only have modest predictive power, all the resulting trees will have this feature at the top and they are likely to yield very similar forecast. In contrast, by forcing some trees to deliberately ignore

Artificial Intelligence Risk Certificate

this strong predictor, the other features are given a chance, and the resulting forecasts are less correlated.

Boosting

Like Bagging, Boosting entail combining the forecasts from many decision trees. However, while in bagging each tree is grown independently from the others, in boosting each tree is grown while exploiting the information from the prediction errors of the previously grown trees.

The 2 main varieties of boosting are gradient boosting and adaptive boosting.

Gradient Boosting constructs a new model on the residuals of the previous one, which then become the target, i.e. the labels in the training set are replaced with the residuals from the previous iteration.

AdaBoost, involves training a model with equal weights on all observations and then, sequentially, increasing the weight or misclassified outputs to incentivize the classifier to focus more on those cases.

4.2 K Nearest Neighbors

Is a simple, intuitive, supervised machine learning model that can be used for either Classification or prediction problems.

To predict the outcome or class, for an observation not in the training set, we search for the K observation in the training set that are closest to it using either Euclidean or Manhattan distance.

Our prediction is the mean of the nearest neighbor's outcomes. If the problem is a classification one, the instance to be classified is assigned to the class which more neighbors belong to (majority voting).

KNN is sometimes termed as a Lazy learner because it does not learn the relationship in the dataset in the way that other approaches do. KNN does not build a model, instead, every time KNN encounters a new instance, it compares it to all the existing instances to make a prediction.

The steps to implement KNN are as follows:

Firstly, select a value of K and a distance Measure.

Secondly, among the points in the training sample, identify the K points in feature space that are closest to the point in feature space for which a prediction is to be made according to the chosen measure.

Thirdly, if it is a prediction problem, compute the mean of the outcomes for the K neighbors that have been

Artificial Intelligence Risk Certificate

identified, and if it is a classification problem, assign the instance to the class to which most of the nearest neighbors belong to.

On the example below, the Balance and Income were already standardized, Euclidean distance is the measure used and 1 means defaulted an 0 not defaulted.

Table 4.2 Balance, income, and a dummy variable equal to one if they default for a notional sample of borrowers

N	Default	Balance	Income	Euclidean Distance
1	0	-0.55	-0.75	1.46
2	0	-0.23	0.19	1.38
3	0	-0.76	0.53	2.01
4	0	-0.08	-0.73	0.99
5	0	-0.72	1.32	2.52
6	0	-0.79	2	3.11
7	0	-0.8	-0.34	1.72
8	1	0.46	-0.71	0.45
9	1	1.95	-0.96	1.11
10	1	1.51	-0.54	0.61

Assume that we must classify a new instance, which is a borrower with a standardized balance of 0.9 and a standardized income of -0.61.

Suppose that we select K equal to four and we use the Euclidean distance to find the nearest neighbors.

Artificial Intelligence Risk Certificate

We should compute the distance from the unclassified with all data points in the training data. For the first we would have:

$$d_E = \sqrt{(-0.55 - 0.9)^2 + (-0.75 + 0.61)^2} = 1.46$$

Observations 4,8,9,10 are the closest ones. Therefore, the new instance would be considered as a Default one.

A crucial choice is the value of K. Small values of K tend to overfit the data whereas large values of K may underfit. A common choice is to set k approximately equal to the square root of N.

Another approach is to use cross-validation to tune K and choose the value that minimizes the error over the validation sample.

KNN is simple and yet tends to yield quite accurate forecasts.

However, its main disadvantage is that it is computationally intensive as the distance between one instance and all the others must be computed before KNN can identify the nearest neighbors.

Another drawback of this method is that it performs poorly when there are a few irrelevant or noisy features, as those can drive similar distances apart in feature space.

4.3 Support Vector Machines

SVM are a class of supervised machine learning models that are particularly well suited to classification problems when there are a large number of features.

To understand how they work, we will start with a sample of linearly separable data points that belong to one or two classes (labelled as -1 or +1). If there were only 2 features, this can be represented on the cartesian plane.

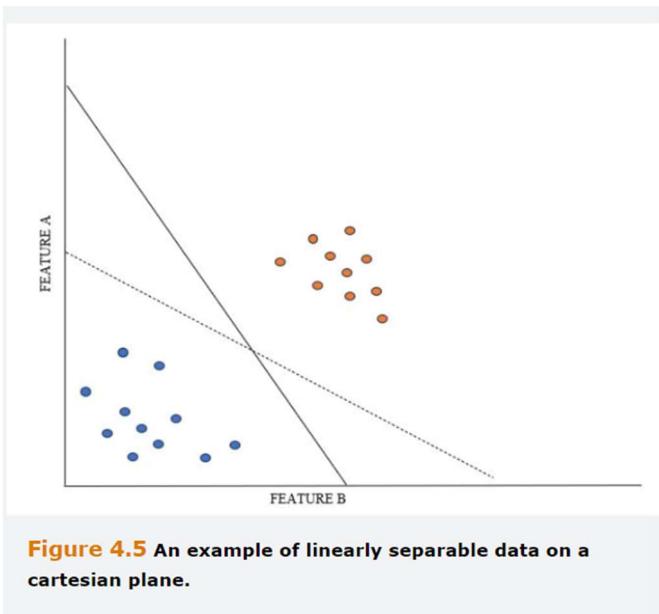


Figure 4.5 An example of linearly separable data on a cartesian plane.

Our main goal is to identify the position of a line that would separate the two groups, the classification boundary, enabling us to predict for an additional data

Artificial Intelligence Risk Certificate

point not in the sample whether the outcome should be -1 or +1. From the above picture, the blue and orange points can be perfectly separated either by using the dotted or solid line.

More generally, there is an infinite number of linear boundaries that perfectly classify the data.

Therefore, we need a metric that helps us to identify which of the boundaries is the most appropriate.

SVM uses a metric called margin. Broadly, the margin is the sum of the distances between the classification boundary and the closest instance in the training data for each of the two classes.

Given a classification boundary it is possible to construct two lines that are parallel to it and that touch the training data of opposite classes, and that have no points between them.

The training data points that are on those parallel lines are called the Support vectors, and the distance between the lines is the margin. The optimal classification boundary, a.k.a. the maximum margin classifier, is such that it is equidistant from each support vector and the margin is maximized.

Artificial Intelligence Risk Certificate

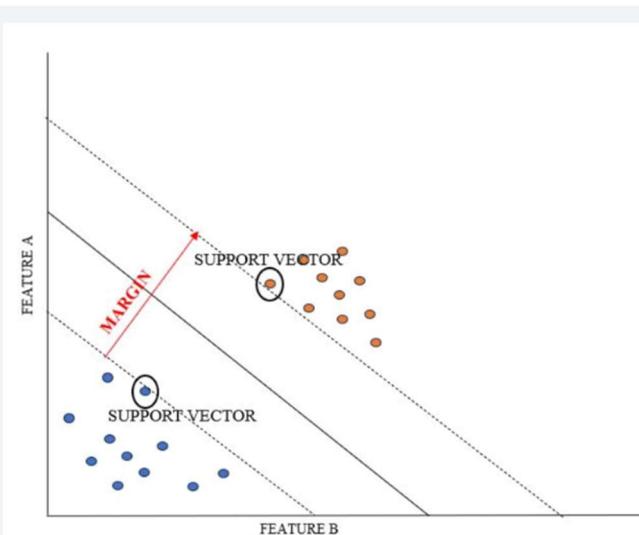


Figure 4.6 A graphical illustration of SVMs.

The maximum classifier creates a decision value $D(x)$ that classifies a new instance such that if $(Dx)>0$ we would predict the instance to belong to class 1, otherwise , we would predict the instance to belong to class two.

Artificial Intelligence Risk Certificate

4.3.1 Support Vector Machines Example

A Bank to grant or not to grant loans.

Table 4.3 Data for SVM car loan example

Applicant number	Monthly income (USD 000s)	Total savings (USD 000s)	Loan granted? (yes = +1; no = -1)
1	2.5	5.0	-1
2	1.8	0.5	-1
3	4.1	1.6	-1
4	0.8	2.0	-1
5	6.2	4.0	-1
6	3.8	6.2	-1
7	2.1	9.0	+1
8	4.6	10.0	+1
9	1.8	13.0	+1
10	5.2	8.0	+1
11	10.5	3.0	+1
12	7.4	8.5	+1

Loans granted are coded as +1 and loans not granted as -1.

The solution of the optimization problem leads to estimation of $w_0 = -12.24$, $w_1 = 0.90$ and $w_2 = 1.26$.

Therefore, the maximum margin classifier is :

$$-12.24 + 0.90(\text{Monthly Income}) + 1.26(\text{total savings}) = 0$$

And the margin constraints are:

$$-12.24 + 0.90(\text{Monthly Income}) + 1.26(\text{total savings}) = +1$$

$$-12.24 + 0.90(\text{Monthly Income}) + 1.26(\text{total savings}) = -1$$

Artificial Intelligence Risk Certificate

If we had to classify a new borrower with Monthly income equal to 5.7 and total savings 3.5, we would obtain:

$$D(x) = -12.24 + 0.90(5.7) + 1.26(3.5) = -2.7$$

As this value is below 0, we would classify the loan as not granted, labeled -1.

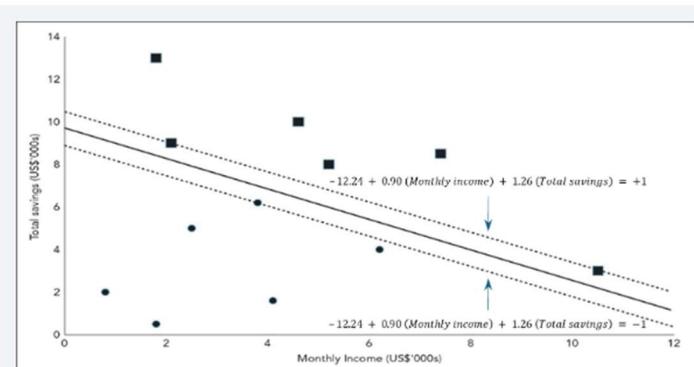


Figure 4.7 Support Vector Machine Illustration for Car Loan Decisions. Squares represent loans granted; circles represent loans not granted for two features (monthly savings and incomes).

4.3.2 Support Vector Machines Extensions

If the Model contemplated more than 2 features, instead of having a line in the center with the biggest margin, we would have a Hyperplane with several dimensions, one less than the number of features.

The two classes had a clear degree of separation on the example provided, but a more likely practical scenario would be that there is some overlap between the two.

In the cases that the data are not linearly separable, the approach described above could not be used and requires some modification.

A more flexible approach would be to use soft margins, which introduces a penalty term into the optimization for incorrect classifications.

4.4 Neural Networks

ANN are a class of machine learning approaches loosely modeled on how the brain performs computation.

By far, the most common type of ANN is a feedforward network with backpropagation, a.k.a. multi-layer perceptron.

The basic unit of a multi-layer perceptron is the neuron, a unit that holds information. The neurons are arranged in layers and a multilayer perceptron consists of several layers of neurons.

Artificial Intelligence Risk Certificate

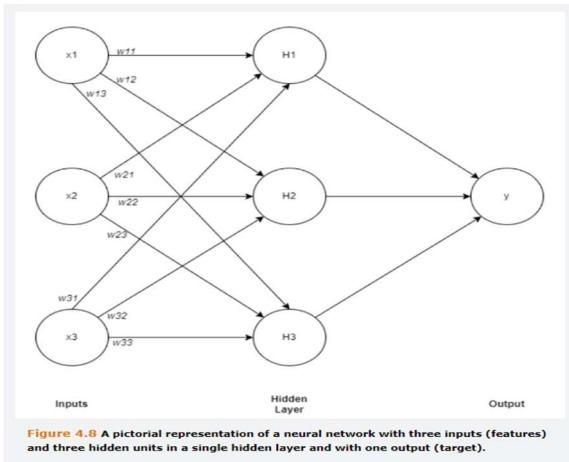
A three-layer perceptron should include an input layer, a hidden layer and an output layer. Each input layer is connected to a hidden layer, and every connection have a weight attributed, $w_{j_i}^{(h)}$, where the notation means that we are connecting neuron j in layer h with the neuron i in the next layer. Each of the inputs is multiplied by the weight associated with each link and passed to the hidden layer, activation function:

$$H_1 = f \left(w_{11}^{(1)}x_1 + w_{21}^{(1)}x_2 + w_{31}^{(1)}x_3 + b_1 \right)$$

$$H_2 = f \left(w_{12}^{(1)}x_1 + w_{22}^{(1)}x_2 + w_{32}^{(1)}x_3 + b_2 \right)$$

$$H_3 = f \left(w_{13}^{(1)}x_1 + w_{23}^{(1)}x_2 + w_{33}^{(1)}x_3 + b_3 \right)$$

The term b is called bias and is often added to the weighted sum of the features, being this one a constant, like the intercept on Linear regression.



Artificial Intelligence Risk Certificate

The activation function introduces nonlinearity into the relationship between the inputs and outputs. Without it, the outputs from the model would merely be linear combinations of the hidden layer(s), which would, in turn, be linear combinations of the inputs. Such structure would be a linear regression, which is antagonist to the idea of neural networks, since the goal of the later is to discover complex nonlinear relationships.

The process of propagating the attributes from the input layer to the output is called feeding forward. A multi-layer perceptron can also contain more than one hidden layer.

Deep Learning refers to machine learning methods utilizing multiple neural network layers to extract the nonlinear relationships embedded in the data being modeled. The deep learning methods are widely used in natural language processing, generative artificial intelligence, image processing and so on.

4.4.1 The choice of Activation Function

The logistic (sigmoid) function we encounter in connection with logistic regression.

$$f(z) = \frac{1}{1 + e^{-z}}$$

This function outputs a value between 0 and 1. Other examples of activation functions are the softmax function, the rectified Linear Unit (ReLU), the leaky ReLU and the Hyperbolic tangent.

The softmax function is a more generalized version of the logistic activation function, that is used in multiclass classification problems.

For $z_i = z_1, z_2, z_k$ where K is the number of classes,

$$f(z_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}$$

Softmax applies the standard exponential function to each element z_i of the vector z of inputs and normalize the result by dividing by the sum of all exponentials. The sum of all $f(z_i)$ is 1. This function exponentially magnifies the importance of the largest member of the input value.

In contrast to sigmoid and Softmax, the ReLU activation function is unconstrained from above. The ReLU activation function takes z as input and returns:

$$f(z) = \max(z, 0)$$

Artificial Intelligence Risk Certificate

When the input is negative, it returns zero and when is positive returns the value itself. The leaky ReLU activation function is a variation of the ReLU function which allows for small negative numbers:

$$f(z) = \begin{cases} 0.01z, & z < 0 \\ z, & z \geq 0 \end{cases}$$

Finally, the Hyperbolic activation function:

$$f(z) = \frac{e^{2z} - 1}{e^{2z} + 1}$$

Has a similar shape to the logistic function but it produces values between -1 and +1 rather than zero and one.

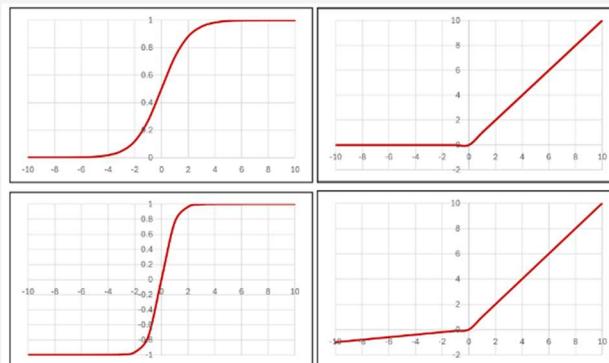


Figure 4.9 The logistic (top left), the ReLU (top right), the hyperbolic tangent (bottom left), and the leaky ReLU (bottom right) activation functions.

Notably, each layer can employ a different activation function, while all the neurons in the same layer apply the same activation function. However, it is common to

Artificial Intelligence Risk Certificate

use the same activation function for all the hidden layers. For the multiple layers perceptron and convolutional neural networks, a common choice is to use the ReLU function for hidden layers.

When a recurrent neural network is employed, popular choices of activation functions are the sigmoid and the hyperbolic. The activation function of the output layer tends to depend on the problem at hand. The sigmoid is a common choice for binary problems, as the softmax is typically employed in multiclass classification problems.

4.4.2 A numerical Example

Example of a basic neural network. There are Three layers, the input the hidden and the output one.

The input one contains four neurons, as much as the number of features., in which $x_1 = 0.6, x_2 = -0.4, x_3 = 0.3, x_4 = -0.2$

The output layer is a binary classification to determine whether each observation belongs to 0 or 1.

In the hidden layers there are only two neurons.

For now, assume the weights as follows:

$$w_{11}^{(1)} = 0.15 \quad w_{12}^{(1)} = -0.2$$

$$w_{21}^{(1)} = -0.4 \quad w_{22}^{(1)} = 0.3$$

$$w_{31}^{(1)} = -0.3 \quad w_{32}^{(1)} = -0.6$$

$$w_{41}^{(1)} = 0.2 \quad w_{42}^{(1)} = -0.2$$

Artificial Intelligence Risk Certificate

The biases are $w_{01}^{(1)} = 0.7$ and $w_{02}^{(1)} = 0.3$ and the activation function for the hidden layer is ReLU.

The value of z_1 is obtained via multiplying the weights by the features and summing their values and the bias.

$$z_1 = w_{11}^{(1)}x_1 + w_{21}^{(1)}x_2 + w_{31}^{(1)}x_3 + b_1 \\ z_1 = 0.15 \cdot 0.6 - 0.4 \cdot (-0.4) - 0.3 \cdot 0.3 + 0.2 \cdot (-0.2) + 0.7 = 0.82$$

Then, the chosen activation function is applied to obtain H_1

$$H_1 = \max(z_1, 0) = \max(0.82, 0) = 0.82$$

Similarly, z_2 is calculated as:

$$z_2 = -0.2 \cdot 0.6 + 0.3 \cdot (-0.4) - 0.6 \cdot 0.3 - 0.2 \cdot (-0.2) + 0.3 = -0.08$$

Which is transformed to:

$$H_2 = \max(z_2, 0) = \max(-0.08, 0) = 0$$

In the output layer there is one neuron, and the activation function is logistic, by assuming that is binary.

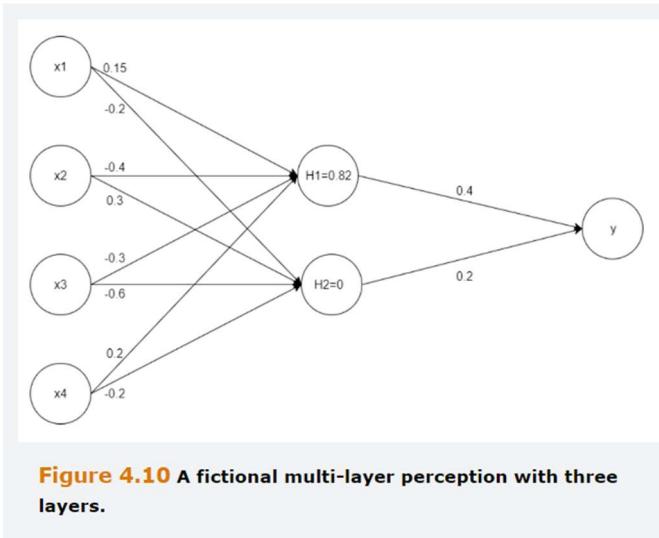
The weights are $w_{11}^{(2)} = 0.4$ and $w_{21}^{(2)} = 0.2$, the bias is $w_0^{(2)} = 0.3$

The value of the Output neuron, our prediction, is obtained by applying the logistic activation function to the weighted sum of H_1 and H_2

$$\hat{y} = \frac{1}{1 + e^{-(0.3 + 0.82 \cdot 0.4 + 0 \cdot 0.2)}} = 0.65$$

Artificial Intelligence Risk Certificate

This can be interpreted as the probability of y is 1, meaning that since 0.65 is higher than 0.5, the observation is assigned to the 1 class.



4.4.3 Backpropagation

The set of weights that link the neurons are parameters. These were given in the previous example, but they must be estimated. The strategy is to find such weights that some measure of classification or prediction error, a loss function, typically residual sum of squares or mean squared error, is minimized.

This procedure is recursive, at the beginning the weights are assigned random values and then a first training example is forward propagated to the networks output.

Artificial Intelligence Risk Certificate

Then the weights are updated using the error between the calculated and actual values of the labels. After updating the weights, a new example is introduced, and the weights are updated again. When the last training example is introduced, one epoch, i.e. iteration is completed.

In general, the successful training of a network involves many epochs and therefore neural networks are usually regarded as a computationally intensive technique.

4.4.4 Architectural Issues

Neural networks are an incredibly flexible tool. It was proved empirically that with the right number of hidden neurons, and under some assumptions, neural networks can approximate any function with arbitrary precision. This result is known as the universal approximation theorem.

However, in practice, neural networks are very prone to overfitting the training data and this problem is worse with large networks.

Usually, one starts with one or two hidden layers and add more layers after assessing the performance of the model. Adding more layers is useful for modeling more complex phenomenon. This leave us with the question on how to determine the current amount of hidden layers and the number of neurons in which layer.

Artificial Intelligence Risk Certificate

One solution could be to experiment with different sizes of network, compute their accuracy over the test sample and pick the network structure that minimizes the error rates.

Consider a network with 100 features, one hidden layer with 100 neurons and an output layer with only one neuron. The estimated weights would be $100*100+1=10,001$. Besides, many epochs are generally needed to reach convergence, which makes this approach highly impractical.

An alternative is to find an appropriate size of network proceeds. We start with a small number of hidden layers. At the end of each epoch, the algorithm computes the value of the loss function over the training set. This value is likely to keep decreasing when the number of epochs increases, until a point is reached where the performance fails to improve. This can happen either because the network lacks the necessary flexibility to make correct predictions, or because a local minimum has been found.

When this is observed, additional layers are added to the network and the training is resumed. If this allows for a reduction in the value of the loss function, the newly added neurons are retained. If not, the smaller model is preferred.

The number of neurons within each layer is dictated by the size of the features set and target. It was common practice to structure the network such that the number of neurons decreased from one layer to the next as the

Artificial Intelligence Risk Certificate

network approaches the output layer. This pyramid style structure has been largely superseded by a more uniform structure with an equal number of neurons in the hidden layers coupled with regularization techniques to ensure that the model is not overfitting.

4.4.5 Overfitting

When the model is large and insufficient training examples have been provided, neural networks tend to learn random artifacts of the training data. This implies that they will fail to generalize well to unseen test instances.

An extreme form of overfitting is memorization, which results in an almost perfect fit to the training data, and which is not uncommon with neural networks.

Overfitting signs are:

Firstly, the same model obtains very different predictions depending on the sample it is trained with.

Secondly, the gap between the prediction error over the training and the test sample is very large.

Apart from avoiding parameter proliferation, there are a few techniques that can be used to reduce overfitting.

Penalty based regularization involves imposing a penalty over the loss function, such as $\lambda \sum_{i=0}^d w_i^2$, where d is the number of neurons.

Artificial Intelligence Risk Certificate

Dropout is an ensemble technique explicitly designed for neural networks. It consists of creating alternative networks by selectively dropping a few neurons each time. The forecast from these networks are then aggregated to create the final prediction.

Early stopping, in which the optimization is stopped before converging to the optimal solution on the training data. A portion of the data is held out and used to determine the optimal stopping point. The training is stopped when the error in the hold out samples starts to rise.

4.4.6 Advanced Neural Network Structures

Convolutional Neural Networks

CNN, are a specialized form of Neural network where the neurons in one layer are only connected to a subset of neurons in the next layer. They are designed to work with inputs that have a grid structure and where adjacent points in the grid exhibit dependencies.

The most obvious application of CNNs is with 2 dimensional images, but they can also be employed for textual, voice or time series data.

CNNs are ideal in such cases because the number of network weights to be trained is drastically reduced, which results in faster training of the model. The most common type of convolutional layer is the 2D or planar convolutional layer.

Artificial Intelligence Risk Certificate

This one applies an $n \times n$ kernel matrix W over a $m \times m$ input grid to obtain a new filtered image that has a smaller size than the original image. This new one is called feature map.

The kernel matrix contains weights that should be learned during the training process, but in the following example, these ones are given for simplicity.

Input is a 4×4 image, the matrix X :

$$X = \begin{pmatrix} 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 3 \\ 2 & 1 & 0 & 1 \\ 0 & 2 & 2 & 1 \end{pmatrix}$$

This can be filtered using a 3×3 Kernel:

$$W = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

The feature map is obtained by sliding the kernel over the image starting from the top left corner to move the kernel through all the positions where it fits entirely within the boundaries of the image.

Each position corresponds to a single cell in the feature map, the value of which is calculated by multiplying together the kernel value and the underlying image for each of the cells in the kernel, and then adding all these numbers together.

In practice, we will start by replacing the area on the top left corner.

Artificial Intelligence Risk Certificate

$$X = \begin{pmatrix} \boxed{1 & 0 & 0} & 2 \\ 0 & 0 & 1 & 3 \\ 2 & 1 & 0 & 1 \\ 0 & 2 & 2 & 1 \end{pmatrix}.$$

The feature map should be:

$$f_{11} = 0 * 1 + 1 * 0 - 1 * 0 + 0 * 0 + 1 * 0 + 1 * 1 + 1 * 2 + 0 * 1 + 1 * 0 = 3$$

Then, we slide to the right:

$$X = \begin{pmatrix} 1 & \boxed{0 & 0 & 2} \\ 0 & \boxed{0 & 1 & 3} \\ 2 & \boxed{1 & 0 & 1} \\ 0 & 2 & 2 & 1 \end{pmatrix}$$

$$f_{12} = 0 * 0 + 1 * 0 - 1 * 2 + 0 * 0 + 1 * 1 + 1 * 3 + 1 * 1 + 0 * 0 + 1 * 1 = 4$$

The down:

$$X = \begin{pmatrix} 1 & 0 & 0 & 2 \\ \boxed{0 & 0 & 1} & 3 \\ 2 & 1 & 0 & 1 \\ 0 & 2 & 2 & 1 \end{pmatrix}$$

$$f_{21} = 0 * 0 + 1 * 0 - 1 * 1 + 0 * 2 + 1 * 1 + 1 * 0 + 1 * 0 + 0 * 2 + 1 * 2 = 2$$

Artificial Intelligence Risk Certificate

It is now intuitive to assume the final position as

$$F = \begin{pmatrix} 3 & 4 \\ 2 & 2 \end{pmatrix}$$

The area in red, blue and green are termed a receptive field. It is the region in the input space that influences a cell in the feature map.

A non-linear layer, where a nonlinear activation function is applied to the feature map, can also be added after the convolutional layer.

It is also common to have a pooling layer. The pooling layer replaces the output of the previous layer at certain locations with summary statistics. For instance, it would be possible to summarize F by taking the average or the maximum of the values in the cells.

CNNs are parsimonious in terms of parameters as the same weights are applied to all receptive fields. Therefore, CNNs are useful to process images, which typically involve millions of pixels.

Recurrent Neural Networks

Differ from a Standard multilayer perceptron as the former models employ a temporal sequence to preserve the order in which the observations occur. In other words, RNN is designed to have some memory. RNNs are often employed in time series applications, and they are at the heart of large language models.

4.5 Autoencoders

Are a class of Artificial Neural Network Models (ANNs) used for unsupervised learning. They are feedforward specifications, but the outputs are the same features as the inputs, and hence there are no labels.

Unlike K-means clustering, autoencoders are primarily used for dimensionality reduction and so are best thought of as non-linear extensions of PCA.

Autoencoders can provide a compact representation of the feature data and are particularly useful for high dimensional systems.

It should be noted that although PCA is commonly discussed in machine learning Contexts, in fact there is no learning involved, because it is merely a decomposition with a unique mathematical solution.

Autoencoders, on the other hand, are trained to learn the relationships present in the Data through model estimation.

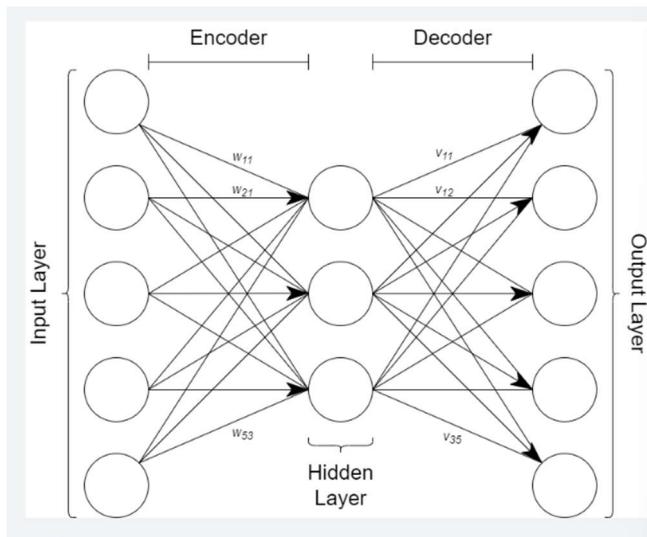
The advantages of auto encoders over PCA is the use of non-linear activation functions which provide the universal approximation property in high dimensional space.

The features are put through the encoder, which is a function, to arrive at the values on the hidden layer. Then, the values in the hidden layer are converted back to the feature values through the Decoder.

Artificial Intelligence Risk Certificate

The optimization objective is to reconstruct the original features as accurately as possible. The weights between the input layer and the hidden layer encode the information from the features, and the weights between the hidden layer and the output layer decode the information.

When the number of neurons on the Hidden Layer is smaller than the number of features, this is called a constricted or bottleneck hidden layer and leads to dimensionality reduction.



The hidden layers are simply calculated as a weighted sum of the inputs, and the outputs (reconstructed features) are a weighted sum of the value on the hidden layers.

Artificial Intelligence Risk Certificate

The values on the hidden layer can be calculated as:

$$H_{1i} = w_{11}^{(1)}x_{1i} + w_{21}^{(1)}x_{i2} + w_{31}^{(1)}x_{3i} + w_{41}^{(1)}x_{4i} + w_{51}^{(1)}x_{5i} + w_1$$

Where w_1 is the bias term.

The values at the hidden layer are the reduced dimension representation of the data, also known as the code.

If we let \hat{x}_{1i} denote the fitted output, i.e. the reconstructed values from the model this would be calculated as:

$$\hat{x}_{1i} = v_1 + v_{11}H_{1i} + v_{21}H_{2i} + v_{31}H_{3i}$$

The weights are chosen by minimizing a loss function, L, akin to the Residual Sum of Squares (RSS) in a linear regression.

$$L = \sum_{j=1}^m \sum_{i=1}^N (x_{1i} - \hat{x}_{1i})^2$$

An alternative would be mean Squared error (MSE)

$$L = \frac{\sum_{j=1}^m \sum_{i=1}^N (x_{1i} - \hat{x}_{1i})^2}{mN}$$

L will be a positive number to reflect that the feature inputs will no be reconstructed precisely after the encoding and decoding processes, but this is the price paid in order to obtain a more parsimonious representation.

To this point, the autoencoder is linear in the weights, and thus it will perform a function comparable to PCA.

Artificial Intelligence Risk Certificate

More specifically, in such a linear model, if there is only one hidden layer with K nodes, if both the encoder and decoder are linear, and if the inputs are suitably normalized, then the encoder hidden nodes will be the first K principal Components.

However it is more common to use a non-linear autoencoder by applying an activation function to the weighted sums in the hidden layers. The activation function, as previously discussed, introduces nonlinearity into the relationship between the inputs and outputs.

Without them, the outputs from the model would merely be linear combinations of the hidden layer(s), which would, in turn, be linear combinations of the inputs.

The new value of the hidden layers would be :

$$H_{1i} = \phi(w_{11}^{(1)}x_{1i} + w_{21}^{(1)}x_{i2} + w_{31}^{(1)}x_{3i} + w_{41}^{(1)}x_{4i} + w_{51}^{(1)}x_{5i} + w_1)$$

Where phi is the activation function.

By capturing these nonlinearities in the data, activation functions can also allow the number of nodes in the hidden layers to be further reduced so that the representation is even more compact than if a purely linear specification was used.

The number of hidden neurons is usually lower than the number of features to allow dimensionality reduction. If the numbers were the same, it would be possible to trivially reconstruct the exact features, but this would be pointless since there will not be any dimensionality reduction.

If the value of hidden Layers is higher than the number of features, we would have a sparse autoencoder. Hence, even though the

Artificial Intelligence Risk Certificate

number of hidden units is large, many of the weights, are set to zero, so that the effective number of weights is much lower and commensurate with a smaller number of hidden units/neurons.

By adding additional hidden layers, we form a deep autoencoder, which has more scope to capture more sophisticated nonlinear patterns between the features. These ones tend to be symmetrical to the center hidden Layer.

We evaluate autoencoders by calculating the Loss of the final fitted model, which is called the reconstruction error.

Table 4.4 Loss Measures (MSE) for PCA and Autoencoders for Modeling the Changes in Treasury Yields

Number of components for PCA	MSE for PCA	Number of hidden units for autoencoder	MSE for autoencoder
1	0.18208	1	0.18270
2	0.04965	2	0.05027
3	0.01708	3	0.01740
4	0.00848	4	0.00854
5	0.00528	5	0.00534
6	0.00295	6	0.00299
7	0.00107	7	0.00113
8	0.00000	8	0.00000

Appendix 4.A Technical details of how SVM are Determined

Considering a two-dimensional case with features x_1 and x_2 . Upper margins constrain is defined as:

$$w_0 + w_1 x_{1i}^{(u+)} + w_2 x_{2i}^{(u+)} = 1$$

With the lower boundary:

$$w_0 + w_1 x_{1j}^{(l-)} + w_2 x_{2j}^{(l-)} = -1$$

Where w are the weights, l- and u+ superscripts denote the data points corresponding to the lower and upper margin respectively, and x_1 and x_2 denote features 1 and 2.

Subtracting the second of these two equations rom the first provides us with an equation for the margin width to be maximized:

$$w_1 \left(x_{1i}^{(u+)} - x_{1j}^{(l-)} \right) + w_2 \left(x_{2i}^{(u+)} - x_{2j}^{(l-)} \right) = 2$$

Note that $(x_{1j}^{(l-)}, x_{2j}^{(l-)})$ and $(x_{1i}^{(u+)}, x_{2i}^{(u+)})$ are two specific instances of features.

We also need to apply two constrains to ensure that all the points lie on or outside of the estimated margins:

$$w_0 + w_1 x_{1i}^+ + w_2 x_{2i}^+ > 1, i = 1, \dots, N_1$$

And:

Artificial Intelligence Risk Certificate

$$w_0 + w_1 x_{1j}^- + w_2 x_{2j}^- \leq -1, j = 1, \dots, N_2$$

The $-$, $+$ superscripts denote the data points corresponding to class 1 and 2, respectively, and N_1 and N_2 the number of observations in each class.

Using the index j to now denote all the observations regardless the class they belong to, we could combine the constraints as:

$$y_i(w_0 + w_1 x_{1j} + w_2 x_{2j}) \geq 1, j = 1, \dots, N$$

Where x_{1j} and x_{2j} combine the positive and negative outcomes, y_i is $+1$ if the observation belongs to class 1 and -1 if the observation belongs to class 2, and $N=N_1+N_2$

By considering the relationship of $(x_{1j}^{(l-)}, x_{2j}^{(l-)})$ and $(x_{1i}^{(u+)}, x_{2i}^{(u+)})$, through w_1 and w_2 one can show that the margin width is given by $2/|w|$, where :

$$|w| = \sqrt{w_1^2 + w_2^2}$$

This is the Euclidean norm of weights. Maximizing the margin width is equivalent to minimizing $|w|$. However, rather than working with $|w|$, it is often easier to minimize $\frac{1}{2}|w|^2$ because its derivative is just w . Therefore, we can write the optimization problem as:

$$\min_{w_0, w_1, w_2} \frac{1}{2}|w|^2 \text{ subject to } y_i(w_0 + w_1 x_{1j} + w_2 x_{2j}) \geq 1, j = 1, \dots, N$$

Artificial Intelligence Risk Certificate

To allow for Overlapping classes that are not linearly separable, we employ a hinge (max) function, which sets the penalty to zero for correct classifications and to distance between the point and the decision boundary for incorrect classifications. We can write the optimization problem as minimizing the function:

$$L(w) = \frac{1}{N} \sum_{j=1}^N \max(1 - y_i(w_0 + w_1 x_{1j} + w_2 x_{2j}), 0) + \frac{\lambda}{2} |w|^2$$

The function max() os zero for all correct classifications, however far the point is from the margin, but will be equal to the distance between the boundary and the point for incorrect classifications.

The regularization here is somewhat like a ridge regression, with the hyperparameter λ controlling the relative weight on the margin width versus incorrect classifications.

Questions and Answers Module 2 Chapter 4 from GARP - Machine Learning Techniques

4.1 What are the main differences between regression and classification decision trees?

The main difference between classification and regression trees is the type of outcome variable. When the outcome variable is continuous, we refer to regression trees and when it is categorical we refer to classification trees.

In both cases, we use a top-down recursive binary split to grow the tree.

However, for regression trees the splits are decided to minimize the RSS, conversely, the splits in classification trees are decided to produce the largest drop either in entropy or the Gini coefficient.

4.2 What are the main differences between bagging, boosting and random forest?

All are ensemble techniques that are based on obtaining a final prediction as the average between the forecast of many trees.

When the problem is classification, a majority vote is generally used to determine the class to which an unseen instance is classified.

Artificial Intelligence Risk Certificate

Bagging relies on the construction of many sub samples by randomly extracting observations from the training sample, with replacement. Trees are fitted on each of the sub samples and the forecast is obtained on the average (or the most popular class) across the predictions of those trees.

Random forests are like bagging, but the correlation among trees is reduced by using a random subset of $p < m$ features at each split of the tree.

Finally, boosting is a sequential procedure where each new tree that is grown uses the information from the residuals of the previously grown trees. Gradient boosting and adaptive boosting.

4.3 In the context of decision trees, what is pruning?

Is a technique to reduce the size of the Tree to avoid overfitting and enhance interpretability. Pre or online pruning are conducted while the tree is growing by imposing a stopping criteria. Post pruning is conducted after the tree has been grown by replacing subtrees with leaves when the substitution does not decrease the predictive accuracy of the learner.

Artificial Intelligence Risk Certificate

4.4 What are the main advantages and disadvantages of decision trees vs other supervised machine learning techniques?

The main advantage of Decision trees are their interpretability and the fact that they resemble the human decision-making process. They are called white box models.

Their Main disadvantage is that they are often less accurate than black box models such as neural networks. To enhance their performance, ensemble techniques are often used, however this costs interpretability, there is a tradeoff between accuracy and interpretability.

4.5 How would K nearest neighbors proceed to classify a new instance given a training sample of 10,000 observations?

The first step involves choosing a measure distance, such as Euclidean or Manhattan and K, the number of nearest neighbors to be considered. For instance, we could set $K = \sqrt{N}$ which would be 100. Then, the distance between the instance to be classified and each of the instances in the training sample are computed and the K nearest neighbors are identified. A majority vote is used to assign the unclassified instance to the class of most of the neighbors.

Artificial Intelligence Risk Certificate

4.6 In the context of SVM, what is the maximum margin classifier

The maximum margin classifier is the optimal decision boundary. It is the line (hyperplane) that is equidistant from the margin that constrains and maximizes the margin.

4.7 In the Context of artificial neural networks, what is an activation function?

An activation function for the neuron, popular choices are the sigmoid, the softmax, the ReLu and the Hyperbolic function.

4.8 Describe some potential ways to address overfitting in a neural network.

Overfitting is a common problem with neural networks, as their great flexibility is also their primary disadvantage. In addition to avoiding parameter proliferation, ways to address overfitting include penalty-based regularization, dropout and early stopping.

The first approach involves imposing a penalty on the loss function. The second relies on the generation of alternative networks by selectively dropping a few neurons.

Artificial Intelligence Risk Certificate

The final forecast is obtained by aggregating the prediction from the various networks.

Finally, early stopping entails stopping the training before the convergence is achieved over the training sample. The stopping point is decided by looking at the error rate over a hold out sample.

4.9 What is the primary purpose of autoencoders

Dimensionality reduction, representing the most important characteristics of a dataset using a smaller number of transformed features.

4.10

A. How do encoders achieve dimensionality reduction?

By constructing a network with fewer nodes in the hidden layer than in the input and output layers.

B. How do autoencoders differ from PCA in the way that they reduce the dimensionality of a dataset?

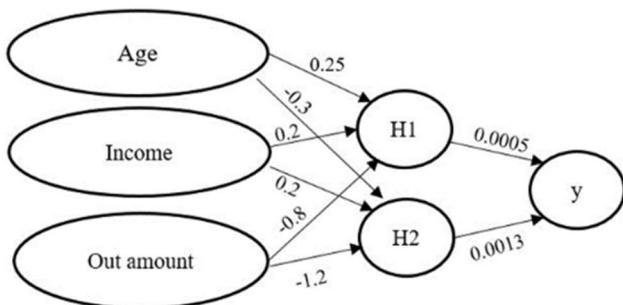
PCA reduces the dimensionality of a dataset by constructing a set of orthogonal (linearly independent) components that are linear combinations of the original features. Although there will be m principal components if there are m features, only the first K of those, which

Artificial Intelligence Risk Certificate

are ordered to explain most of the variation within the features are retained.

Autoencoders use a neural network specification, usually with a non-linear activation function, which will identify a non-linear combination of the original features that is able to closely approximate them.

4.11



[View Answer](#)

A. Describe the Structure of the Neural network

Multi-layer perceptron with 3 layers, an input, a hidden and an output layer. The input layer contains three neurons and the hidden layer two neurons.

Artificial Intelligence Risk Certificate

B. Using the ReLU function at the hidden layer and the logistic function at the output layer, make a prediction for the default (=1) of a credit holder who is 35, has an income of 100,000 and an outstanding amount of 20,000. Biases are equal to zero.

We shall first compute the value assigned to each the two neurons on the hidden layer.

$$z_1 = w_{11}^{(1)}x_1 + w_{21}^{(1)}x_2 + w_{31}^{(1)}x_3 + b_1 z_1$$

$$H_1 = \max(z_1, 0)$$

$$H_1 = \max(35 * 0.25 + 100,000 * 0.2 + 20,000 * -0.8; 0) = 4,008.75$$

$$H_2 = \max(35 * -0.3 + 100,000 * 0.2 + 20,000 * -1.2; 0) = 0$$

Then applying the Sigmoid to the weighted sum of the inputs of the hidden layers

$$F(y_i) = \frac{1}{1 + e^{-y_i}}$$

$$\hat{y} = \frac{1}{1 + e^{-(0.0005 * 4,0008.75 + 0.0013 * 0)}} = 0.88$$

Because the predicted probability is high the risk that credit card holder would default is quite high.

C. Why is the Sigmoid adequate for the output layer?

Because is a classification problem. Values between 0 and 1, being 1 default and 0 not to default, and the value we obtain is the probability of such event to occur.

5. Semi supervised Learning

Learning Objectives:

Explain how semi supervised learning differs from unsupervised and supervised learning.

Discuss the assumptions required for effective semi supervised learning.

Compare and contrast self-training and co training methos of semi supervised learning.

5.1 Introduction to Semi Supervised Learning

When we have unlabeled data there are two possible approaches.

Firstly, either to remove the labels and treat the whole dataset as if is unlabeled, using appropriate techniques for unsupervised learning.

Secondly, drop the observations that are not labeled, and the use supervised learning on the remainder.

Neither of these options would be ideally because they both imply throwing away information. A third option could be to try to identify labels for the unlabeled data, but the labels could either be impossible to obtain or very costly to collect.

For instance, imagine that a bank asks its customers to write brief comments on the service they received. The

Artificial Intelligence Risk Certificate

bank would like to classify either as happy or unhappy and then correlate these classes with costumer specific info. This body of data would be unlabeled , because the comments would not have been classified yet, and creating labels would require a human to read through them individually and make a classification.

Doing this process could be infeasible if the data set is large and in other scenarios might require subject matter expertise. But in either case, should be possible to determine appropriate labels for a small subset of costumer, 30 for example.

This would create an hybrid data set with some labeled and some unlabeled observations. We could still employ one of the two approaches described above, but there is a third category of structure to consider, semi supervised learning also known as Weak supervision.

5.2.1 Semi Supervised Learning Assumptions

Invariably, weak supervision is used in the context of classification rather than prediction scenarios. The technique makes use of parallels between classification and clustering and for it to work well several assumptions about the nature of the data need to hold.

Firstly, the **Clustering assumption**, i.e. the unlabeled data fall naturally into separable clusters (locally dense regions in feature space).

Artificial Intelligence Risk Certificate

Secondly, the **Smoothness assumption**, or continuity assumption is assuming there is a smooth and continuous boundary separating the classes that can be used for deciding the classes of unlabeled instances.

Thirdly, the **Manifold Assumption**, i.e. the observed data point in the High-dimensional feature space are often concentrated along lower dimensional substructures that are topological manifolds. A topological manifold is a topological space that locally resembles the Euclidean Space \mathbb{R}^n . A way to understand the manifold assumption is to think about a sphere, 3D object, where all datapoints are concentrated on the surface (a two-dimensional object). The surface of a sphere is a two-dimensional manifold embedded in a three-dimensional space. The manifold assumption states that the input space is composed of many manifolds on which all the datapoints lie, and all the datapoints in the same manifold belong to the same class.

These assumptions imply that the clusters in the unlabeled data, map naturally onto the classifications on the labeled data.

For instance, a bank wants to develop a default model and had a dataset with two subjects, one with labels, i.e., if defaulted or not and a second subset without any labels where there is only info about the mortgage, but the bank does not know whether those customers defaulted or not.

Artificial Intelligence Risk Certificate

The bank might build a classification model on the labeled part of the dataset and a clustering model for the unlabeled part of the dataset.

Weak supervision would work best if the clusters that formed in the unlabeled data naturally captured the same characteristics as the classification of the labeled data. For instance, two clusters, one with low income, high borrowing low collateral and the other the other way around.

If the different clusters and classifications separate the features in very different ways, the additional benefit from employing the unlabeled data to bolster the labeled data is much diminished.

This link between classification and clustering provides a foundation for how semi supervised learning works, specifically the assumption is that if a set of instances are clustered closely together, they would likely share the same label if they were labeled. On the other hand, points far apart in feature space are less similar and therefore less likely to share a label.

5.2.2 Semi Supervised Learning Techniques

There are two main techniques.

Transductive methods, which do not aim to build a generalise model and are therefore sometimes considered to arise from a “close world view”. In this case, because there is no model, the objective is solely to identify labels for the unlabeled data already observed. All instances need to be specified at the time of conducting the analysis, and no new instances can be incorporated into the study and classified at a later stage, so there is no separate test data.

One transductive technique is label propagation, which is a graphical technique that assigns labels to unlabeled data based on how close they are to labeled data points using a metric such as the Euclidean Distance.

Inductive methods, on the other hand, involve building a model that links the features to the labels, and that can be applied to other instances. Common inductive methos include self-training and Co-training.

5.2.3 Self-Training

The most popular due to its intuitiveness and simplicity.

It is sometimes referred to as a heuristic technique, because it employs unlabeled data from a supervised perspective, using models and methos for the latter, rather than using both labeled and unlabeled data together in learning. Self-Training is included in the Wrapper family.

Firstly, generate a classification model using any preferred technique (KNN, Logistic, etc) applied to the labeled part of the data.

Secondly, apply the model generated in the first step to all the unlabeled data and generate predicted labels for each instance in the unlabeled part of the data.

Thirdly, select the single instance for which the model's predicted label has the highest probability of being correct based on the probabilities output from logistic regression, neural networks...

Fourthly, apply the predicted label to the instance selected at stage 3 and shift that datapoint from the unlabeled to the labeled portion of the data set.

Fifthly, return to stage 1 with the labeled set having now been enlarged by one observation and the unlabeled set reduced by one.

Artificial Intelligence Risk Certificate

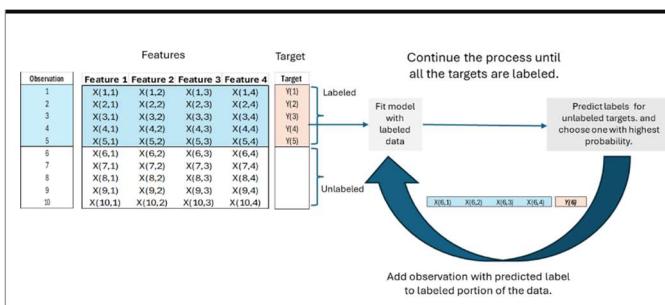
Sixthly, repeat stages 1 to 5 until all unlabeled data points have been labeled, the stop and that would be the final classification model.

These labels assigned are a.k.a **pseudo-label**.

The disadvantages of these are a couple.

Firstly, it is very **computationally intensive** because the model is retained as many times as there are instances in the unlabeled data. If computational resources are constrained, this problem can be mitigated by selecting the best predicted k observations at stage 3 and shifting all k observations, along with their predicted labels, in stage 4. For instance, if $K=10$, this will reduce by tenfold the number of rounds of training required.

Secondly, retraining the model after the addition of each individual datapoint can result in severe overfitting. Overfitting can be guarded against by a process known as co-training.



Artificial Intelligence Risk Certificate

A self-training example, in which we have 4 unlabeled data points, where the borrowers default status is unknown. In this example we employ logistic regression for classifying the unlabeled borrowers into two classes, default or no default.

Table 5.1 Initial Data for Borrowers (Standardized)

Borrower #	Default	Balance	Income
1	0	-0.55	-0.75
2	0	-0.23	0.19
3	0	-0.76	0.53
4	0	-0.08	-0.73
5	0	-0.72	1.32
6	0	-0.79	2
7	0	-0.8	-0.34
8	1	0.46	-0.71
9	1	1.95	-0.96
10	1	1.51	-0.54
11	?	-1.95	0.83
12	?	-0.03	0.05
13	?	1.99	-1.04
14	?	-0.01	-0.02

Step 1 we fit the logistic model using only the labeled data points, i.e. 1 to 10.

$$P(i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \text{Feature } A_i + \beta_2 \text{Feature } B_i)}}$$

Artificial Intelligence Risk Certificate

Step 2, utilizing this trained model, we predict outcomes, the probability of default, and the probability of non-default for the unlabeled data.

Based on the results, the 13th borrower has the clearest split between the two categories, i.e. the biggest difference in the predicted probabilities.

Table 5.2 Predicted Probabilities for Unlabeled observations after First regression

Borrower #	Probability of non-default	Probability of default	Predicted outcome
11	1.0000	8.0819×10^{-29}	0
12	0.9934	6.6148×10^{-3}	0
13	0.0000	1.0000	1
14	0.9896	0.0104	0

Step 3, we include the 13th observation in the labeled sample and label it as a default. Subsequently, we repeat the process using the updated data to fit the logistic regression model and predict the remaining unlabeled data.

This time the 11th borrower has the biggest difference in the predicted probabilities. Hence we include this observation into the labeled sample and label it as non-default.

We run the process again until obtain all labels.

5.2.4 Co-Training

Can be applied when we have two different views of an example. Can utilize both views to build two classifiers that teach each other on unlabeled data.

Let us divide the feature set x into two disjoint subsets X_A and X_B representing two different views of the dataset. Co-training assumes that either X_A or X_B are individually sufficient to learn if we have enough labeled data and thus classifiers can be built for each of them.

Firstly, split feature set x into disjoint subsets X_A and X_B corresponding to two different views, A and B, both for the labeled and unlabeled data.

Secondly, generate classification models (model A and model B) for the two feature sets of the labeled data.

Thirdly, apply the models generated in step 2 to the two unlabeled subsets of data and generate predicted labels for each instance in the unlabeled subsets.

Fourthly, select the predicted observation from the unlabeled subset for each model with the highest probability score.

Fifthly, assign the predicted labels to the instances selected on point 4 and shift those from the unlabeled to the labeled sets. The key difference with self-training is that the data points move to the labeled dataset of the other feature set. So the best predicted

Artificial Intelligence Risk Certificate

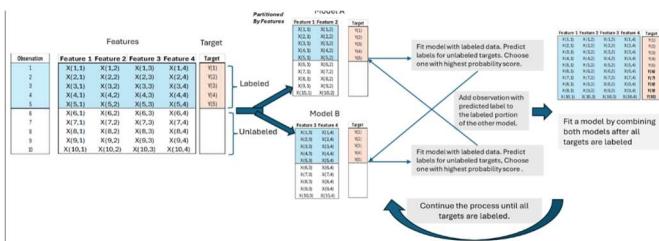
instance from unlabeled subset A moves to the labeled subset B and vice-versa.

Sixthly, return to stage 2 with the labeled sets having now been enlarged by one observation each, and the unlabeled sets reduce by one each.

Seventhly, repeat step 2 and 6 until all unlabeled data have been labeled, then stop and those would be the final classification models, one for each of the two disjoint sets.

Eighthly, estimate a single supervised model that reunites the two subsets A and B now that all instances have been labeled.

Because the Co-training use different subsets of features to build two different models to augment the training set, It reduces the risk of overfitting. Co-training is a.k.a disagreement-based method, because it exploits differences in the predictions based on the two subsets of features to improve the training classifications of both as they lean from one other.



Artificial Intelligence Risk Certificate

Same example as in Self-training.

Initially, we separate the dataset into labelled and unlabeled samples. Subsequently we split the features into two distinct groups, A and B.

Table 5.6 Data split into two groups based on features

Group A			Group B		
Borrower #	Default	Feature A balance	Borrower #	Default	Feature B income
1	0	-0.55	1	0	-0.75
2	0	-0.23	2	0	0.19
3	0	-0.76	3	0	0.53
4	0	-0.08	4	0	-0.73
5	0	-0.72	5	0	1.32
6	0	-0.79	6	0	2
7	0	-0.8	7	0	-0.34
8	1	0.46	8	1	-0.71
9	1	1.95	9	1	-0.96
10	1	1.51	10	1	-0.54
11	?	-1.95	11	?	0.83
12	?	-0.03	12	?	0.05
13	?	1.99	13	?	-1.04
14	?	-0.01	14	?	-0.02

Two Sigmoid modes A and B are trained independently using labeled data, focusing on each future group separately.

Artificial Intelligence Risk Certificate

$$P(i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \text{Feature } A_i)}}$$

$$P(i) = \frac{1}{1 + e^{-(\beta_0 + \beta_2 \text{Feature } B_i)}}$$

These models are then employed to predict the labels of the unlabeled observations.

Table 5.7 Predicted Probabilities for Unlabeled Observations after First Regression

Group A

Borrower #	Probability of non-default	Probability of default	Predicted outcome
11	1.0000	1.3211×10^{-30}	0
12	0.9991	9.1305×10^{-4}	0
13	0.0000	1.0000	1
14	0.9983	0.0174	0

Group B

Borrower #	Probability of Non-default	Probability of default	Predicted outcome
11	0.9989	1.1419×10^{-3}	0
12	0.9635	0.0365	0
13	0.1655	0.8346	1
14	0.9507	0.0493	0

Based on the prediction from the Sigmoids, the 13th borrower has the biggest difference in the probabilities in Model A, whereas the 11th borrower has the biggest difference in the probabilities of model B. Consequently, the 13th observation is moved to group B and labeled has



Category
5.2!

Artificial Intelligence Risk Certificate

having defaulted, whereas the 11th observation is moved to group A with no default label.

Utilizing the updated labeled data, we repeat the process of fitting the sigmoid models and predicting the remaining unlabeled observations.

Table 5.8 Predicted Probabilities for Unlabeled Observations after Second Regression

Group A

Borrower #	Probability of non-default	Probability of default	Predicted outcome
12	0.9993	7.3523×10^{-3}	0
14	0.9986	0.0143	0

Group B

Borrower #	Probability of non-default	Probability of default	Predicted outcome
12	0.8944	0.1057	0
14	0.8784	0.1216	0

Since borrower 12th has the greatest difference in both models, it will be moved to both A and B group. Then we calculate the 14th.

5.2.5 Unsupervised Pre processing

Involves working with the unlabeled data portion first before dealing with the labeled data in the subsequent stage. Within this family of approaches, there are at least three possibilities.

Feature extraction, which involves employing techniques such as Principal Components analysis or autoencoders to reduce dimensionality of the unlabeled data and to represent it more efficiently.

Cluster then label, which is the combined unlabeled plus labeled datasets are subjected to clustering algorithms such as K-Means, and then resulting clusters are used to train a classifier model. If most of the labeled instance with a given label appear in the same cluster, then that label is assigned to all the unlabeled data points in the same cluster. This is another example of pseudo labeling.

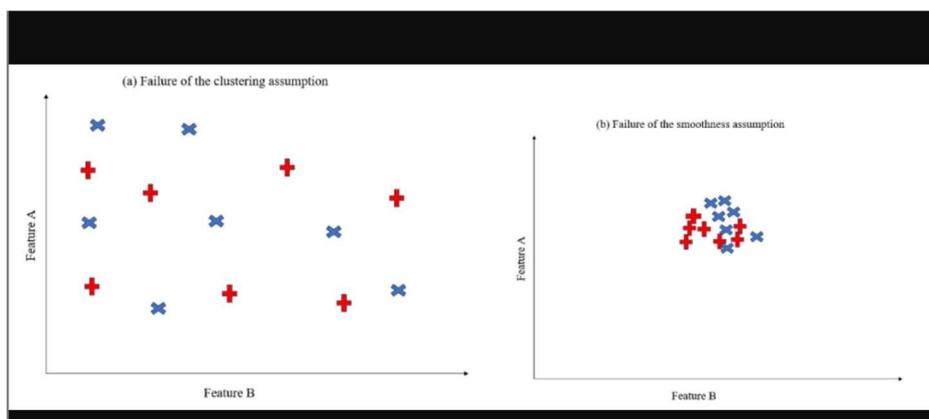
Pre training, here the unlabeled data points are formed into clusters that are useful to develop preliminary decision boundaries prior to applying supervised learning.

Appendix 5.A Semi Supervised Learning Assumptions

Its success relies on the clustering assumption, the smoothness or continuity assumption ad manifold assumption.

The figure below provides a pictorial representation of the first two assumptions. In panel (a) on the left the data belong to two classes, but fail to form clusters, and thus the clustering assumption is violated. The data points are dispersed in the feature space and do not form dense regions that can be separated.

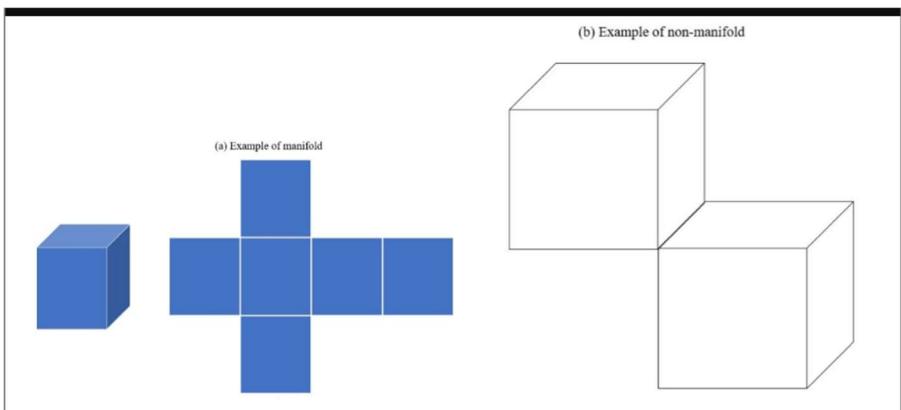
In panel (b) on the right, the data form clusters, but these are partly overlapping and cannot be separated using a line, and the smoothness assumption fails.



Artificial Intelligence Risk Certificate

The figure below shows the difference between manifold and non-manifold. The cubic box can be unfolded on a 2-dimensional space, and therefore each of its faces is a two-dimensional manifold. However, we will not be able to unfold the two cubes linked only by one hedge represented in panel b on the right in the same way as we did previously, meaning b is a non-manifold structure.

Notably if each face of the cube in panel (a) contained datapoints that belong to a single class, the manifold assumption would be satisfied. However, if one face contained datapoints belonging to different classes, the assumption would be violated as the datapoints in manifold would not belong to the same class.



Questions and Answers Module 2 Chapter 5 from GARP - Semi Supervised Learning

5.1

A- When should we consider using semi supervised learning?

Is most valuable when the dataset contains both labeled and unlabeled instances, but where the labeled portion is relatively small and where it is infeasible or excessively costly to manually add labels to the currently unlabeled instances.

B- How does self-training differs from Co-training in semi supervised learning?

Both are methods for applying pseudo labeled to the unlabeled part of the data in semi supervised learning. With self-training once the labeled data have been classified using a machine learning technique, the unlabeled data are assigned labels one at the time and added to the labeled portion of the data.

Co training adopts similar principals, but it involves splitting the features of the labeled and unlabeled data into two separate sub-datasets. Then, two separate classification models are built, and the most confident outcomes generated for each group of features become labeled instances for the other group of features.

6. Reinforcement Learning

Is a machine learning technique that applies a trial-and-error feedback loop to train models to optimize actions that maximize a defined long-term or cumulative reward.

The output from reinforcement learning applications is a recommended action based on defined parameters rather than a prediction, classification or cluster produced in unsupervised or supervised learning.

Chapter learning objectives are as follows.

Explain the key principals and frameworks behind reinforcement learning

Compare and contrast exploration, exploitation and e-greedy strategies.

Describe reinforcement learning in the context of the Multi Armed Bandit (MAB) problem.

Explain Markov decision processes.

Differentiate between the Monte Carlo and temporal difference methods.

Describe how neural networks can be used in reinforcement learning.

6.1 The Principles of Reinforcement Learning

Is concerned with developing a policy for a series of decisions to maximize a long-term reward. In reinforcement learning, the learner is presented with feedback on the quality of the reward in a process analogous to trial and error. The technique is advantageous when decisions need to be made repeatedly so that the algorithm can learn based on the rewards or sanctions received in previous rounds.

Unlike both unsupervised and supervised learning, the output from reinforcement learning applications is a recommended action given the circumstances, rather than a prediction, classification or cluster.

A typical example of a situation where reinforcement learning can be applied would be a video game in which players can hone their strategies based on whether they won or lost, and by how much, in prior turns.

A further frequently employed game is that of teaching a dog to do tricks.

These ones are great examples because each one of the players/dogs behave differently, and therefore a fixed set of universal instructions covering all steps in the process cannot be developed.

For these instances is easy to define success but hard to specify a priori what is the appropriate action in every situation.

Artificial Intelligence Risk Certificate

This technique has had very successful applications, in Chess and GO for instance. The algorithm learns by playing against itself many times and using a systematic trial and error approach.

More recent applications are on controlling movements in robots, self-driving cars, traffic light control, and inventory management.

There are many potential uses of reinforcement learning in finance, including for technical trading, determining how to split a large volume of trades to sell quickly while minimizing the adverse price effect, and determining how much of a position to hedge using derivatives.

A disadvantage is that they tend to require larger amounts of training data than other machine learning approaches. The improvement function rises exponentially, starting poorly and then improving to the point of maximum optimization.

6.2. The Multi-Arm Bandit Problem

This problem involves a gambler, agent, who can choose to play one of several different slot machines. The gambler believes that the machines have different probabilities of winning, but is unsure which machine is more generous than the others.

The gambler objective is to maximize the total payout from a fixed number of rounds.

In each trial, the gambler picks one machine and the outcome, reward, is either that they win, or they lose. Therefore, each machine has its own probability distribution of rewards, and each round has only one step, i.e. one action, one state and one reward.

There is no other gambler/agent involved, i.e. it's a single agent framework.

The gambler action in the current round does not affect the states in the subsequent rounds. Therefore, the current action only affects the current reward, not feature rewards.

6.2.1. Terminology in MABs

Such models are determined in terms of states, actions and rewards. The states define the environment, i.e., the slot machine in each round, an action is the decision taken, decision on which slot machine to play, and

Artificial Intelligence Risk Certificate

rewards are the goal of the problem, payout from the chosen slot machine.

The aim is to choose the decision that maximizes the value of total subsequent rewards that are earned, possibly applying a discount rate to the rewards.

Agent is the person or algorithm making the decision. Usually there is a single agent, although in some models it is possible to have more than one, in which case the agents could be working together or in competition.

Actions, A are the possible choices that an agent can select from at each timestep. In MAB, the agent is free to choose which slot machine to play.

State, S are the circumstances or a description of the environment, in which the decision is being made at each time step. Because it is assumed that our action do not change the slot machines in any way and we are always playing on those machines, there is only one single state for our slots machines, and it does not change.

Reward, R, is the feedback that the agent receives based on its previous action. This could be either positive or negative, reward or sanction.

Expected future rewards, G, are the expected value of future rewards. The objective is to maximize this one. In our MBA case, the objective is to maximize the payout we get in the future, whenever we choose a slot machine.

Artificial Intelligence Risk Certificate

Policy, π is the plan of action that the agent takes based on observing the current state. The policy maps the states to actions that will maximize the reward. Because there is only one single state in MAB problems, we only need to consider the policy in this state.

Value function, V, measures how good a state is. This is a.k.a, the state-value function. It relates the expected reward to a given state. It measures how good a state is. Because there is only one state in MABs, it is not relevant in this case, but it can be very useful for other problems with many states to guide the agent on policy improvement.

Action-value Function, Q, measures how good an action is, given a certain state, like our slot machines. This relates the expected reward to the actions and the state. It measures how good an action is, given a certain state. In the MBA context, it measures how rewarding it is if we choose a certain slot to play. This is very important and useful for us to compare the different actions and use them to optimize our policy.

The capital letters S and A are used to denote the set of states and the set of actions in general, whereas their lowercase counterparts denote specific states and actions. Time subscripts are generally suppressed unless they are specifically required for clarity, such as when describing the transition from the state in one period to the next.

Artificial Intelligence Risk Certificate

The relationship between the agent and the environment is depicted in [Figure 6.1](#), and [Figure 6.2](#) shows the linkages between the state, policy, and action.

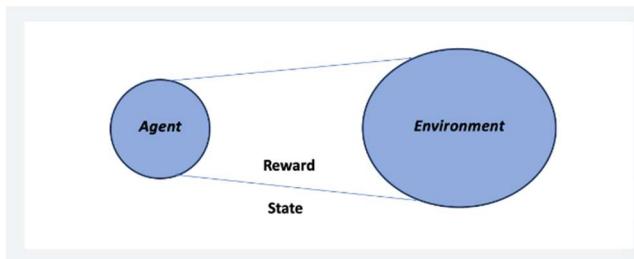


Figure 6.1 The Agent and the Environment

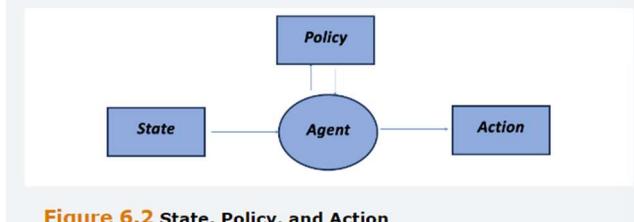


Figure 6.2 State, Policy, and Action

6.2.2. Strategy in MAB

A very intuitive way of playing is to always choose the best actions identified so far, which is called greedy strategy.

Greedy strategy, based on exploitation, is a simple strategy in which the agent always chooses the actions with the best rewards seen so far. In the MAB problem, it means that we always choose the slot machine that give us the best payout in a greedy way.

Artificial Intelligence Risk Certificate

This strategy focuses on the idea of exploitation of the information gained from the agent's experience so far. It may appear like a good strategy, but it has problems.

If we find one slot that seems to pay out well and stick with it, it may produce suboptimal results because we did not experiment with other slot machines, which may be better.

Random strategy, based on exploration, is an intuitive strategy where we randomly select a slot machine to play. Whereas the greedy strategy only chooses the action with the best payout up to that point, the random strategy is useful for exploring other possible actions.

Its problem is that does not exploit knowledge gained from the past rewards, to make more informed decisions over time.

We can see from the preceding two strategies that neither exploitation nor exploration alone are promising in the MAB problem. To address the shortcomings, can be employed a strategy of combining the two, called ε -greedy.

ε -greedy combines the exploitation and exploration. Epsilon is a hyperparameter, between 0 and 1, that determines whether a random selection is made to explore, or a greedy selection is made to exploit.

Usually, a random number between 0 and 1 is drawn. If that number is below Epsilon, we explore by selecting a

Artificial Intelligence Risk Certificate

random slot machine, otherwise we choose the machine that had the best payoff up to that point.

This helps us to continue to exploit the machine that provided best rewards so far, while still exploring other options.

Usually, a small value like 0.05 or 0.1 are attributed to Epsilon, in order for the agent to rely more on existing accumulated knowledge, than experimenting with new strategies , i.e. we want to exploit more than explore.

Although the Epsilon greedy strategy does not use random selection and it is more adaptive in the face of diverse varying reward structures, it might still select an obviously suboptimal action in many random trials.

A refinement to the Epsilon greedy strategy is to allow Epsilon to vary systematically throughout the exercise, so that it is initially larger, allowing a lot of experimentation while the amount of accumulated knowledge about the relationships between actions and rewards is low.

Then, the hyperparameter is gradually reduced as more information becomes known and the benefit of additional exploration is diminished because the algorithm has already learned more about the test strategy.

A popular approach is to use a decay factor, β , and set $\epsilon = \beta^{t-1}$, where t is the trial number and with Beta between 0 and 1.

Artificial Intelligence Risk Certificate

As an example, assume slot machine A,B and C and that the game can be played many times. The payoff from the i th machine is normally distributed with mean μ_i and Standard deviation of one. The means are known in advance.

$$\mu_A = 0.8 \quad \mu_B = 0.5 \quad \mu_C = 0.7$$

And decay factor $\beta = 0.85$ for Epsilon.

In the fist simulation, Epsilon equals one, as we know nothing about the slot machines and we must explore. We choose machine A and receive a payoff equal to 1.2. At this point, the value of Epsilon is now 0.85. Therefore, we drawn a random number between 0 and 1.

If the number below 0.85, we explore, vice versa we exploit.

Suppose that the Random number is 0.99, then we return to machine A to exploit. The payoff is now 0.8, so we will update our expected reward for machine A by averaging the two outcomes.

From the pictures below, although many other trials would be generally, needed, if we had to stop the process at this stage the best strategy identified would be to play slot machine A, because it was the highest payout ratio after 10 trials.

Also, as the number of trials increase, the use of exploration reduces thanks to de decay factor. The parameter Epsilon is close to 0 after about 57 trials, which implies that any new random number drawn will

Artificial Intelligence Risk Certificate

be greater than Epsilon and exploration stops. We juts paly the machine with the highest expected reward.

Table 6.1 The outcome of 10 trials in a multi-arm bandit problem.

Trial	ϵ	Random number (0,1)	Decision	Choice	Payoff	Reward
1	1	0.7	Explore	A	1.2	$Q(A) = 1.2$ $Q(B) = 0$ $Q(C) = 0$
2	0.85	0.99	Exploit	A	0.8	$Q(A) = 1$ $Q(B) = 0$ $Q(C) = 0$
3	0.72	0.65	Explore	B	1.7	$Q(A) = 1$ $Q(B) = 1.7$ $Q(C) = 0$
4	0.61	0.5	Explore	C	0.6	$Q(A) = 1$ $Q(B) = 1.7$ $Q(C) = 0.6$
5	0.52	0.7	Exploit	B	-0.7	$Q(A) = 1$ $Q(B) = 0.5$ $Q(C) = 0.6$
6	0.44	0.4	Explore	C	1	$Q(A) = 1$ $Q(B) = 0.5$ $Q(C) = 0.8$
7	0.38	0.8	Exploit	A	-0.2	$Q(A) = 0.6$ $Q(B) = 0.5$ $Q(C) = 0.8$
8	0.32	0.55	Exploit	C	0.5	$Q(A) = 0.6$ $Q(B) = 0.5$ $Q(C) = 0.7$
9	0.27	0.25	Explore	A	1.2	$Q(A) = 0.75$ $Q(B) = 0.5$ $Q(C) = 0.7$
10	0.23	0.45	Exploit	A	1	$Q(A) = 0.8$ $Q(B) = 0.5$ $Q(C) = 0.7$

Artificial Intelligence Risk Certificate

In general, the new value of action-value function, Q_k, for the kth slot machine after it was chosen for n number of trials is:

$$Q_k^n = \frac{1}{n} \sum_{j=1}^n R_j$$

Here R_j is the reward for the jth trial. It can be shown with simple algebra that each time slot machine K is chosen, its new Q-Value will be a weighted average of its old Q-value and the new reward, with weights $\frac{n-1}{n}$ and $\frac{1}{n}$ respectively.

$$\begin{aligned} Q_k^n &= \frac{1}{n} \sum_{j=1}^n R_j = \frac{1}{n} \sum_{j=1}^{n-1} R_j + \frac{1}{n} R_n = \frac{n-1}{n} Q_k^{n-1} + \frac{1}{n} R_n = \\ &= Q_k^{n-1} + \frac{1}{n} (R_n - Q_k^{n-1}) \end{aligned}$$

The MAB problem is clearly much simpler than the setup for most potential applications of reinforcement learning because there is only one state here and the slot machines do not change. One example of problems with multiple states is the Markov decision process.

6.3 Markov Decision Processes

Markov Decision Processes, MDPs, are simple settings for environment dynamics. In this case, the environment changes based on the actions of the agent. MDPs are processes that have no memory, which means that only the current state is relevant for determining the most

Artificial Intelligence Risk Certificate

appropriate current action and not any of the previous states.

MDPs are useful for modeling decision making in the cases where the agent is not fully in control of the evolution of the states. The use of MDPs establishes a straightforward framework where there are m states, denoted s , each of which will occur with a given probability, and there is also a fixed probability of being in a particular state S_{t+1} at time t+1 given that the state at time t was S_t .

$$Prob(S_{t+1} = s_{t+1} | S_t = s_t)$$

The assumption that each state follows a Markov process greatly simplifies the analysis because such processes have no memory, as described earlier. We can express Markov property as:

$$\begin{aligned} Prob(S_{t+1} = s_{t+1} | S_t = s_t, S_{t-1} = s_{t-1}, \dots, S_1 = s_1) &= \\ &= Prob(S_{t+1} = s_{t+1} | S_t = s_t) \end{aligned}$$

That is, the future state at t+1 is only dependent on the current state at t, and independent of past states at t-1, t-2 etc. We can then specify a transition probability matrix, P, that shows the probabilities of moving from any state in one period to any other state in the next period.

The probabilities of being in each state n timesteps into the future, given an initial state, are then simply given by the elements P of n, i.e. the transition matrix P multiplied by itself n times.

Artificial Intelligence Risk Certificate

The Markov assumption therefore provides a simple way for the algorithm to determine how likely each future state is given the current state.

The agent will select an action at time t, $A_t = a$, based on observing state $S_t = s$, and receives a reward, R_{t+1} in the next period at $t+1$ as a function of the state $S_{t+1} = s'$ in that period and the action in the previous period:

$$R_{t+1} = f(s', a)$$

The agent goal will be to select the policy that maximizes this expected return aggregated over all feature time periods, note that the current and previous returns are “sunk costs” and hence are not included in the objective function).

Denoting the discount factor by γ , with γ between zero and 1, we can define the goal at time t, G_t as:

$$G_t = R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \dots)$$

The term in parentheses is the equation for G_{t+1} and so the expression G_t could be further redrafted as a recursion:

$$G_t = R_{t+1} + \gamma G_{t+1}$$

Given the definition of G_t , we can define the state-value function V which measures how good a certain state s is following a certain Policy, π as :

$$V_\pi(s) = E_\pi[G_t | s]$$

Artificial Intelligence Risk Certificate

We can also define the action-value function, Q , which measures how good a certain action a is, in a state s , following a certain policy, π , as its expected return:

$$Q_\pi(s, a) = E_\pi[G_t | s, a]$$

Clearly, the agent's objective will be to choose the optimal policy, π^* , that maximizes Q .

$$\pi^* = \text{argmax } Q_\pi(s, a)$$

Where argmax is the set of values of π for which Q_π is maximized.

6.4 Approaches to Reinforcement Learning

Reinforcement Learning algorithms can be classified as model-based and model-free algorithms. When there are a limited number of states, with well defined actions and the transition probabilities are well defined, a dynamic programming technique can be used to obtain a solution.

Typically, however, we only have partial information about the model. In such cases, the algorithm search for an optimal solution to maximize the reward. There are two different approaches to find the policy that maximizes Q .

Value-based approaches that work on maximizing the reward by determining the best action in each state. Several algorithms such as Temporal difference method,

Artificial Intelligence Risk Certificate

Q-learning, SARSA and Deep Q-learning belong to this category.

Policy-based approaches that find the optimal policy to map a state into an action. The Policy gradient approach is a commonly used policy-based algorithm.

6.5.1 The Bellman Equations

Provide a framework for evaluating policies. The equations can be written either in terms of values, V, or in terms of action-values, Q.

They establish an updating mechanism, or in other words a recursive algorithm that sets the current value or action-value function for a given policy equal to the reward from the function a at time t, plus the discounted value function of future rewards at time t+1.

The equations make it possible to pinpoint actions that will lead to greater future values of R for each given state. The Bellman optimality equation for the value function V can be stated as:

$$V_t^*(s) = \max E[R_{t+1} + \gamma V_{t+1}(s')]$$

Where s and s' are the states at t and t+1 respectively.

The optimality equation for Q, the action value is:

$$Q_t^*(s, a) = \max E[R_{t+1} + \gamma V_{t+1}^*(s')]$$

Or because $V_t^*(s) = \max Q_t^*(s, a)$

Artificial Intelligence Risk Certificate

$$Q_t^*(s, a) = \max E[R_{t+1} + \gamma \max Q_{t+1}^*(s', a')]$$

Here, a and a' are the actions taken at time t and $t+1$. Solving the Bellman equations give the optimal policy, π^* . If all the rewards and transition probabilities are known, then dynamic programming can be used to determine π^* .

But they are unlikely to be known in practice, in which case an iterative technique is required. There are two common ways to solve reinforcement learning problems iteratively: Monte Carlo and Temporal difference Methods.

6.5.2 The Monte Carlo Method

In MC we conduct trials, or episodes, repeatedly, using random initialization for each state and estimating the average reward for each state over all episodes for each policy π_i .

Through repeated trials, the algorithm develops an estimate of the expected value of acting A in state S . Usually, we perform trials until we meet convergence criterion which depends on the problem domain or the environment. The resulting action-value Function, $Q(s,a)$, is the value of taking action A in state S .

In Monte Carlo, we directly work with $Q(s,a)$, the action-value function.

Artificial Intelligence Risk Certificate

Suppose that the algorithm acts a in state s and the total subsequent discounted rewards prove to be G. Under the Monte Carlo method, Q (s,a), instead of equally weighting the trials we use exponentially moving average for the trials and update as follows.

If we assume equal weights:

$$\begin{aligned}Q_k^n &= \frac{1}{n} \sum_{j=1}^n R_j = \frac{1}{n} \sum_{j=1}^{n-1} R_j + \frac{1}{n} R_n = \frac{n-1}{n} Q_k^{n-1} + \frac{1}{n} R_n = \\&= Q_k^{n-1} + \frac{1}{n} (R_n - Q_k^{n-1})\end{aligned}$$

In Monte Carlo with Exponential Moving Average:

$$Q(s, a) \leftarrow Q(s, a) + \alpha [G_t - Q(s, a)]$$

Where alpha is the smoothing factor that is chosen after some experimentation. The quantity of alpha controls how much Q is updated at each iteration when a new reward value R is observed.

The MC method for reinforcement learning has two mains' disadvantages. First, it can only be used for processes that have a finite horizon, in other words, the rewards do not continue into the indefinite future. Secondly, convergence to the true reward for each state could be infeasibly slow, particularly for long episodes with many states.

6.5.3 The Temporal Difference TD Method

The technique begins by assuming that the agent is only aware of the states and possible actions that could be taken. The likelihood of each state and the transition probabilities are initially unknown.

The objective is to estimate $V_t(S)$, but unlike Monte Carlo, there is no need to wait until the end of the episode before assigning the rewards. Instead, the temporal difference method can be used for very long episodes or those with infinite lifetimes.

The temporal difference method combines the immediate reward and the discounted value of the next state. Suppose that at time t the algorithm acts in a state s and a reward of R is earned. It then moves to s', where the value is $V_{t-1}(s')$. Under the temporal difference method, the value function is updated towards its target $R_{t+1} + \gamma V_{t+1}(s')$.

$$V_t(s) \leftarrow V_t(s) + \alpha[R_{t+1} + \gamma V_{t+1}(s') - V_t(s)]$$

The updating process is controlled by the hyperparameter alpha. Instead of updating the value function, we can work with Q, the action-value function and update it as:

$$Q_t(s,a) \leftarrow Q_t(s,a) + \alpha[R_{t+1} + \gamma \max Q_{t+1}^*(s',a') - Q_t(s,a)]$$

This is called Q-learning.

Artificial Intelligence Risk Certificate

6.5.4 Illustrative Example

Suppose that there are 4 states and 3 actions, and that the current Q(S,A) values are the following ones.

Table 6.2 Current Q values

	State 1	State 2	State 3	State 4
Action 1	0.1	0.2	0.4	0.2
Action 2	0.8	0.3	0.5	0.1
Action 3	0.3	0.7	0.9	0.8

Suppose that on the next trial, action 3 is taken in state 4 and the total subsequent reward, G, is 1.0

If alpha=0.05, using the updating equation for the Monte Carlo method, would lead to Q(4,3) being updated from 0.8 to:

$$Q(s, a) \leftarrow Q(s, a) + \alpha[G_t - Q(s, a)]$$

$$0.8 + 0.05(1.0 - 0.8) = 0.81$$

Suppose we are going from State 4 in the current trial to State 3 in the next trial after taking action 3.

Suppose further that a reward R, of 0.2 is earned between the two dimensions. Assuming that γ is 0.9 and V is 1.0 in state 3, γV , the discounted value is 0.9.

Artificial Intelligence Risk Certificate

Using temporal difference method, if alpha is 0.05, this would lead to Q(4,3) being updated from 0.8 to:

$$Q_t(s,a) \leftarrow Q_t(s,a) + \alpha [R_{t+1} + \gamma \max Q_{t+1}^*(s', a') - Q_t(s,a)]$$

$$0.8 + 0.05(0.2 + 0.9 - 0.8) = 0.815$$

6.5.5 Curse of Dimensionality and Neural Network Approximation

For the methods previously discussed we are still able to define the value function and action-value function as lookup tables, where the values are stored in a table corresponding to their states and actions.

This is possible for simple reinforcement learning problems with a small number of actions and states. But it becomes impractical when the number of actions and states becomes very large, even infinite sometimes in problems such self-driving cars, which would require an infinitely large table to store all possible actions and states. This also leads to an exponentially growing computational cost.

In many real world applications, there are simply too many states, making it impractical to tabulate them all, so a different approach is required. In such circumstances, the task of the model will be to learn the value of each action as a function of the current state by averaging over many possible future rewards arising after the action taken at this stage.

Artificial Intelligence Risk Certificate

Instead of calculating and then using the possibly infinitely large lookup table, we can estimate the Q value function for the given state and action, with a significantly smaller number of parameters we need to store compared to a tabular approach.

Neural Networks can be used to estimate the complete table from the observations that are available. This approach is called deep reinforcement learning. In this approach, the Q value function is estimated by training a neural network.

The gradient descent method is used to estimate the weights of the network by minimizing a loss function which is the sum of the squared difference between the targeted and neural networks estimate of the Q values. Once the network is trained, it can be used to generate Q values for any input set of state variables. This replaces the need to generate a very large lookup table for the Q values.

6.6 Chapter Summary

We started from a simple one-state reinforcement learning example using MAB, and introduced and formalized basic concepts in reinforcement learning problem setting with a discussion of how to optimize the policy with the E-selection Strategy. Then we tackled MDPs with multi states, introducing a dynamic programming solution if the exact model for state transition is known, and other sampling solutions such

Artificial Intelligence Risk Certificate

as Monte Carlo and Temporal Difference methods. Finally, we discussed real-word level reinforcement learning problems with a very large, even infinite, number of states, where we have too many states for storage and calculation, which requires using neural networks for approximating the value function or the action value function.

Appendix 6.A Markov Transition Probabilities

Consider a simple scenario where there are 3 states, A, b and C in a Markov decision process. The transition probability is the probability of moving from one state at time t to another state at next period, t+1.

Typically, a transition probability matrix is used to summarize the probabilities.

$$\Pi(t, t + 1) = \begin{pmatrix} P_{aa} & P_{ab} & P_{ac} \\ P_{ba} & P_{bb} & P_{bc} \\ P_{ca} & P_{cb} & P_{cc} \end{pmatrix}$$

In the above matrix, the first element of the first row, P_{aa} is the probability of remaining in state A from time say t to t+1.

The next element, P_{ab} is the probability of moving from state A at t to state B at t+1. And so on.

Using this matrix, we can calculate the probabilities of moving from one state to another state in two time periods, e.g., t=0 and t=2. This requires the multiplication of transition probability matrix by itself.

Artificial Intelligence Risk Certificate

$$\Pi(0,2) = \Pi(0,1)\Pi(0,1)$$

In general, the n step transition probabilities can be calculated as:

$$\Pi(0,n) = \Pi^n(0,1)$$

Consider the matrix below:

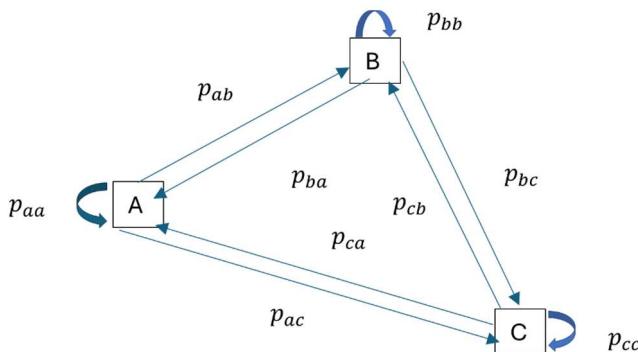
$$\Pi = \begin{pmatrix} 0.9 & 0.09 & 0.01 \\ 0.09 & 0.8 & 0.11 \\ 0.01 & 0.11 & 0.88 \end{pmatrix}$$

The transition probability matrix for moving from t=0 to t=2, in two-time steps is:

$$\Pi^2 = \begin{pmatrix} 0.82 & 0.15 & 0.03 \\ 0.15 & 0.66 & 0.19 \\ 0.03 & 0.19 & 0.79 \end{pmatrix}$$

Similarly, in 3-time steps:

$$\Pi^3 = \begin{pmatrix} 0.75 & 0.20 & 0.05 \\ 0.20 & 0.56 & 0.24 \\ 0.05 & 0.24 & 0.71 \end{pmatrix}$$



Appendix 6.B Detailed Reinforcement Learning Example- The game of Nim

For simplicity we will be using a total of 6 coins. These coins are placed on a table and each player can remove either 1,2 or 3 coins at the same time. The winner is the player who removes the last coin.

We begin without any strategy and employ reinforcement learning to develop an optimal one. We assume that the opponent, behave randomly rather than optimally. To setup the framework, we employ the Epsilon greedy policy, with initial epsilon equal to 1 and the decay factor, beta, of 0.9995. Additionally, we chose the learning rate, alpha, as 0.05 and discount factor γ ,as 1.

Monte Carlo Method

First, we initialize the Q-table by setting all Q-values to 0.

Table 6.B.1: Initial Q-Table

Coins picked up	State (Number of coins left)					
	1	2	3	4	5	6
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0

Artificial Intelligence Risk Certificate

First Episode

Step 1

State: 6 coins, Action: Player1 picks 3 Coins

State: 3 coins, Action: Player 2 picks 2 coins.

Step 2

State: 1 coin; Action: Player 1 picks 1 coin.

After completing this episode and player 1 received a reward of 1, we update the Q-table using the following formula:

$$Q^{new}(S, A) = Q^{old}(S, A) + \alpha[G - Q^{old}(S, A)]$$

where G is equal to the total expected discounted return

$$G = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3}$$

In this episode, players only receive reward at the end of the episode. At t=2, Player 1 receives a reward of $R_2 = 1$. Since $R_1 = 0$ and γ assumed to be 1, $G = R_1 + \gamma R_2 = 1$

Therefore, after the 1st step:

$$\begin{aligned} Q^{new}(6,3) &= Q^{old}(6,3) + \alpha[G - Q^{old}(6,3)] \\ &= 0 + 0.05 * (1 - 0) = 0.05 \end{aligned}$$

Therefore, after the 2nd step:

$$\begin{aligned} Q^{new}(1,1) &= Q^{old}(1,1) + \alpha[G - Q^{old}(1,1)] \\ &= 0 + 0.05 * (1 - 0) = 0.05 \end{aligned}$$

Artificial Intelligence Risk Certificate

Table 6.B.2: Q-Table after First Episode

Coins picked up	State (Number of coins left)					
	1	2	3	4	5	6
1	0.05	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0.05

Second Episode

Step 1

State: 6 coins, Action: Player1 picks 1 Coins

State: 5 coins, Action: Player 2 picks 1 coins.

Step 2

State: 4 coins; Action: Player 1 picks 3 coins.

State:1 coin, Action: Player2 picks 1 coin.

Player 2 wins

After completing this episode, player 1 receives a reward of G=-1.

Therefore, after the 1st step:

$$\begin{aligned}Q^{new}(6,1) &= Q^{old}(6,1) + \alpha[G - Q^{old}(6,1)] \\&= 0 + 0.05 * (-1 - 0) = -0.05\end{aligned}$$

Artificial Intelligence Risk Certificate

Therefore, after the 2nd step:

$$Q^{new}(4,3) = Q^{old}(4,3) + \alpha[G - Q^{old}(4,3)] \\ = 0 + 0.05 * (-1 - 0) = -0.05$$

Table 6.B.3: Q-Table after Second Episode

Coins picked up	State (Number of coins left)					
	1	2	3	4	5	6
1	0.05	0	0	0	0	-0.05
2	0	0	0	0	0	0
3	0	0	0	-0.05	0	0.05

Third Episode

Step 1

State: 6 coins, Action: Player1 picks 3 Coins

State: 3 coins, Action: Player 2 picks 2 coins.

Step 2

State:1 coin, Action: Player 1 picks 1 coin.

Player 1 wins

After completing this episode, player 1 receives a reward of G=1.

Therefore, after the 1st step:

$$Q^{new}(6,3) = Q^{old}(6,3) + \alpha[G - Q^{old}(6,3)] \\ = 0.05 + 0.05 * (1 - 0.05) = 0.0975$$

Artificial Intelligence Risk Certificate

Therefore, after the 2nd step:

$$\begin{aligned} Q^{new}(1,1) &= Q^{old}(1,1) + \alpha[G - Q^{old}(1,1)] \\ &= 0.05 + 0.05 * (1 - 0) = 0.0975 \end{aligned}$$

Table 6.B.4: Q-Table after Third Episode

Coins picked up	State (Number of coins left)					
	1	2	3	4	5	6
1	0.0975	0	0	0	0	-0.05
2	0	0	0	0	0	0
3	0	0	0	-0.05	0	0.0975

After 1,000,000 episodes, the Q-values converge into the following.

Table 6.B.5: Q-Table after 1,000,000 Episodes

Coins picked up	State (Number of coins left)					
	1	2	3	4	5	6
1	1	-1	-0.071	0.380	0	0.774
2	0	1	-1	-0.110	0	1
3	0	0	1	-0.858	0	0.020

Artificial Intelligence Risk Certificate

The program eventually learns the correct strategy which is:

Number of Coins	Optimal Strategy
6	Pick 2 coins
5	Not applicable
4	Pick 1 coin
3	Pick 3 coins
2	Pick 2 coins
1	Pick 1 coin

Temporal Difference Method

We update the Q-values within each episode by considering one time-step ahead of the state we are in. We update the Q values using the following formula.

$$Q^{new}(S, A) = Q^{old}(S, A) + \alpha[R_{t+1} + \gamma V(s') - Q^{old}(S, A)]$$

Where $V(S) = \max(Q(S, A))$ is the state value function that measures how good a state is.

Note that in our example, R_{t+1} will always equal 0 unless we are at the last step, because players only receive rewards at the end and not an intermediary step within an episode.

Artificial Intelligence Risk Certificate

Table 6.B.6: Initial Q-Table

Coins picked up	State (Number of coins left)					
	1	2	3	4	5	6
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0

First Episode

Step 1

State: 6 coins, Action: Player1 picks 3 Coins

State: 3 coins, Action: Player 2 picks 2 coins.

Player 1 will move from S=6 to S'=1

Step 2

State:1 coin, Action: Player 1 picks 1 coin.

Player 1 wins

Therefore, after the 1st step player 1 moved to S'

$$V(s') = V(1) = \max(Q(1, A)) = 0$$

$$\begin{aligned} Q^{new}(6,3) &= Q^{old}(6,3) + \alpha[R_{t+1} + \gamma V(1) - Q^{old}(6,3)] \\ &= 0 + 0.05 * (0 + (1 * 0) - 0) = 0 \end{aligned}$$

Therefore, after the 2nd step, no coins are left, and Player 1 won and receive the reward R_{t+1}=1, V(0)=0.

$$\begin{aligned} Q^{new}(1,1) &= Q^{old}(1,1) + \alpha[R_{t+1} + \gamma V(0) - Q^{old}(1,1)] \\ &= 0 + 0.05 * (1 + (1 * 0) - 0) = 0.05 \end{aligned}$$

Artificial Intelligence Risk Certificate

Table 6.B.7: Q-Table after First Episode

Coins picked up	State (Number of coins left)					
	1	2	3	4	5	6
1	0.05	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0

We subsequently update $V(1)$ with the maximum value of 0.05.

Second Episode

Step 1

State: 6 coins, Action: Player1 picks 1 Coins

State: 5 coins, Action: Player 2 picks 1 coins.

Player 1 will move from S=6 to S'=4

Step 2

State:4 coin, Action: Player 1 picks 3 coin.

State:1 coin, Action: Player 2 picks 1 coin.

Player 2 wins

Therefore, after the 1st step player 1 moved to S'

$$V(s') = V(4) = \max(Q(4, A)) = 0$$

$$\begin{aligned} Q^{new}(6,1) &= Q^{old}(6,1) + \alpha[R_{t+1} + \gamma V(4) - Q^{old}(6,1)] \\ &= 0 + 0.05 * (0 + (1 * 0) - 0) = 0 \end{aligned}$$

Artificial Intelligence Risk Certificate

Therefore, after the 2nd step, no coins are left, and Player 1 lost and receive the reward R_{t+1}=-1, V(0)=0.

$$Q^{new}(4,3) = Q^{old}(4,3) + \alpha[R_{t+1} + \gamma V(0) - Q^{old}(4,3)] \\ = 0 + 0.05 * (-1 + (1 * 0) - 0) = -0.05$$

Table 6.B.8: Q-Table after Second Episode

Coins picked up	State (Number of coins left)					
	1	2	3	4	5	6
1	0.05	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	-0.05	0	0

Third Episode

Step 1

State: 6 coins, Action: Player1 picks 3 Coins

State: 3 coins, Action: Player 2 picks 2 coins.

Player 1 will move from S=6 to S'=1

Step 2

State:1 coin, Action: Player 1 picks 1 coin.

Player 1 wins

Therefore, after the 1st step player 1 moved to S'

$$V(s') = V(1) = \max(Q(1,A)) = Q(1,1) = 0.05$$

$$Q^{new}(6,3) = Q^{old}(6,3) + \alpha[R_{t+1} + \gamma V(1) - Q^{old}(6,3)] \\ = 0 + 0.05 * (0 + (1 * 0.05) - 0) \\ = 0.0025$$

Artificial Intelligence Risk Certificate

Therefore, after the 2nd step, no coins are left, and Player 1 won and receive the reward $R_{t+1}=1$, $V(0)=0$.

$$\begin{aligned} Q^{new}(1,1) &= Q^{old}(1,1) + \alpha[R_{t+1} + \gamma V(0) - Q^{old}(1,1)] \\ &= 0.05 + 0.05 * (1 + (1 * 0) - 0.05) \\ &= 0.0975 \end{aligned}$$

Table 6.B.9: Q-Table after Third Episode

Coins picked up	State (Number of coins left)					
	1	2	3	4	5	6
1	0.0975	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	-0.05	0	0.0025

We update this 1,000,000 times to obtain:

Table 6.B.10: Q-Table after 1,000,000 Episodes

Coins picked up	State (Number of coins left)					
	1	2	3	4	5	6
1	1	-1	-0.105	0.383	0	0.845
2	0	1	-1	0.058	0	1
3	0	0	1	-0.842	0	0.011

Number of Coins	Optimal Strategy
6	Pick 2 coins
5	Not applicable
4	Pick 1 coin
3	Pick 3 coins
2	Pick 2 coins
1	Pick 1 coin

Questions and Answers Module 2 Chapter 6 from GARP - Reinforcement Learning

6.1 State whether the following statements are true or false and explain why

A- Reinforcement Learning algorithms could never outperform a human in competitive games such as chess.

False. Although reinforcement learning mimics the way that humans learn to perform tasks effectively by trial and error, the ability of computers to be able to simulate future states and evaluate the effectiveness of different possible actions across those states means that reinforcement learning can outperform human decision makers.

In fact, there are many situations where reinforcement learners have outperformed even the most skilled humans.

B- The Bellman Equations can specify the links between the states and the agent's action

False. The Bellman equations define the value of a policy, in other words, taking a particular action in a particular state. The equation specifies that the value of taking a particular action in a particular state is given by the expected sum of the next period reward plus the

Artificial Intelligence Risk Certificate

discounted value for the next state reached after that action.

C- The Monte Carlo involves simulating the entire episode before updating the Q function whereas Temporal Differences updates Q function one time step at a time.

True. This is precisely the difference between the two approaches to solving reinforcement learning problems. The MC update strategies using the total future rewards of one episode. Temporal difference method looks only one decision ahead when updating strategies.

D- Reinforcement Learning is underpinned by the assumption that the agent's goal is to maximize their total (discounted) future rewards.

True. To form a policy, in other words, what action the agent should take in each state, we need to assume about the agent's goal, and this is usually that they want to maximize their return over the longer term, i.e., over the entire exercise. Future Rewards are usually discounted in the same way as cashflows, because they are less attractive the further ahead in the future that they are received.

Artificial Intelligence Risk Certificate

E- The ε -Greedy strategy involves both exploration and exploitation whereas the random strategy only exploits.

False. The epsilon greedy is correct but the Random strategy is based on exploration instead of exploitation.

F-In a Markov process, the actions taken in the current state are dependent on the current state and previous states.

False. The MDPs are simple settings for environment dynamics. In this case the environment changes based on the actions of the agent. MDPs are processes that have no memory, which means that only the current state is relevant for determining the most appropriate current action and not any of the previous states.

6.2 What are the three strategies that can be used in MAB? How do they differ?

Greedy, random and epsilon greedy. The greedy strategy is simple in which the agent always chooses the actions with the best rewards seen so far. So, it tends to stick to one action. The random strategy is to randomly select an action to take. So, it tends to change actions. The Epsilon Greedy combines both Exploitation and Exploration and uses a hyperparameter epsilon, between 0 and 1 to decide if exploits or explores. If Random number >Epsilon Exploits, otherwise explores.

Artificial Intelligence Risk Certificate

6.3 Consider a MAB with Beta equal to 0.99 and five slot machines.

A- What Is the value of Epsilon on the 89th trial assuming an exponential decay factor is used?

The formula is the following:

$$\varepsilon = \beta^{t-1} = 0.99^{89-1} = 0.41$$

B- Assume that a random number of 0.8 has been extracted at the 89th trial. Should you explore or exploit.

Exploit since $0.8 > 0.41$

C- At trial 150, you explore the third machine for the 50th time. The current value of the expected reward for this machine is 1.8. You obtain a payoff of 1. What is the updated expected reward for the machine?

The formula is the following:

$$Q_3^n = \frac{n-1}{n} * 1.8 + \frac{1}{n} * 1 = 1.8 + \frac{1}{50}(1 - 1.8) = 1.784$$

Simply is the weighted average.

7.0 Supervised Learning – Model Estimation

Learning objectives:

Compare and contrast the Ordinary Least Squares and Maximum likelihood methods.

Explain how gradient descent method is used to optimize parameter estimates

Explain how backpropagation is used to determine the weights in neural networks.

Discuss the difference between underfitting and overfitting and potential remedies for each.

Describe the tradeoff between bias and variance

Explain the use of regularization techniques to simplify models.

Describe cross validation and its uses

Describe the accuracy-interpretability tradeoff.

Describe how grid search and bootstrapping can be used to optimize hyperparameter estimation.

Calculating the parameters, often called weights, is an essential step in the process of building a model. There are three families of techniques that could be potentially used for this purpose

Least Squares, we choose the parameter values that minimize the residual sum of squares.

Artificial Intelligence Risk Certificate

Maximum Likelihood, in which we form a likelihood function and choose the parameter values that maximize it. This provides parameter estimates that maximize the likelihood that we would observe the data that occurred.

The method of Moments, in which we construct a set of “moment restrictions” based on an assumed distribution for the data and we solve them to choose the parameters.

In fact, under certain conditions, both the least Squares and the maximum likelihood techniques are nested within the method of moments framework.

Consequently, for simple models, the three approaches will yield identical parameter estimates. The third approach would not be very useful in machine learning context and so is rarely used for such applications and will not be discussed further here. The remainder of this chapter therefore concentrates on least squares and maximum likelihood. These two methods and their various extensions, constitute virtually the only tools we might need to estimate machine learning models.

In all cases, we form an objective function and either maximize, i.e. the likelihood, or minimize, i.e. OLS. For least squares, the objective function is also known as the loss function. Optimization is the process of finding the parameter estimates that best fits the data.

An essential distinction that we need to draw is between analytical and numerical methods for optimization.

Artificial Intelligence Risk Certificate

Analytical methods are those where there is a closed-form solution to the optimization problem and therefore the optimal parameter values will be unique and can be calculated using a formula or a set of formulas.

Analytical solutions will usually be available when the objective function can be differentiated with respect to the parameters. On the other hand, when such differentiation cannot be conducted or is too complex to be evaluated, because of high dimensionality for instance, then a numerical procedure can be employed to estimate the parameters.

Numerical processes usually start with an initial guess for each of the parameters and then the optimizer tries to improve on these initial values to gradually move towards a set of estimates that fit the data well.

Many of the estimation methods we discuss in this chapter involve the use of hyperparameters. These are parameters that are used to specify the configuration of a model or the learning process while training the model. Examples are the number of hidden layers in a neural network and the learning rate used in the optimization methods discussed later in the chapter.

Artificial Intelligence Risk Certificate

7.1.1 Ordinary Least Squares

Is the most straightforward of the available estimation methods and is an analytical approach used to estimate the parameters in linear regression models.

Salary and Experience of bank employees.

Observation, i	Experience, x_i	Salary, y_i
1	1	15.46
2	3	22.40
3	7	27.47
4	12	34.31
5	9	33.08
6	3	18.70
7	20	35.06
8	22	35.78
9	0	14.81
10	4	14.84
11	6	25.78
12	8	28.17
13	4	26.36
14	2	11.12

In this case, suppose that we use a simple linear regression model so that there is one future, x_i (experience) and one label target, y_i , salary.

Artificial Intelligence Risk Certificate

We would like to identify the intercept and slope parameters in the model that best describe the relationship between the variables.

OLS finds this line by forming the residual sum of squares as a function of the parameters and minimizing it.

$$y_i = b_0 + b_i x_i + u_i$$

u of i is the error term and is assumed to have zero mean and constant variance, also known as a disturbance term.

The error term allows for the fact that it is very unlikely there would be a perfect linear relationship between the two variables. We define the residual, \hat{u}_i , as the difference between the actual value of series, y_i and the fitted value from the model for the data point i , \hat{y}_i .

$$\hat{u}_i = y_i - \hat{y}_i$$

Then we can construct the RSS, which is obtained by taking each of the residuals, squaring them, and then adding them over all N data points.

$$RSS = \sum_{i=1}^N \hat{u}_i^2 = \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Note that we square the residual first and then simply add, otherwise the positive and negative residuals, would cancel each other out, and squaring makes all of them positive.

Artificial Intelligence Risk Certificate

The mean squared error, MSE averages the residual sum of squares.

$$MSE = \frac{1}{N} \sum_{i=1}^N \hat{u}_i^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

We want to find the values of the parameters that minimize this RSS, and so we substitute \hat{y}_i , with the fitted equation for the line, $\hat{y}_i = \hat{b}_0 + \hat{b}_i x_i$ where the above parameters denote the estimated values.

$$RSS = \sum_{i=1}^N (y_i - \hat{y}_i)^2 = \sum_{i=1}^N (y_i - \hat{b}_0 - \hat{b}_i x_i)^2$$

We partially differentiate this expression for the RSS separately with respect to each of the parameters, set the derivatives to zero and then rearrange the formula to obtain expressions for the optimal intercept and slope estimates.

Intercept:

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

Slope:

$$\hat{b}_1 = \frac{\sum_{i=1}^N y_i x_i - N \bar{y} \bar{x}}{\sum_{i=1}^N x_i^2 - N \bar{x}^2}$$

Where \bar{x} and \bar{y} denote the mean of the observations x and y.

Any other regression line having a different intercept or slope coefficient would fit the data less well and would lead to a higher RSS/MSE.

Artificial Intelligence Risk Certificate

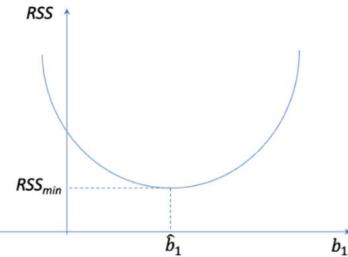


Figure 7.1 How RSS changes as the parameter moves away from its optimal value

OLS minimizes the sum of squares of the distances from all the points together to the fitted line.

From our exercise $\hat{b}_0 = 16.88$ and $\hat{b}_1 = 1.06$, so if we add for instance observation 5 to the chart, we would obtain a residual of.

$$33.08 - (16.88 + 9 * 1.06) = 6.66$$

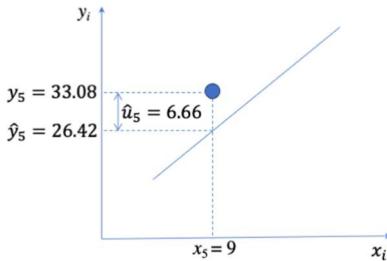


Figure 7.2 The residual, actual and fitted values for data point number 5

Artificial Intelligence Risk Certificate

Observation, i	Experience, x_i	Salary, y_i	Fitted value, \hat{y}_i	Residual, \hat{u}_i
1	1	15.46	17.94	-2.48
2	3	22.40	20.06	2.34
3	7	27.47	24.30	3.18
4	12	34.31	29.59	4.71
5	9	33.08	26.42	6.66
6	3	18.70	20.06	-1.36
7	20	35.06	38.07	-3.01
8	22	35.78	40.19	-4.40
9	0	14.81	16.88	-2.07
10	4	14.84	21.12	-6.28
11	6	25.78	23.24	2.54
12	8	28.17	25.36	2.81
13	4	26.36	21.12	5.24
14	2	11.12	19.00	-7.88

Although OLS will minimize the collective distances from the data points to the line, none of the residuals is very close to zero, and some points are well above the line, while others are well below. If we believe the noise level is low, then we may suspect that the model fit is not very good, which might motivate us to consider alternative or additional predictors such as square of the experience term, as previously discussed.

A similar approach would be used in the multiple linear regression context where we have more than one

Artificial Intelligence Risk Certificate

explanatory variable. If there were two explanatory variables, i.e. three parameters to estimate, one intercept and 2 slopes, we could derive a set of three equations like the two above. But in the case of more explanatory variables than 2, although the formula can still be derived, the become increasingly long and unwieldy.

Therefore, it is more common to use a matrix notation for the model and the estimation formula, will extend naturally and straightforwardly to however many explanatory variables there are.

7.1.2 Nonlinear Least Squares

OLS can be used in situations where the underlying model is linear in the parameters, but does not apply to many machine learning models, such as neural networks. In these cases, a more flexible approach is needed.

Nonlinear least squares (NLS) is an approach that can be used when the model is nonlinear, and it works using the same principles as OLS, i.e. by minimizing the residual sum of squares.

$$RSS = \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

But in this case, $\hat{y}_i = f(x_1, x_2, \dots, x_m; w)$ where f can be any nonlinear function of the m explanatory variables or features, which are denoted by x_i , and the

Artificial Intelligence Risk Certificate

correspondent parameters are denoted by w_i . Same for MSE.

Because the relationship between the features and the output could in principle take any form, it is often not possible to derive a set of closed form solutions to this minimization problem. Therefore, NLS usually uses a numerical approach to finding the optimal parameter estimates, and it proceeds with the following steps.

Firstly, begins with a set of initial values for the parameters, these could be either randomly generated or the best preliminary guess.

Secondly, evaluates the objective function (RSS or MSE)

Thirdly, modifies the parameter estimates and re-evaluate the objective function.

Fourthly, if the improvement in the objective function is below a pre-specified threshold, then stop looking and report the current parameter values at the chosen ones. If not, return to step 2.

The third of the above steps is the crucial aspect, and usually a gradient descent algorithm is employed.

7.1.3 Hill Climbing

A simple form of optimizer for estimating the parameters of a nonlinear model is hill climbing which involves starting with initial guesses for each parameter and then making small changes in both directions to each parameter one-at-a-time.

The aim is to maximize the value of an objective function, until no further improvement in its value is observed.

Hill Climbing is very straightforward because it does not require the calculation of derivatives, and therefore it can be applied to non-differentiable functions. It is also simple to implement and for this reason it is a.k.a. heuristic optimizer.

Some of the disadvantages are, as follows.

Firstly, of all the optimization techniques available, hill climbing is the most susceptible to getting stuck in local optima.

Secondly, convergence to the optimal solution can be very slow.

Thirdly, only one parameter at the time can be adjusted, meaning that it is easy for the algorithm to miss optimal parameter combinations, particularly for complex and highly interconnected models.

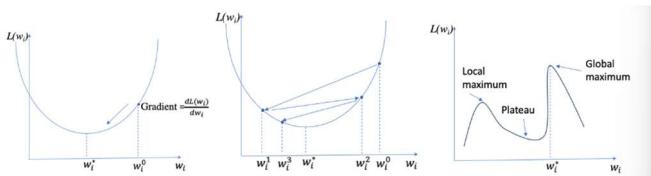
Given these limitations of hill climbing, an approach based on gradient descent instead usually forms the

Artificial Intelligence Risk Certificate

workhorse of parameter optimization for machine learning models.

7.1.4 The gradient Descent Method

In this method, the objective function, for example the residual sum of squares is minimize. Suppose that all the parameters to be estimated are stacked into a single vector w and the objective function in this case is known as the loss function and is denoted as $L(W)$. at each iteration, the algorithm chooses the path of steepest descent, slope, which is the one that will minimize the value of the loss function the most. The method works similarly when the maximum likelihood method is used, in which case the negative of the log-likelihood function is minimized.



Starting with an initial guess w_i^0 , the aim is to move towards the optimal value, w_i^* , which occurs where the loss function is minimized. The gradient of the function is given by its partial derivative with respect to the parameter, evaluated at the initial guess.

$$\left. \frac{\partial L(w_i)}{\partial (w_i)} \right|_{w_i = w_i^0}$$

Artificial Intelligence Risk Certificate

The guess of weight is then updated to a revised value w_i^1 according to the formula:

$$w_i^1 = w_i^0 - \eta \frac{\partial \mathcal{L}(w_i)}{\partial (w_i)}$$

Here, $\frac{\partial \mathcal{L}(w_i)}{\partial (w_i)}$, is the partial derivative of the loss function with respect to w_i and η is called the learning rate, usually in the range of 0-1.

We can also write this partial derivative of the loss function with respect to the weight w_i .

$$\frac{\partial \mathcal{L}(w_i)}{\partial (w_i)} = \frac{2}{N} \sum_{i=1}^N (y_i - \hat{y}_i) \frac{\partial \hat{y}_{in}}{\partial w_i}$$

Determining the gradient of the loss function with respect to the weight involves calculating the partial derivative of the output with respect to the weight.

Note that the adjustment to weight W_i takes place in the opposite direction to the gradient or slope, and hence the negative sign in the equation of w_i^1 .

When the gradient is positive, the estimate of the weight W_i will be reduced at the next step. At the same time, if the gradient had been negative, the current value of W_i would be too small, and hence it would be increased on the next step.

As previously described, the process of calculating the gradient and updating the weights continues until the value of the loss function can no longer be improved, indicating convergence has been achieved. To avoid the

Artificial Intelligence Risk Certificate

iteration going into infinite cycles, a maximum number of iterations is assigned to the algorithm. In machine learning parlance, each iteration, comprising calculating the loss function and gradient then adjusting the weights using the whole training data sample, is known as Epoch.

We usually use the entire training data sample and minimize the loss function with respect to all of it, which is known as batch gradient descent. A slightly less common alternative approach is to apply gradient descent to each data point at a time, individually selected at random from the training set. This is known as stochastic gradient descent, or applying it to subsets of the training data, known as mini-batch gradient descent.

The Benefit of the stochastic gradient descent is that it does not require the entire database to be loaded into memory simultaneously, which reduces the required computational resources compared with the batch approach. However, because updating the weight will take place after the algorithm see each new data point, the convergence on the optimal values will require more iterations and will be less smooth.

The parameter eta, η , is a hyperparameter that defines how much adjustment to weight take place. This hyperparameter must be chosen judiciously, in the sense that if it is too small, each iteration will yield only modest improvements in the loss function and will result in a slow movement towards the optimum, requiring many iterations in total. On the other hand, if η , is too large,

Artificial Intelligence Risk Certificate

the is the danger of overshooting and an erratic path towards the optimal w_i^* . The algorithm can overshoot the minimum, oscillate around it, or may not converge.

For Batch Gradient descent, eta is fixed a priori, but for stochastic gradient descent, the learning rate can be made a diminishing function of the number of epochs that have already occurred so that the weight updating slows down as the algorithm comes closer to the optimum. In other words, we could employ dynamic learning, which entails starting with a larger eta to get close to the optimal solution faster, but to then reduce eta as the learning proceeds to avoid overshooting, decay function of eta.

$$\eta_t = \eta_0 e^{-kt}$$

Or the inverse decay

$$\eta_t = \frac{\eta_0}{1 + kt}$$

Where the parameter k controls the rate of decay and t is an epoch.

Gradient descent is a neat technique that usually works well, but it can run into problems. One such issue occurs when the function does not have a single, global optimum but rather a series of local optima or an extended plateau away from the true optimum. In such instances the optimizer can get stuck at the local optimum or plateau and never reach the optimal solution w_i^* .

7.1.5 Illustration of the Gradient Descent Method

Although we known that the OLS estimator can be derived analytically, it might be useful for pedagogical purposes to illustrate how gradient descent would work if we had used it to derive the estimates for the regression of salary on experience. To keep complexity at a minimum, and focus on illustrating how the optimization works, assume that we already know $b_0 = 16.88$ and hence we only need to optimize over b_1

$$MSE(b_1) = \frac{1}{N} \sum_{i=1}^N (y_i - 16.88 - b_1 x_i)^2$$

We also know that the gradient of this function is given by:

$$\frac{\partial MSE(b_1)}{\partial b_1} = \frac{2}{N} \sum_{i=1}^N (y_i - \hat{y}_i) \frac{\partial \hat{y}_i}{\partial b_1} = -\frac{2}{N} \sum_{i=1}^N (y_i - \hat{y}_i) x_i$$

Suppose that we start from an initial value of 0.8 and we use a learning rate equal to 0.005. Using the initial value, we compute the gradient, which is equal to -48.66. Given the learning rate of 0.005 the changes that we need to apply to b1 is $-0.005 * -48.66 = 0.243$, because we move in opposite direction of the gradient.

Therefore, the new value parameter is $0.8 + 0.243 = 1.043$. After 5 iterations the algorithm converges to 1.059. If we calculate the gradient of this parameter value, it is approximately zero as we have reached the minimum point, and no further adjustments are needed.

7.1.6 Backpropagation

Although the previous discussion focused solely on one parameter, the gradients will be calculated for all weights at each iteration. Determining the optimal weights in a neural network model is particularly challenging because, even with a single layer, the output is a function of a function. It is like running a logistic regression on the output from another logistic regression.

Therefore, to use a technique such as gradient descent would require repeated use of the chain rule of differentiation.

The backpropagation algorithm involves starting on the right hand side of the neural network and then successively working backward through the layers to update the weights estimates at each iteration. This begins by calculating the errors, actual minus fitted values, for each target data point, then these errors are “assigned” to each of the weights in the layer before it. Gradient descent can then be applied to the weight to calculate improved values. The output layer error in the target is determined via a feedforward of the feature values with the updated weights, and the process continues again.

The derivatives are computed starting from the output layer and moving backward, with an application of the chain rule. The algorithm stops when convergence is achieved, i.e. when updating the weights no longer

Artificial Intelligence Risk Certificate

reduces the cost function. The key to back propagation is to consider each layer separately rather than trying to do all the computation in a single step because breaking it down in this way greatly simplifies the mathematics.

In summary, the steps to implement the backpropagation approach to learning the optimal ANN weights are to,

Firstly, generate initial guesses for all weights, including the biases.

Secondly, given these weights, feedforward the values of the inputs to calculate the values at each neuron and then finally the value of the outputs. This is done separately for each of the N data points in the training sample.

Thirdly, once the fitted values of the outputs are determined the error which is the difference between the network output and the actual value, can be calculated for each observation. For the 1st iteration, see step 4. If the residual sum of squares is below a particular threshold or has not improved much since the previous iteration, or the number of iterations has reached a pre-specified maximum value, stop and fix the weights at their current values. Otherwise step4.

Fourthly, During the backward pass, the gradient descent method is used to calculate improved values of the weights. In this process, the error is propagated through the network, and the weights are updated to

Artificial Intelligence Risk Certificate

minimize the loss function. Return to step 2 and run through a further iteration.

A solution is obtained after several iterations (epochs).

7.1.7 Computational Issues

There is a risk that the algorithm will fail to find a global minimum but will instead get trapped in a local optimal. Sometimes a momentum term is added to the optimizer, which increases the learning rate if the previous change in the weights and the current change are in the same direction but reduces it if they are in opposite directions. The benefit of such approach is that it speeds up convergence but at the same time reduces the probability of overshooting the optimal parameter values. Incorporating momentum involves a modification of the updating scheme discussed in the previous chapter.

$$w_i^{new} = w_i^{old} - \eta \frac{\partial E}{\partial w_i^{old}} + \mu |w_i^{old} - w_i^{older}|$$

Where w_i^{older} is the weight before w_i^{old} and Miu is the momentum rate, which can be chosen between 0 and 1. The parameter Miu controls how much of the previous weight change we will keep in the next iteration. This works by overshooting the target, which helps prevent the algorithm from getting stuck in local minima.

This problem is less critical in networks with many hidden layers, deep networks. However, the deep

Artificial Intelligence Risk Certificate

networks are plagued by another computational issue, the so called vanishing gradient. Backpropagation is an application of the chain rule, which entails multiplication of several derivatives (as many as the layers in the network). When the derivatives are numbers between 0 and 1, like for sigmoid, their product tends to become very small very quickly. An opposite problem is the exploding gradient, where the product of the derivatives becomes larger and larger. When one of these problems emerges, the only way to find the optimum is by using an extremely large number of small updates, which of course makes the learning very slow. Vanishing or exploding gradients are by far the most relevant problems for deep network structures such as convolutional and recurrent neural networks. Some solutions are:

An appropriate choice of activation function. In recent years, the use of Sigmoid as an active function for hidden layers has been abandoned in favor of the ReLu function that is less prone to the vanish gradient problem.

Batch Normalization, which consists of adding the “normalization layers” between the hidden layers where the features were normalized by structuring the mean and dividing by the standard deviation. Here, it is the new inputs originating from the hidden layers that are normalized.

Specific Network Architectures, in which some networks have been developed to be resistant to the vanishing or exploding gradient problem such as LSTM.

7.2 Maximum Likelihood

This is another estimation technique that can be used for nonlinear models as an alternative to NLS. Maximum Likelihood can also be used for linear models, in which case will give identical estimates for the intercept and slope parameters of the OLS if the random errors are independently normally distributed with zero mean and constant variance. To a certain extent, maximum likelihood is a more flexible framework that enables it to be used to estimate the parameters for a wide range of specifications if the distribution of the data generation mechanism is known. For example, is used on GARCH family, to forecast volatility. Is also used to estimate Logit and Probit, which like GARCH, are nonlinear models.

Logit model we have two possible outcomes, either 1 or 0 which might, correspond to customers not being granted a loan or being granted loan.

If F_i denotes the Logistic Function, and P_i denotes the likelihood, $P_i = F_i$. The probability distribution for Y_i given the data on the explanatory variables can be written as:

$$L(y_i) = F(y_i)^{y_i} (1 - F(y_i))^{1-y_i}$$

Because Maximum Likelihood works by selecting the parameters that maximize the chances of the training data occurring, we want the joint likelihood of all N data point taken together, not just a single observation. This joint likelihood, denoted as L , will be obtained by multiplying all the individual probability distributions together.

Artificial Intelligence Risk Certificate

$$L(y_i) = \prod_{i=1}^N F(y_i)^{y_i} (1 - F(y_i))^{1-y_i}$$

Where Π denotes that the functions are multiplied because the joint probability of all N data points is the product of the $F(y)$ across the positive outcomes (=1) in the training set multiplied by the product of the $(1 - F(y))$ across the negative outcomes (=0) in the training set, as long as they are independent. If y_i is 1, the i th function reduces to $F(y_i)$, and if it is zero, the i th function reduces to $1 - F(y_i)$.

It is easier to maximize the log likelihood function, $\log(L)$, than to maximize the likelihood function because the logarithmic transformation turns multiplication into a sum. The log-likelihood is obtained by taking the natural logarithm of the above expression.

$$\log(L) = \sum_{i=1}^N [y_i \log(F(y_i)) + (1 - y_i) \log(1 - F(y_i))]$$

Once the parameters that maximize this expression have been estimated, predictions can be constructed from the model by setting a threshold, Z , estimating the value of P_i using the equation above, and then specifying the category that observation i is predicted to belong to.

$$\hat{y}_i = \begin{cases} 1 & \text{if } P_i \geq Z \\ 0 & \text{if } P_i < Z \end{cases}$$

If the costs of being wrong are the same for both categories we might set Z to be 50%. But in other cases a different threshold is more useful. For instance in Loan Classification, the cost of assuming someone will pay

Artificial Intelligence Risk Certificate

and defaults is higher than assuming someone will not pay, a priori, and lose business due to that.

7.3 Overfitting, Underfitting and Bias Variance trade off.

We will never know the true process generating the data, and we will only have a sample of data upon which to select an appropriate model and estimate the parameters. Consequently, it is an empirical choice as to the size of model we estimate, which leads to the possibility that the model contains too many (overfitting) or too few (underfitting) parameters.

7.3.1 Overfitting

Every time a model is too large or excessively parameterized. A simple example is when a high-dimensional polynomial is used to fit a data set that is roughly quadratic. The most obvious sign of an overfitted model is that it performs considerably worse on new data. When building a model we have the training and validation data sets. The training set is used to estimate the model parameters, and the validation set is used to evaluate the model's performance on a separate data set. An overfitted model captures excessive random noise in the training set rather than just the relevant signal. Overfitting causes the false impression of an excellent specification because RSS of the training set

Artificial Intelligence Risk Certificate

will be very low. However, when applied to other data not in the training set, the model's performance will likely be poor, and the model will not be able to generalize well.

Overfitting is usually a more severe issue with machine learning than with conventional econometrics due to a larger number of parameters in the former. For instance, a standard linear regression model generally has a relatively small number of parameters. By contrast, it is not uncommon for neural networks to have several thousand parameters.

7.3.2 Underfitting

Occurs when the relevant pattern in the data remains uncaptured by the model. For instance, we might expect the relationship between the performance of hedge funds and their size to be quadratic. A linear model would not be able to capture phenomena such as, *for funds that are too small they would have insufficient access to resources with costs thinly spread, and funds that are too big may struggle to implement their strategies in a timely fashion without causing adverse price movements in the market*, and would estimate a monotonic relationship between performance and size, and so would be underfitted. A more appropriate specification would allow for a nonlinear relationship between the fund size and performance.

Artificial Intelligence Risk Certificate

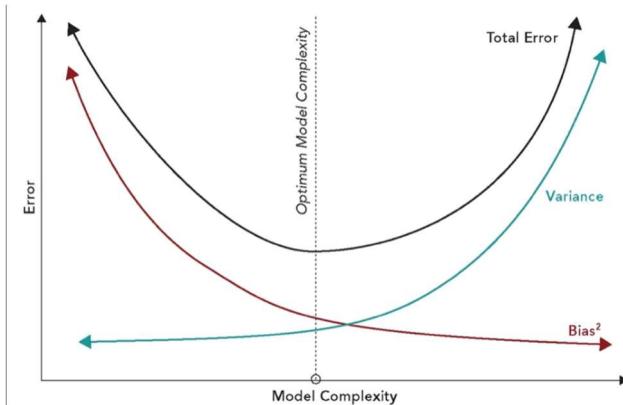
Failure to include relevant interaction terms, would be a further example of underfitting. It is clear that from these examples that underfitting is more likely to happen in conventional models than on machine learning models, where only a minimal assumption (such as smoothness) on the signal is imposed.

However, it is also possible for machine learning approaches, as well as econometric models, to underfit the data. This can happen either when the number or quality of inputs is insufficient, or if steps taken to prevent overfitting are excessively stringent. In such cases, the model fit to the training data will be poor, and there will be characteristic of the output variable that remain uncaptured by the model postulated. This will also likely lead to biased estimates of the parameters on the variables that are included in the model.

7.3.3 Bias variance Trade Off

The choice of the size of the machine learning model, which will determine whether the data are overfitted, underfitted or appropriately fitted, involves what is termed a bias-variance tradeoff. If the model is underfitted, the omission of relevant factors or interactions will lead to biased predictions but with low variance. On the other hand, if the model is overfitted there will be low bias but a high variance in predictions.

Artificial Intelligence Risk Certificate



An example, consider an analyst is assigned with the task of predicting the price at which a house will sell using the age of the house as a predictor. A simple approach would be to estimate a linear regression of the house prices on their ages.

$$\text{House price} = \hat{\beta}_0 + \hat{\beta}_1 \text{Age}$$

The regression line fitted uses a training sample of 201 points, picture below in Blue. Visibly, a linear regression is insufficient to fully capture the relationship between house prices and their ages. In fact, both very new and very old houses (with a historic value) seem to be more expensive, while the linear regression forces the predicted prices to be strictly decreasing in age. Fortunately, the linear regression model can accommodate polynomials of higher degree, as the linearity requirement concerns the parameters, not the explanatory variables. Therefore, the analyst could

Artificial Intelligence Risk Certificate

include a quadratic term to the regression above and estimate.

$$\text{House price} = \hat{\beta}_0 + \hat{\beta}_1 \text{Age} + \hat{\beta}_2 \text{Age}^2$$

The red line on the picture below is the estimated quadratic regression. This looks like a much more accurate representation of the data. Finally, the green line, is the result of fitting a polynomial of the night degree to the data. This highly complex can capture some highly nonlinear patterns in the data. However, it also captures noise, and it does not generalize well when we deploy the estimated model to predict unseen data, as it is possible to assess on the Panel B.

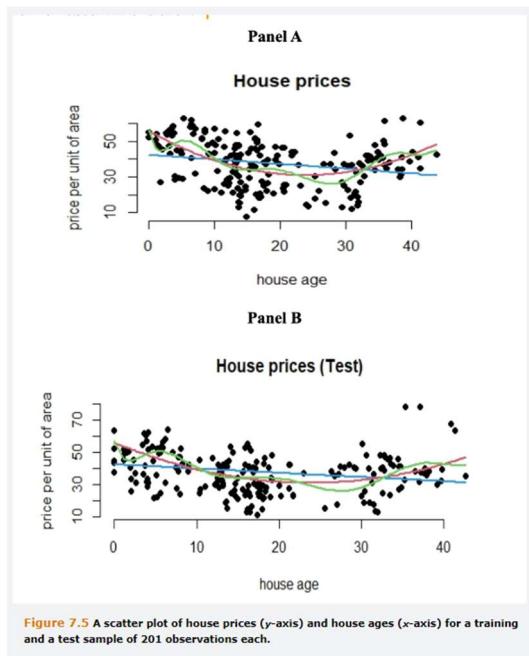
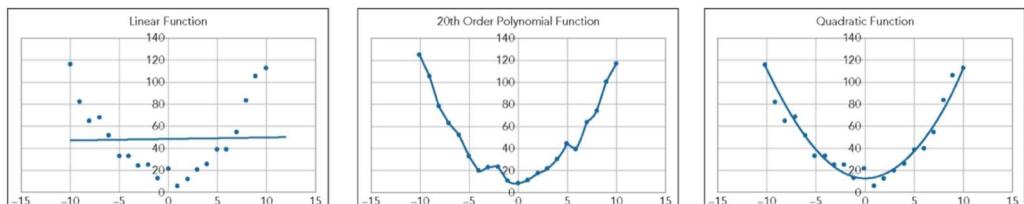


Figure 7.5 A scatter plot of house prices (y-axis) and house ages (x-axis) for a training and a test sample of 201 observations each.

Artificial Intelligence Risk Certificate

A highly complex model minimizes the MSE in the training sample but may fail to do so when deployed to new data. In the example above, the MSE on the training sample for Linear is 157.67, 130.34 for quadratic and 121.73 for the ninth order polynomial regression. However, when we use the training model to predict observations on the test data, the MSE is 156.37 for linear, 137.08 for quadratic and 139.82 for ninth order polynomial. Therefore, this simple example, the quadratic model strikes the right balance between fitting the training data accurately while also generalizing well to new data.

The picture below shows another example of how underfitting and overfitting can manifest themselves. Here, a single feature is plotted on the x-axis and an output variable on the Y-axis. The left panel shows a linear regression fitted to the data, which is clearly insufficient to describe the series and will give rise to predictions that are highly biased. The center panel shows the result of applying a high-order polynomial to the fit. This line contours perfectly with the training data set but is evidently overfitting. The right panel shows a quadratic polynomial, which has a better balance between over and under fitting.



Artificial Intelligence Risk Certificate

Overfitting is particularly likely and a severe problem with neural networks. Often there are several hidden layers and many nodes per layer which leads to an enormous number of parameters and the likelihood that there will be overfitting unless specific steps are taken to avoid it. One approach to limit the chances of overfitting is to carry out calculations for the validation set at the same time as the training data set. As the algorithm steps down the multi-dimensional valley, the objective function will improve for both data sets, but at some stage, further steps down the valley will start to worsen the value of the objective function for the validation set, while still improving on the training one. This is the point at which the gradient descent algorithm should be stopped because further steps down the valley will lead to overfitting the training data and therefore poor generalization and poor predictions for the test sample.

7.3.4 Prediction Accuracy vs. Interpretability

There is often an unrecognized trade off between a model's prediction accuracy and its interpretability. Machine learning models are often highly complex and heavily parameterized so that they have often been accused of being black boxes. More flexible models often deliver more accurate predictions, as they can generate a wider range of shapes for the function that maps the features to the outcome. Therefore, they can fit the highly complex and nonlinear patterns in real-world data. However, these models often lack interpretability.

Artificial Intelligence Risk Certificate

The linear regression model is often inadequate to model the complex nature of real-world relationships between the predictors and the target variable. Yet, its popularity among financial economist remains unchallenged, thanks to its ability to deliver an easy to understand relationship between the predictors and outcome that resonates with financial theory. Generally, less flexible but more interpretable models are preferred when the goal is to investigate causal relationships. In contrast, more flexible models tend to be the obvious choice when the goal is to make accurate predictions.

7.4 Regularization

The stepwise selection method discussed previously add or remove predictors to a regression with the aim of finding the combination that maximizes the model performance. An alternative is to fit the model on all m features using a regularization technique that shrinks the regression coefficients towards zero. Regularization can be used for standard linear regression models and many other machine learning models. The two most common regularization techniques are ridge regression and least absolute shrinkage and selection predictor, LASSO. Both work by adding a penalty term to the objective function being minimized. The penalty term is the sum of squares of the coefficients in ridge regression and the sum of absolute values of the coefficients in LASSO. Regularization can simplify models, making them easier to interpret, and reduce the likelihood of overfitting.

7.4.1 Ridge Regression

Suppose that we have a dataset with N observations on each of m features in addition to a single output variable y, and, for simplicity, assume that we are estimating a standard linear regression model with hats above the parameters denoting their estimated values. The relevant objective function, referred to a loss function, in ridge regression is:

$$L = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \dots - \hat{\beta}_m x_{1m})^2 + \lambda \sum_{j=1}^m \hat{\beta}_j^2$$

The first sum in this expression is the usual regression objective function, i.e. the residual sum of squares, and the second is the shrinkage term that introduces a penalty for large-slope parameter values. The parameter lambda controls the relative weight given to the shrinkage versus model fit, and some experimentation is necessary to find the best value in any given situation. Parameters that are used to determine the model but are not part of a model are referred to as hyperparameters. In this case, lambda is a hyperparameter and Beta(s) are model parameters.

7.4.2 LASSO

Is quite similar to ridge regression but takes the absolute value instead of the square.

$$L = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \dots - \hat{\beta}_m x_{1mi})^2 + \lambda \sum_{j=1}^m |\hat{\beta}_j|$$

Whereas there is an analytical approach to determining the values of the Bs for the ridge regression, a numerical

Artificial Intelligence Risk Certificate

procedure must be used to determine these parameters for LASSO because the absolute value function is not everywhere differentiable. Ridge Regression and Lasso are a.k.a L2 and L1 regularization due to the order of penalty terms in these methods. The key differences are:

Ridge Regression, L2, tends to reduce the magnitude of beta parameters, making them closer to, but not equal to 0. This simplifies the model and avoids situations in which for two correlated variables, a large positive coefficient is assigned to one and a large negative coefficient is assigned to the other. **LASSO, L1**, is different in the sense that it sets some of the less important estimates of Betas to zero. The choice of one approach over the other, depends on the situation and on whether the objective is to reduce extreme parameter estimates or remove some terms from the model altogether. LASSO is sometimes referred to as a feature selection technique, because de facto it removes the less important features by setting their coefficients equal to 0. As the value of Lambda, penalty, increases, more features are removed.

Ridge Regression and Lasso can be used with Logistic Regression. Maximizing the likelihood is equivalent to minimizing its negative. Therefore, to apply a regularization, we add lambda times the sum of the squares of the parameters or Lambda times the sum of absolute values of the parameters to the negative of the expression for the log-likelihood. Then the objective would be to find the value of the parameters that jointly minimize this composite of the negative of the log-likelihood and the sum of absolute values of the parameters.

7.4.3 Elastic Net

Is a combination of both loss functions

$$L = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_m x_{1m})^2 + \lambda_1 \sum_{j=1}^m |\hat{\beta}_j| + \lambda_2 \sum_{j=1}^m \hat{\beta}_j^2$$

By appropriately selecting the two hyperparameters, λ_1 and λ_2 , it is sometimes possible to obtain the benefits of both ridge regression and LASSO, by reducing the magnitude of some parameters and removing some unimportant ones entirely.

7.4.4 Regularization Example

Suppose that we are interested in running a regression of a time series of stock index returns on a set of Treasury yields for different maturities using the data from the PCA on an above example. As mentioned, the features are usually rescaled before using ridge or LASSO. But in this case, the magnitudes are similar and so, to keep the example simple, we will skip the rescaling process.

The figures from the table below, demonstrate that an OLS regression provides Beta parameters that are quite large in magnitude, with some Betas having a sign that is opposite to what is expected. This indicates that the features in this model are highly correlated. The ridge regression reduces the magnitude of the parameters, with the higher value of lambda shrinking them more. Some of Betas changed signs going from OLS to Ridge Regression. LASSO, on the other hand, reduces some

Artificial Intelligence Risk Certificate

coefficient values to 0. When $\lambda=0.1$, only one coefficient, apart from the intercept, is non-zero with a negative sign, indicating an inverse relationship between stock returns and treasury yields.

Conducting a regularized regression efectevely requires selecting the hyperparameter carefully. Often, this involves choosing a value of Lambda that produces a model that is easy to interpret while still producing accurate forecasts. The data can be split into training, validation and test dataset, in which the training is used to determine the coefficients for a particular value of lambda. The Validation set is used to determine how well the model generalizes to new data, and the test set is used to provide a measure of the accuracy of the chosen model. Sometimes, the simpler models produced using regularization generalize better than the original OLS linear regression model.

Table 7.4 OLS, Ridge, and LASSO Regression Estimates. An illustration of ridge regression and LASSO applied to a regression containing highly correlated features, with two different hyperparameter values.

Feature	OLS	Ridge, $\lambda = 0.1$	Ridge, $\lambda = 0.5$	LASSO, $\lambda = 0.01$	LASSO, $\lambda = 0.1$
Intercept	5.17	2.67	2.46	2.61	2.39
USTB1M	-23.22	-6.55	-2.00	-1.13	0
USTB3M	50.64	10.00	2.45	1.35	0
USTB6M	-37.64	-3.82	-0.51	0	0
USTB1Y	11.00	0.70	0.40	0	0
USTB5Y	-5.55	-1.75	-1.41	-1.22	-0.71
USTB10Y	9.13	0.57	-0.11	0	0
USTB20Y	-5.88	-0.08	0.36	0.14	0

7.5 Cross Validation and Grid search

The previous segments have concentrated on training a model using the dataset, validating it and then testing it. This section presents a more advanced technique for developing machine learning models which use the model fitting methods discussed so far repeatedly in a systematic manner as they search for models with better fit or to identify optimal values of parameters that will result in better models. The motivation for these techniques come from the necessity of making an efficient use of available data and to identify optimal values for hyperparameters such as the regularization term in regressions or the number of layers in neural networks, so computational costs can be optimized.

7.5.1 Cross Validation

Sometimes the Dataset is not sufficient allow for reasonably divide the dataset into training, validation and Test. For example, suppose that we have 100 instances. To split this into three sub-samples in the conventional way might entail a training sample of 70 instances and just 15 for validation and 15 for testing. In such circumstances, cross validation is a technique that can be employee to use the data more efficiently. It involves combining the training and the validation data into a single sample, with only the test data held back. This means that there is effectively not a separate validation sample, only a combined sample, which we

Artificial Intelligence Risk Certificate

now call the training sample. Then, this training data are split into equally sized sub-samples, with the estimation being performed repeatedly and one of the sub-samples being one of the sub-samples being left out each time.

The technique known as K-fold cross validation splits the total data available, N, into K samples and it is common to choose k=5 or 10, i.e. either 20% or 10%. If we define the sub-samples as K_i , $i=1.2.3.4.5$, the first estimation would be repeated with sub-samples K1 to K4, excluding K5. Next, the estimation would be repeated with sub-samples K1 to K3 and K5, leaving out k4. This is repeated 5 times, so we will have a 5x5 matrix, and will have K= validation samples, one for each iteration, and the model performance will be assessed based on the average of the validation sub-samples individual performances.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Iteration 1	Training	Training	Training	Training	Validation
Iteration 2	Training	Training	Training	Validation	Training
Iteration 3	Training	Training	Validation	Training	Training
Iteration 4	Training	Validation	Training	Training	Training
Iteration 5	Validation	Training	Training	Training	Training

Figure 7.7 Five-fold cross-validation

A larger value of K will imply an increased training sample size, which might be valuable if the overall number of observations is low. The limit as K increases would be K=N, which would correspond to having the same amount of folds as data points on the training data set. This situation is known as N-fold Cross Validation, jack-knifing, or leave-one-out-cross-validation (LOOCV).

Artificial Intelligence Risk Certificate

Cross-validation represents a very resourceful used of the data, but it also has disadvantages, such as using LOOCV will increase the size of the matrix to $N \times N$, which maximizes the sizes of the training and the validation samples but will be computationally expensive as the data are trained at each of the N th iterations. For nonlinear models, cross-validation exercises always increase the computational costs, which grow with K . For unordered data, the points allocated to each fold will usually be selected randomly. This implies that K -fold Cross-Validation cannot be used when the data have a natural order, such as in Time series, for instance. In this case, the appropriate frameworks would be to use a rolling window.

7.5.2 Stratified Cross Validation

When the overall sample is very small, there is a heightened risk that one or more of the training or validation sub-samples will, purely, by chance, comprise a set of datapoints that are atypical compared with the other sub-samples. Some classes or types of data will be overrepresented in the training sample and underrepresented on the validation one, and vice versa for the other classes or types of data. The use of Cross-validation will mitigate this issue to some extent, but an extreme case of it could cause even more severe problems. Specifically, if the output data are categorical but unbalanced between categories, it might be the case that there are no instances of one or more categories in one or more of the Sub-samples. For instance, suppose that we have $N=100$ sample observations on whether a car loan borrower will default, =1, or not, =0., with only

Artificial Intelligence Risk Certificate

ten costumers who defaulted. If we use cross-validation with K=10 it is likely to be the case that for some of the ten iterations, there will be no defaulting costumers in the validation sample, which will lead to a distorted evaluation of the models applied to the validation sample in those cases.

A potential solution would be to use stratified k-fold cross validation. In that case, instead of drawing k samples, without replacing, to comprise the validation data, the positive and negative outcomes would be sampled separately in proportion to their presence in the overall sample. So if, for example, k=10 and we have N=100, with 90 non-defaults and 10 defaults, we would select nine non-defaults and one default at random from the overall sample, so that each of the 10 folds would contain the same 9:1 ratio of the two classifications. This approach guarantees that each class is represented to an equal degree in all training and validation samples. An alternative way to deal with imbalanced classes is to generate further artificial instances of the minority, underrepresented class, using what is known as SMOTE, synthetic minority Over-sampling Technique. Or to use an asymmetric loss function that puts more weight on any incorrect predictions for this class.

7.5.3 Bootstrapping

Is a simulation technique where new data distributions are created by sampling with replacement from the original data. In this context, it would involve, for each iteration, drawing a sample of size N, the combined size of training and validation set, with replacement. It is highly likely that this sample will contain some instances more than once from the original sample and some instances will not appear at all, typically around a third of the original data will not be sampled in each iteration, value given by $\frac{1}{e}$. Those instances not appearing in the bootstrapped training sample, out-of-bootstrap data, then comprise the validation sample for that iteration. Many iterations would be performed, 10.000 or more and the results averaged over the iterations as for K-folds Cross-validation.

Cross-validation and bootstrapping represent more efficient ways to deal with the data than an arbitrary separation between training and validation sets, because, effectively, every observation appears in both the training and validation samples for different folds. Cross-validation is also straightforward to implement, but a disadvantage is that it might be computationally intensive if the model is complex, or the number of folds is large, or the total sample size is large. If there are K folds and H different possible values for the hyperparameter to consider, this will involve estimating KH separate models each for a sample of size N, which could be computationally infeasible. Bootstrapping with

Artificial Intelligence Risk Certificate

a large number of iterations will be even more computationally demanding, although recent advances in computing have made this less onerous than previously.

7.5.4 Grid Searches

The purpose of Cross-validation might be to determine the optimal value of a Hyperparameter. To do this, the researcher might use a grid search procedure, which involves selecting a set of possible parameter values. Suppose that the model under study involves specifying one hyperparameter, lambda. This might, for example, be the hyperparameter that controls the strength of the penalty term in the LASSO regularization. Assume that the researcher determines that a range of 0 to 100 is plausible to investigate, with a step size of 1. Using a 5-fold cross-validation to determine the most appropriate value of Lambda could be achieved following the next steps.

Firstly, separate the composite training sample into five randomly assigned sub-samples.

Secondly, set lambda to be 0.

Thirdly, combine four of the sub-samples and estimate the model under study on that composite and using the remaining sub-sample, calculate a performance measure, such as the percentage of correct classifications or the MSE.

Artificial Intelligence Risk Certificate

Fourthly, repeat the third step for the other four combinations of sub-samples.

Fifthly, calculate the average of the performance measure across the five validation folds.

Sixthly, add one to lambda, and if lambda < or = to 100, repeat steps 3 to 5, otherwise proceed to the Seventh step.

Seventhly, there will now be 101 performance statistics, one for each value of lambda, 0 to 100. Select the optimal value of Lambda corresponding to the best value of the performance statistic.

Eighthly, perform one final estimation of the model, this time using the entire training sample with the hyperparameter set to λ^* .

Constructing a Grid search framework has some drawbacks. A first issue is that the researcher may have no idea of even the scale of the hyperparameter, so a power scale might need to be used for Lambda, such as $10^1, 10^2, 10^3$ and so on. It would be possible to search over a coarse grid, including relatively few points over a wide range, but that could leave the best hyperparameter value from the grid search still a long way from the optimal value. Even getting close to the latter, using a more refined grid could impose a vast computational burden, but it is becoming less of a concern with the recent advances in computational technology. A second problem is that searching over too

Artificial Intelligence Risk Certificate

many grid points is another manifestation of overfitting and could lead to weaker test sample performance.

An alternative to grid search for hyperparameter selection would be to use random draws. This could reduce the computational time significantly and seems to work surprisingly well compared with more structured approaches. But if the researcher is unlucky, it could be that none of the randomly selected hyperparameter values come close to the optimum.

Appendix 7.A Genetic Algorithms

Another family of alternative approaches for machine learning model parameter optimization is based on genetic algorithms. These techniques, sometimes referred to as evolutionary algorithms, apply thinking from evolutionary biology that capture the fundamental aspects of how populations of animals evolve over the very long term according to Darwinian principles to be more resilient to hazards in their environment. Hence, they are loosely analogous to the genetic processes of reproduction and survival of the fittest. GAs have found widespread applicability in decision trees, support vector machines and neural networks. They can also aid feature selection. GAs treat each individual parameter as a chromosome, and combinations of parameters are people, with all possible parameter combinations considered to be the human population. The parameters are optimized over many generations. The process is

Artificial Intelligence Risk Certificate

initialized by establishing a first-generation population of randomly assigned parameter combinations. Then each generation will follow the given path.

Firstly, we define a fitness measure in terms of how good the optimization is for that combination of parameters, this could be for instance the RSS or the value of a likelihood function.

Secondly, each individual person, set of parameters, is entered into a mating pool and two people are selected at random to combine, but with the selection process such that individuals having high levels of fitness are more likely to be selected.

Thirdly, the genetic information, parameter values, is combined between these two individuals to create offspring, i.e. new parameter combinations. This is termed Crossover or Recombination.

Fourthly, random changes, a.k.a as mutations are also made to the parameter values at the time of combination to create more diversity in the population, estimates. A hyperparameter controls the extent of mutation.

This process continues until either a pre-specified number of generations has been reached, or the improvement in the fitness of the fittest individual, i.e. the best parameter combination so far, falls below a threshold and the solution has converged. Genetic algorithms constitute a very unstructured approach to parameter optimization, as they do not require gradient

Artificial Intelligence Risk Certificate

calculations and so they are particularly useful for very complex models where solution using gradient-based methods would be challenging. GAs are more robust with respect to local optima than more conventional techniques because they have the potential to search over a wider range of the parameter space, although they can still suffer from this problem when “long-term fitness” is dominated by short-term fitness and parameter combinations that maximized the former. GAs also do not require estimates of the derivatives of the loss function with respect to the parameters, which means that they can be employed in a variety of contexts and used with complex model structures.

In practice, encoding parameter values into strings, chromosomes, efficiently could pose challenges. If there does not exist an efficient encoding scheme, the length of the chromosomes could be large, and it would make GA to be an infeasible choice of optimization algorithm. GAs can demand considerable computational resources because the required number of iterations could be large, and for each iteration the fitness level must be calculated for the whole population of parameter combinations.

Questions and Answers Module 2 Chapter 7 from GARP – Model estimation

7.1 Define the term bias-variance trade off

The trade-off between choosing a model that is not complex enough to capture the true nature of the relationship between the outcome and the features, underfitting, and choosing a model that is too complex and will model the random noise in the training, overfitting, is generally known as the bias-variance trade-off.

7.2 Explain how the Gradient Descent Algorithm works.

Is designed to minimize an objective function. By first setting the starting values for the weights, the algorithm computes the initial value of the gradient function. Then the values of the weights are updated by going one small step in the direction of the gradient. Such a series of computations is repeated until convergence, i.e. the magnitude of the gradient is smaller than a prespecified threshold or the change in the objective function's value is within a prespecified threshold, or the maximum number of iterations is reached.

Artificial Intelligence Risk Certificate

7.3

A. Why does OLS minimize the sum of the squared residuals rather than the sum of the residuals

Some residuals will be positive and other negative, and if we sum these residuals, they will cancel out. Squaring the residuals ensures that all values are positive and they don't cancel out. In fact, the sum of the residuals is always zero for the optimal choice of parameter values.

B. When it is not possible to use OLS?

OLS cannot be applied when the target is a Binary variable instead of a continuous one.

7.4 Given that it is applicable to a wider range of model classes, why do we not typically use maximum likelihood estimation to estimate the parameters of linear regression models?

MLE, maximum likelihood estimation, is a more general and flexible method than OLS. But MLE is less tractable, in other words, it is inherently more complex to understand, and parameter estimation is also more complex. Both methods give the same optimal coefficients in linear regression, but OLS is a simpler model. Hence, it is usually preferred to use OLS when the model structure allows it.

Artificial Intelligence Risk Certificate

7.5 What is the role of the learning rate Eta in determining the weights in a neural network?

Eta is an hyperparameter that regulates the speed of adjustment of the weights. If ETA is too small, the learning can take considerable time, on the other hand, if Eta is to large, the algorithm may fail to achieve convergence. A common solution is to start with a large ETA but then let it decay as the learning process progresses. Popular choices are to subject the learning rate to exponential or inverse decay.

7.6 What is the vanishing Gradient problem?

Tends to occur, specially in deep networks, when many hidden layers are present. It is a consequence of the application of the chain rule for derivatives when the gradients are small numbers, between 0 and 1. When a long chain of small numbers is multiplied together, the gradient quickly becomes very close to zero, so it vanishes. An opposite problem is an exploding gradient, which is the result of multiplying large numbers.

7.7

A. Explain the benefit of regularization for regression models and how it works

Is good for Linear Regression and Machine Learning models, where there are several highly correlated features that make coefficient determination difficult. For instance, where coefficient estimates are offsetting

Artificial Intelligence Risk Certificate

one another to some extent, and are unstable in the face of minor changes in the specification. Regularization works by adding a penalty to the loss function, that penalizes the model for including large parameter values of either sign. Depending on the nature of the penalty term, i.e. LASSO or Ridge regression, some parameters are either shrunk toward zero or set to zero. This process will make the fitted model more parsimonious, which will usually improve its performance when applied to a validation or test data set.

B. Explain how LASSO and Ridge Regression Differ

The only difference between them is the Penalty term, in which, for LASSO, this takes the sum of the absolute values of the coefficients, the so called L1 measure, and ridge regression it is the sum of the squared coefficients, Betas squared. LASSO can set coefficients to zero, whereas ridge regression will simply push their values towards zero.

7.6 What is Overfitting, and how can it be detected?

Where a model fits not only to the signal in the training data set, but also to the noise. In such circumstances, although the training sample fit might be very good, the estimated model will not generalize well to the validation or test sets and so the test sample predictions would be poor.

8.0 Supervised Learning – Model Performance Evaluation.

Learning Objectives are:

Discuss metrics used to evaluate the performance of a model when the outcome variable is continuous.

Evaluate the performance of a classification model using a confusion matrix and related metrics.

Explain the relationship between true and false positive rates and how this trade off can be illustrated using the receiver operating curve (ROC).

8.1 Model Evaluation when the Output is continuous

When the data output is continuous the most common measure of the predictive ability of a model is the MSE.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Another measure is the RMSE, Square root of MSE.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$$

Because the forecast errors are squared, the MSE scales with the square of the units data, whereas the RMSE

Artificial Intelligence Risk Certificate

scales with the units of data and so is easier to interpret. Nothing prevents us from computing the MSE Over the Training sample. However, we are generally more interested in computing performance measures over the validation sample, to tune the model's parameters, or over the test sample, to evaluate the model's performance over unseen instances, than on the training sample. These are a.k.a out of sample forecasts. Often, models overfit to the training data sample and examining a model's accuracy in predicting observations that it has already seen and used to determine the parameters is not a true test of its performance. We will assume that the we are computing the measures of forecast accuracy over the test sample, which comprises observations not used on determining the parameter estimates or tuning the hyperparameters.

It is apparent that MSE measures how close on average the predictions are to the actual data and the square is taken to remove the sign from negative distances, which occur when the prediction overestimates the target value. Taking the squares ensures that the negatives and positives do not cancel out. The closer the predictions are to the target, the more accurate the model is. Therefore, among a set of different predictive models, the most accurate is the one with the lowest MSE in the test sample.

Going back to the previous example to predict salary.

$$\widehat{\text{salary}}_i = 16.88 + 1.06 * \text{experience}_i$$

Artificial Intelligence Risk Certificate

Consider the following test data sample, and be aware that this data was not used to compute the parameter estimations or to tune any hyperparameter.

Table 8.1: Illustration of calculation of MSE for a test sample of bank employees

Observation, <i>i</i>	Experience, <i>x_i</i>	Actual salary, <i>y_i</i>	Predicted salary \hat{y}_i	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	4	20.15	21.12	-0.97	0.94
2	6	21.48	23.24	-1.76	3.10
3	9	31.88	26.42	5.46	29.81
4	10	32.88	27.48	5.40	29.16
5	21	49.92	39.14	10.78	116.21
6	25	57.46	43.38	14.08	198.25
7	3	21.13	20.06	1.07	1.14
8	1	12.18	17.94	-5.76	33.18
9	2	19.47	19.00	0.47	0.22
10	7	28.66	24.30	4.36	19.01

First, we obtain the difference between the predictions and the observed values, then we square that difference. A MSE equal to 43.1 is obtained, which leads to a RMSE of 6.57, squarer root of MSE.

Although it remains the most widely used measure of performance in practice, MSE has some drawbacks. The most notable one is that, because of taking the squares of distances between the actual data and predictions, MSE overweights large deviations from the observed

Artificial Intelligence Risk Certificate

values. This implies that, when MSE is used to pick the best model, one that is generally very accurate but displays a few occasional large deviations from the observed value might be less preferable, than a model that is in general less accurate but where no prediction largely under or overestimates the true value. MSE is highly sensitive to Outliers in the test sample.

To overcome this limitation, we could use the MAE.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

RMSE and MAE are both unnormalized measures. If we consider the salary example above, this means that if we underestimate the salary by 2.000, this will contribute to the RMSE and MAE in the same way irrespective of the level of salary that we are predicting, same level of greatness. In other words, predicting a salary of 148.000 when the salary is 150.000 or predicting a salary of 10.000 when the salary is 12.000 is treated symmetrically. However, in some practical situations, the first error might be considered much less important than the second one, as it represents about 1% of total salary and the other one more than 10% of the salary.

To obtain the measure that allows to understand relative importance of error measures, we use the MAPE, mean absolute percentage error.

$$MAPE = 100 * \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Artificial Intelligence Risk Certificate

Where the multiplication by 100 is to express the metric as a percentage of the actual value. Similarly, the MSE and RMSE can be expressed in percentage of the actual by dividing the squared forecasted errors by the corresponding actual values.

Table 8.2: Illustration of calculation of MAE and MAPE for a test sample of bank employees

Observation, i	Experience, x_i	Actual salary, y_i	Predicted salary \hat{y}_i	$y_i - \hat{y}_i$	$ y_i - \hat{y}_i $	$\frac{y_i - \hat{y}_i}{y_i}$	$\left \frac{y_i - \hat{y}_i}{y_i} \right $
1	4	20.15	21.12	-0.97	0.97	-0.05	0.05
2	6	21.48	23.24	-1.76	1.76	-0.08	0.08
3	9	31.88	26.42	5.46	5.46	0.17	0.17
4	10	32.88	27.48	5.40	5.40	0.16	0.16
5	21	49.92	39.14	10.78	10.78	0.22	0.22
6	25	57.46	43.38	14.08	14.08	0.25	0.25
7	3	21.13	20.06	1.07	1.07	0.05	0.05
8	1	12.18	17.94	-5.76	5.76	-0.47	0.47
9	2	19.47	19.00	0.47	0.47	0.02	0.02
10	7	28.66	24.30	4.36	4.36	0.15	0.15

The calculation of MAE yields a 5.01. The MAPE is computed to 16%. Of all the forecast accuracy measures, MAPE is the most intuitive one. The figure of 16% can be interpreted as implying that the average forecast error is 16% of the actual salary. When used to evaluate alternative models, MSE, MAE and MAPE can all yield different model rankings. The choice of measure that should be considered as the most important one, ultimately depends on the problem at hand. For instance, a risk manager is likely to evaluate the expected loss in units of dollars. On the contrary, an index manager is interested in controlling the deviation of a portfolio's percentage return from the index is training.

8.1.1 An example of Continuous Variable Model Performance Comparison

Compare two different models employed to predict the house price per unit area. Model A is a simple linear regression and model B is a tree-based regression. The sample consist of 414 observations, equally split between training and Testing sample. The features are House age, the number of convenience stores in the area, and the longitude and latitude coordinates. The performance metrics calculated by estimating the model using the training sample and using it for predicting the output values in the test sample are reported below.

Table 8.3 Performance measures for two alternative models for house price prediction

	MSE	MAE	MAPE
Model A	94.85	6.56	18.25
Model B	77.35*	5.41*	14.45*

Notably, a nonlinear, tree-based regression is a more accurate predictor to the house price per unit are according to all performance measures in this case because the values on the second row are all smaller than the corresponding ones in the first.

8.2 Model Evaluation: Classification

These types of models tend to yield two types of predictions, a continuous value that would be a score or a probability and a discrete value, which is the predicted class. The continuous value is typically transformed into a discrete one using a threshold, Z. For instance, suppose that we had to predict whether a firm will pay a dividend the following year, positive outcome, or not, negative outcome. We could use sigmoid to obtain the predicted probabilities of dividend payment, \hat{y}_i , between 0 and 1. We then predict a positive outcome if $\hat{y}_i \geq Z$ and a negative if $< Z$. Although an obvious and popular choice for the threshold Z is 0.5, 50%, it does not need to be, and a different value might be more appropriate.

Although metrics such as the MSE could be used to evaluate the continuous prediction, we are often interested in evaluating the discrete prediction of the outcome derived from the probability. Similarly, it is often the case that continuous predictions are distilled into discrete values for use in a decision rule. For instance, a hedge fund analyst might have used a neural network model to make time-series forecasts of future returns but wishes to turn these into an automated trading rule. The approach would be the same as for predicted probabilities, establish a threshold and generate a binary dummy variable indicating whether the prediction is above or below the threshold. In this case, the threshold might be zero so that a buy signal, 1, is generated if the return is predicted to be positive and a sell signal, 0, if the forecast is negative. To evaluate discrete class predictions, different performance measures are necessary. When the outcome is a class, a common way to evaluate the model is through calculations based on a confusion matrix, which is a simple

Artificial Intelligence Risk Certificate

cross tabulation of the observed and the predicted classes. The main diagonal elements of the matrix denote cases where the correct outcome has been predicted, and off-diagonal elements illustrate all the possible cases of misclassification.

		Prediction	
		Firm will not pay	Firm will pay dividend
Outcome	No dividend	279 (27.9%) – TN	121 (12.1%) – FP
	Pays dividend	168 (16.8%) – FN	432 (43.2%) – TP

Based on the Four elements above, i.e., True Negative and True positive, when the outcome matches prediction, and False Negative False Positive, when the model fails to correctly predict the outcome, is possible to obtain the following performance metrics.

$$\text{Accuracy} = \frac{TN + TP}{TP + TN + FP + FN}$$

$$\text{Error rate} = 1 - \text{Accuracy} = \frac{FN + FP}{TP + TN + FP + FN}$$

Accuracy is the sum of the diagonal elements of the matrix divided by the total elements on the matrix, and it is straightforward to interpret as is simply reflects the agreement between the predicted and observed classes or equivalently, the percentage of all predictions that were correct. Similarly, the error rate is just 1- Accuracy, which is the proportion of misclassified elements. Despite both of them being very intuitive, they both

Artificial Intelligence Risk Certificate

ignore the type of error that has been made. In other words, they do not allow to distinguish between errors of Type I and Type II. Type I errors occur when an outcome predicted to be true is actually false, False Positive, and the Type II when the outcome prediction is false, but the actual outcome is True, False negative. In practical situations, the cost of committing each type of error is not the same. For instance, for a bank extending credit, misclassifying a borrower as solvent and then default, is typically more costly than the opposite, i.e. losing out on profit from a non-default client.

Additionally, error rate and accuracy are problematic as performance measures when the classes are largely unbalanced. For instance, suppose again that we want to classify a pool of borrowers as solvent, positive outcome, or insolvent, negative outcome. If 98% of the borrowers are solvent, a model that classifies 100% of the borrowers as solvent would have only have a 2% error rate, however this model would be practically useless.

To overcome such limitations, two other metrics are available.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

In which Precision is the number of correctly positives among all instances that have been classified as positive. In other words, precision is the estimate of the

Artificial Intelligence Risk Certificate

probability that a model is right when labelling an outcome as true. Recall, also known as Sensitivity, is the True positive rate, that is, the number of correctly classified positives over the total number of positives. It is also possible to compute the True negative rate, also known as Specificity, which is the proportion of negative outcomes that were correctly predicted as negative. Either Precision or Recall, can be more useful depending on the context. For instance, a bank prediction whether a borrower is solvent might be more interested in precision. Conversely, an analyst predicting whether a company would pay a dividend might be more interested in Recall.

A further measure known as F1 combines Precision with recall.

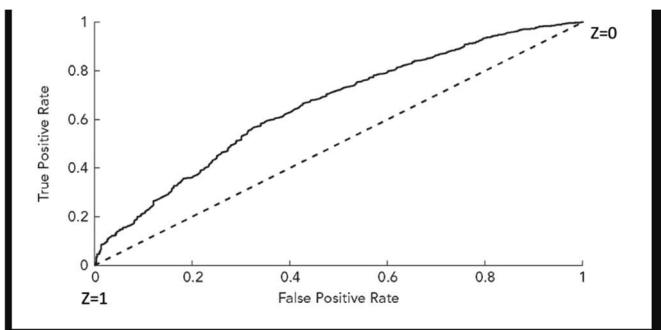
$$F1 \text{ score} = \frac{2 * Precision * Recall}{Precision + Recall}$$

The F1 Score will be bounded between 0 and 1, and it will be close to 1 if the model has both high precision and high recall, but a low score can arise from either poor precision, poor recall, or both. This is also sensitive to class imbalances.

More generally, there is a tradeoff between the True and False positive rates when setting the decision threshold, Z. As the True positive rate increases, the false positive rate also increases. For instance, taking the example above, we can identify more dividend paying firms at the cost of also misclassifying more non-dividend paying firms as dividend paying ones. In other words, we can

Artificial Intelligence Risk Certificate

increase the Recall at the cost of decreasing the precision and vice-versa. A way to illustrate this tradeoff is the Receiver Operating Curve, ROC, which plots the True and False-Positive rates against different choices of Z . Considering the same dividend example, if we had chosen $Z=0$, we would have classified all the firms as dividend paying. This implies that all positives are accurately classified but all the negatives are misclassified so both the true and false-positive rates are 100%. At the other end, if $Z=1$ all the firms are classified as non-dividend payment ones. Therefore, both the True and False-positive rates are 0%. For Values of Z between 0 and 1, the true positive rate increases when the false-positive rate increases.



The greater the area under ROC, AUC, the better the predictions from the model. A completely accurate set of predictions gives an AUC of 1. A value of AUC of 0.5 corresponds to the dashed line and indicates that the model has no predictive ability. Below 0.5 indicates that the model performs worse than randomly guessing.

Artificial Intelligence Risk Certificate

The formulae for the performance metrics described above implicitly assumed only two possible outcomes, 0 or 1. However, the formulae can be extended naturally to situations where there are several classes, such as when credit ratings are being predicted. For instance, in the multi-class case, the precision measure would be calculated by summing all the true positive classifications and dividing by the sum of all the true positive and all the true negative classifications. Likewise, the other performance evaluation metrics would be generalized in a similar fashion.

8.2.1 An Example of Model Evaluation: Classification

Suppose that we had to build a model to classify loans in terms of whether they turn out to default or repay. Several different models can be used to this end, and we want to test whether a single hidden layer feedforward neural network, outperforms a simple logistic regression. The Neural network contains ten units in the hidden layer and a logistic activation function.

From the figure below, the first panel represents the matrix for Sigmoid on the training sample and the second panel for the Sigmoid on the test sample. The third and fourth panel are the same but for Neural network. Interpreting or evaluating a neural network is harder than for more conventional econometric models. It is possible to examine the fitted weights,

Artificial Intelligence Risk Certificate

looking for very strong or weak connections or where estimates are offsetting, which could be indicative of overfitting. However, in the spirit of machine learning, the focus is on how useful the specifications are in classifying the validation sample.

Given that the same data and features have been employed for both the logistic regression and neural network, the results from the models can be compared. For simplicity, a threshold of 0.5 is applied, so that any predicted probability of default greater than or equal to 0.5, the fitted value is of default, and less than 0.5 is of no default.

Table 8.4 Confusion matrices for predicting defaults on personal loans

Logistic regression training sample			
Outcome		Prediction	
		No default	Default
	No default	400	11
Default		68	21
Logistic regression test sample			
Outcome		Prediction	
		No default	default
	No default	104	10
Default		38	15
Neural network training sample			
Outcome		Prediction	
		No default	default
	No default	406	5
Default		74	15
Neural network test sample			
Outcome		Prediction	
		No default	default
	No default	97	17
Default		33	20

Artificial Intelligence Risk Certificate

Table 8.5 Comparison of logistic regression and neural network performance for a sample of loans from the LendingClub

	Training sample (500 data points)		Test sample (167 data points)	
Measure	Logistic regression	Neural network	Logistic regression	Neural network
Accuracy	0.842	0.842	0.713	0.701
Precision	0.656	0.750	0.600	0.541
Recall	0.236	0.169	0.283	0.377

The performance summary measures show that, as expected, the fit of the model is somewhat weaker on the test data than on the training data. This result could be interpreted as slightly overfitting, and it might be worth removing some of the least empirically relevant features or applying a regularization to the fitted models. The confusion matrices shown on figure 8.4 show that the classifications from the two models are more divergent than the summary measures suggested. The Sigmoid predicts more defaults for the Training sample, whereas the neural network predicts more defaults for the test sample. Hence, the Sigmoid has a higher true positive rate but a lower true negative rate for the training data, whereas the situation is the other way around for the test data.

When comparing both models, there is very little to choose between them. On the training sample, their accuracies are identical, and although neural networks perform better in terms of precision, its recall is weaker.

Artificial Intelligence Risk Certificate

But when applied to the test sample, the Sigmoid does better on accuracy and precision, but worse on recall. Overall, these contradictory indicators illustrate the importance of fitting the evaluation metric to the problem at hand.

Questions and Answers Module 2 Chapter 8 from GARP – Model Performance Evaluation

8.1 Define MSE, MAE and MAPE

Performance measures used to rank models predictive capacity when the output variable is continuous. MSE is simply the average of the squared differences between the observed values and the model predictions. The square is taken to eliminate the sign of the difference, as under and overestimated are penalized symmetrically. MAE is the average of the absolute value of the differences between the observed values of the target variable and their model predictions. MAPE is a relative measure of performance that divides each difference between the actual and predicted value by the value of the actual observation.

Artificial Intelligence Risk Certificate

8.2 Specify the meaning of terms True Positive, TN, FP and FN

A TP is when the model and the outcome are both positive, a TN is when both are negative, a FP is when the model is positive, but the outcome is negative and FN is when the model is negative and the outcome is positive.

8.3 Explain the definitions of accuracy and error rate and how they are interrelated.

Accuracy is the percentage of correctly classified instances, and it is computed as the sum of true positives and true negatives over the total number of instance. The error rate is 1-accuracy, which is the % of incorrectly classified instances over the total number of instances.

8.4 Discuss the many limitations of accuracy and error rate as measures of model performance.

Accuracy and error rate do not account for the difference between type I and type II errors, which can be really relevant depending on the problem at hand. In addition, they fail to convey meaningful information when the classes are largely imbalanced. For instance if 98% of the instances were positive, classifying all instances as positive would only have an error rate of 2%, but the model would hardly be useful.

Artificial Intelligence Risk Certificate

8.5 Explain Precision and Recall

Precision is the estimate of the probability that a model will correctly classify an outcome as true. It is computed as the number of True positives over the number of total instances classifies as positive, TP and FP. Recall is the proportion of correctly classified positives among all positive instances, TP and FN.

8.6 Explain What ROC and AUC stand for and how they could be used in making lending decisions.

ROC stands for Receiver Operating curve and AUC is area under the curve. The ROC plots the true positive rate on the y-axis against the false positive rate on the x-axis and the points on the curve emerge from varying the decision threshold. The ROC shows the tradeoff between the true positive rate and the false positive rate when selecting Z.

The AUC shows pictorially how effective the model has been in separating the data points into clusters, with a higher AUC implying a better fit of the model, and so the AUC can be used to compare between models. An AUC of 1 would indicate a perfect fit, whereas a value of 0.5 would indicate an entirely random set of predictions and therefore the model would have no predictive ability.

One possible application of the ROC and AUC would be in the context of comparing models to determine whether a loan application should be rejected or

Artificial Intelligence Risk Certificate

accepted. A better model would be one with higher AUC of the test set.

8.7. A risk manager is evaluating the performance of two separate default prediction models for a sample of corporate loans in particular town. The models predict that the borrower will default or not default in the following year, which is then compared with the realized outcome as summarized in the following tables:

Model 1

		Predicted result	
		No default	Default
Actual result	No default	592	4
	Default	63	2

Model 2

		Predicted result	
		No default	Default
Actual result	No default	498	98
	Default	5	60

A. Calculate the True Positive and True Negative rates as well the precision and accuracy of each model.

Artificial Intelligence Risk Certificate

Model 1

$$Accuracy = \frac{2 + 592}{2 + 592 + 63 + 4} = 89.9\%$$

$$Precision = \frac{2}{2 + 4} = 33.3\%$$

$$Recall = \frac{2}{2 + 63} = 3.1\%$$

$$Error rate = 1 - \frac{2 + 592}{2 + 592 + 63 + 4} = 10.1\%$$

Model 2

$$Accuracy = 84.4\%$$

$$Precision = 38.0\%$$

$$Recall = 92.3\%$$

$$Error rate = 15.6\%$$

B. Comment on models' differences

Model 1 has higher accuracy, lower error rate, but model 2 has better precision and recall. Model 1 would not be preferred because it misses so many defaults. This shows that accuracy can be a misleading statistic when the data set is imbalanced.

9.0 Natural language processing

Also known as content analysis, text mining or computational linguistics, is one of the most exciting and fast developing applications of machine learning. NLP works with data with an unstructured, free text format to understand and analyze human language, both written and spoken. The U.S. SEC was an early adopter of NLP in its effort to detect accounting fraud.

Learning Objectives

Preparation of textual information for use in NLP models, construction of NLP models, a comparison of non-machine learning approaches to NLP and how NLP fit can be evaluated.

Discuss applications of NLP, describe the pre-processing steps for NLP, discuss the bag of words and n-grams approaches, explain how the Naïve Bayes classifier is used to categorize documents.

Illustrate how term frequency-inverse document frequency can be used to determine the appropriate weighting to assign to words in a document.

Describe and contrast different approaches to sentiment analysis.

Artificial Intelligence Risk Certificate

Uses of NLP

Recognition of specific words to determine the purpose of the message, categorization of a particular piece of text, determining the sentiment of the statement, assessing the readability of the document and evaluating the similarity of two documents.

More generally, NLP lies behind the operation of many everyday computational tool such as internet search engines and e-mail spam detection mechanisms. The main benefit of NLP over manual reading is the vastly superior speed with which NLP can complete a task. There is also a guarantee that all documents will be considered identically with no scope for biases or inconsistencies.

NLP is more challenging than modelling purely numerical data because textual data are often noisy with errors and ambiguities, particularly when dealing with information from social media or informal settings as opposed to formal communications from a firm or central bank. Is quite hard for the machine to identify sarcasm or even the tone of voice, as well abbreviations. Such issues mean that NLP fails to fully capture the essence of a particular document, although its power and flexibility are improving rapidly alongside advances in computational capacity for storing and processing information.

The most used NLP data sources are like, the corporate disclosures such as mandatory fillings, social media posts such messages on Twitter and so on, and news wires,

Artificial Intelligence Risk Certificate

newspapers and magazines. These documents might be available to download as a collection of pdfs using an Application Programming Interface (API) or by Web Scraping, which uses code such as python to find and download the documents in a batch. The collection of all documents that have been retrieved and are available for analysis by NLP is known as Corpus.

9.1 Data Pre-Processing

Prior to processing textual information statistically, the first stage is to convert and store the data in machine-readable raw text files if they are not already in that format either by removing text from pdf or if the files are web-scraped, removes any hypertext markup language, HTML. The HTML code contains tags such as “<p>” which indicates a paragraph. These have no content value and can be dropped. Sometimes, rather than complete removal, symbols such as emojis can be converted into a numerical representation in a process of text encoding and retained for later analysis. Various software tools exist that can identify emojis and replace them with their written definitions or numerical values.

The initial quality of the data will depend upon its source. Newswires or central bank communications will usually be drafted in a formal and error free language, but social media or bulletin board messages will likely contain numerous spelling errors, colloquialism and abbreviations. It will be necessary to edit the most

Artificial Intelligence Risk Certificate

common repeated instances of these to retain as much useful information as possible, for example, using a spell checker to guard against an artificial proliferation of separate words that were intended to be the same, “their and Thier”. On the other hand, spell checkers often change correct words, that they do not recognize or replace misspelled words wrongly. More straightforward, it is worth editing the documents to expand out contractions, so that negations are correctly defined.

Once the data have been initially cleaned, the next stage involves several pre-processing steps to ensure that the text is as amenable to accurate analysis as possible. The steps are:

Firstly, **Tokenize the passage**, meaning to separate the document into sentences at each period, question mark or exclamation mark. Then, each sentence is split into words, special symbols, and so forth. Any letters or words in capitals would all be modified to lower case.

Secondly, **stop word removal**, which are those that have no informational value but that are included in sentences to make them flow and so that they are easier to follow, such as has, a, the, also and so on. Precisely what constitutes a stop word will vary from one application to another depending on the purpose of it, because particular words may have no value in one context but convey useful information in another.

Thirdly, **replace with words with their stems**, a process also known as stemming, where words are replaced with

Artificial Intelligence Risk Certificate

their core or root forms to be treated equally in subsequent analysis. An example is that disappointing and disappointed would be replaced with disappoint.

Fourthly, **replace words with their lemmas**, also known as lemmatization, where words such as good, better, best are all replaced with good. This step is undertaken so that words expressing a very similar perspective on the topic are considered equally. For instance, the words good, better and best, all convey a positive sentiment and separating them would cause unnecessary complexity.

Fifthly, **feature extraction**, a process that involves turning the text processed using the previous steps into numeric vectors that can be analyzed by machine learning models, also known as text representation. It is an important and involved part of the process.



Figure 9.1 Preprocessing Steps

A few further notes noting the limitations and disadvantages of the process above. Although the removal of punctuation simplifies the task at hand, we should note that in many cases, by doing so, valuable context will be lost. For instance, the phrase, “Is the CEO performing well?” is written as a question and therefore, arguably, has little useful information content. Yet a

Artificial Intelligence Risk Certificate

textual analysis without the punctuation would conclude that the phrase expressed a positive view owing to the presence of the terms performing and well in the same sentence without a negation word. Relatedly, an exclamation mark at the end of a sentence may denote a phrase that is written in humor or sarcasm and therefore might have less information value than a similar sentence ended with a period.

Lowercasing can lead to inaccuracies or misclassification of words because an uppercase first letter is usually used to denote a proper noun, for instance Reading or reading. The errors caused by this can be addressed by labelling each word according to its class prior to lower casing, which is also known as part-of-speech tagging. In this case, Reading would be classified as a proper noun and reading would be treated as a verb.

Stemming and lemmatization are used so that similar words are treated the same as one another to simplify the analysis. However, again the simplification can lead to loss of nuance and degrees of positivity or negativity, such as the terms good and outstanding, which could be lemmatized to be the same. It might be the case that words are so common across documents that they have little value in classification. For instance, the term like might convey positive sentiment but has more than one meaning and will be ubiquitous (onipresente). Rather than seeking out such words individually for removal, the researcher can rank words by their frequency of occurrence across all documents and then drop, say, the top 5%.

Artificial Intelligence Risk Certificate

Once these steps have been undertaken, the remaining text segment can be subject to examination. Most straightforward NLP tasks treat the processed text as a bag of words, which mean that the ordering of words and any linkage between them are ignored to simplify the task. It is also necessary to examine the context in which a word is used to understand the intended meaning. Techniques like n-grams are used for this purpose.

9.2 NLP Models

Broadly we can separate the various techniques into two categories, heuristic approaches that primarily involve a direct analysis of the vector representations of the documents, and machine learning approaches that use a similar technique to those discussed in previous chapters, naïve bayes, support vector machines and neural networks.

9.2.1 Feature Extraction

Whether heuristic or ML models are adopted, the first stage is to transform the text into numerical values. The most common approach involves treating each sentence or document as a bag of words, where each word in a document is considered distinctly and equally. We begin with a vocabulary, which is a collection of all words to look for, and then we examine whether each word from

Artificial Intelligence Risk Certificate

the vocabulary occurs in the document. The simplest technique is the binary BoW, where we assign a value of one if the word appears and zero otherwise, storing the result into a vector. The standard BoW instead counts how many times each word appears and record these non-negative integers into a vector. Let W be a vector containing all the words in the vocabulary, which is of length $|W|$.

One possibility is that the vocabulary is imposed exogenously, it is created a priori and then applied to the document. Alternatively, we can combine every distinct word in the corpus of all documents under study and this long list will comprise the vocabulary.

For instance, suppose that we had constructed a small vocabulary of 15 words to classify predictions about market movements from bulletin boards that contains the following words (for simplicity at this stage ignoring all the pre-processing steps outlined above): $W = \{ \text{'high'}, \text{'low'}, \text{'buy'}, \text{'sell'}, \text{'stock'}, \text{'bond'}, \text{'profit'}, \text{'loss'}, \text{'gain'}, \text{'volatility'}, \text{'up'}, \text{'flat'}, \text{'down'}, \text{'rise'}, \text{'fall'} \}$. Suppose further that we have two documents available:

$d_1 = [\text{'I expect the stock market to rise, creating profit opportunities, offsetting yesterday's fall, and reaching a high of 1550 by lunchtime with a high of 1560 through the whole day'}]$

$d_2 = [\text{'We anticipate that the bond market will experience volatility but will finish flat or down'}]$

Artificial Intelligence Risk Certificate

we start off with the vocabulary W and create vectors of the same length for each document, where each element of the vector represents the number of times the word is mentioned in that document. In the document d1, high occurs two times, whereas stock, profit, rise and fall occur one, so the vector equals

$$v_1 = [2 \ 0 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 1]$$

Similarly, the vector corresponding to d2

$$v_2 = [0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 0 \ 1 \ 1 \ 0 \ 0]$$

It is evident that the vectors corresponding to each document are of the same length, irrespective of the number of words on the original document, which makes them amenable to quantitative analysis in the same way as we require there to be the same number of observations on each variable in the regression model. It is also clear from this illustration that the document word vectors will be sparse – that is, they contain a lot of zeros. This will make statistical modeling slow and inefficient. The vector representations of the documents can be collected into a matrix known as the document term matrix of dimensions $|W| * D$ where D is total number of documents in the corpus.

Artificial Intelligence Risk Certificate

9.2.2 Vector Normalization

If individual words are used repeatedly in the same document, that can cause problems with the analysis akin to the issues with outliers in econometrics. For instance, consider a bank that is trying to classify its customer feedback as either positive or negative and comes across the review, their service is just awful, bad people, bad rates and so on. This respondent's repeated use of the word bad could skew the analysis. To circumvent this issue, it is common to normalize the word vectors prior to analysis. Using what is known as the L2 Norm would involve creating vectors of length one via dividing each element by the square root of the sum of the squares of the elements

$$v_{1,norm} = \frac{v_1}{\|v_1\|^2}$$

$$\begin{aligned} v_{1,norm} &= \frac{v_1}{\|v_1\|^2} = \frac{[2 \ 0 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 1]}{\sqrt{2^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 1^2}} \\ &= \frac{[2 \ 0 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 1]}{2.83} = [0.71 \ 0 \ 0 \ 0 \ 0.35 \ 0 \ 0.35 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0.35 \ 0.35] \end{aligned}$$

This process has scaled all the elements in v_1 so that they lie between 0 and 1 and no individual words would dominate even if they were assessed repeatedly in a document. A similar scaling would be used separately for each document so that they are all of unit length.

9.3 Dictionary Comparison approaches

Are by far the most used way to analyze text in finance research. The dictionary is a pre-defined list of words that share a common characteristic. Typically, the objective is to use NLP to assess the sentiment expressed in a document by providing separate counts of the number of positive words and the number of negative words it contains, divided by the total number of words. Then the proportion of negative words is subtracted from the proportion of positive word to obtain a net sentiment score. Neutral words can be placed into a third category or simply ignored, along with stop words and any other having little relevant information content for measuring the documents tone. In example:

“Firm XYZ has just reported a year-on-year rise in earnings before tax of just 0.1%, disappointing investors, despite total sales growth in double-digits. The Company also highlighted that the previous safety worries with the new release had been resolved, which should underpin future growth. An analyst expressed relief, suggesting that there had been concerns that the accidents would have led to a decline in the firm’s share of this competitive market”

A dictionary of sentiment words that have already been classified under the positive and negative headings could be employed. Informally applying this approach to the sample above, we would classify the words positive and negative, as per the table below. Because there are a total of five positive words and four negative words, we

Artificial Intelligence Risk Certificate

might conclude that the sentiment of this fee is slightly positive. However, the example highlights some of the challenges of text mining because many of the negative words occur in counterfactual sentences explaining that things are better than feared, “would have led to decline”. These sorts of issues indicate that the research design requires meticulous construction, particularly where the sentence structure is formal or complex.

Table 9.1 Positive and negative word stems for sample newswire text

<i>Positive word stems</i>	<i>Negative word stems</i>
Rise	Disappoint
Grow (occurs twice)	Worry
Resolve	Concern
Relief	Decline

Dictionary approaches usually rely on existing lists of words because constructing a new dictionary from scratch would be very time consuming and would require examining many sample documents to be assured that most relevant words were included. For finance, Loughran and MacDonald, developed comprehensive dictionaries of words classified as positive, negative, uncertain, litigious, strong modal and weak modal, based on extensive examination of 10-k fillings. The research claims that the accuracy of this dictionaries is around 75% in terms of classifying Thomson Reuters new articles.

9.3.1 Advantages and Disadvantages of the Dictionary Approach

The advantages are its transparency and its easy and speedy implementation. Dictionaries also facilitate comparison across different corpuses if the same word lists are used in each case. However, it also entails several key limitations.

Firstly, **the methods are only as good as the words lists on they are built in**, If the lists are incomplete or include ambiguities or errors, classification accuracy might suffer.

Secondly, **the dictionaries were designed for the analysis of formal documents**, and they are likely to perform much less well when used for the analysis of bulletin board posts or social media.

Thirdly, **dictionaries are only available for a narrow range of subjects**. If the researcher did not want to assess a document's tone but something else, the dictionaries are unlikely to exist ex ante. Creating new dictionaries for other applications is likely to be infeasibly time consuming, in which case using a machine learning technique would be preferable.

Fourthly, **negations, and ambiguous words**, can often confound dictionary applications.

9.3.2 N-Grams

Despite its intuitive appeal and simplicity, a major limitation of the basic BoW is that it treats each word independently, irrespective of where they occur in the document, and simply counts the number of times they occur. However, there are sometimes pairs or groups of words with a specific meaning when placed together that need to be consider rather than individually.

Moreover, negation words are frequently used in sentences to reverse the meaning of a word, for example, not good, nothing useful, neither up nor down. Negation words can turn an otherwise positive sentence into a negative one, or less cleanly, turn a negative sentence into a slightly positive one. Ignoring these negation words could severely compromise the accuracy of any analysis of a document.

A technique for dealing with these issues is to use “n-grams” where n, is an integer equal to one or more, contiguous words in a document are treated as if they were a single word. A single word is sometimes known as “unigram”, an n-gram of two words is a bi-gram and so on. For example, if the document is “Credit spreads narrowed today”, the bi-grams words would be “credit Spreads”, “Spreads Narrowed”, “Narrowed Today”. When the BoW approach is applied in the context of n-grams, it is known as the bag of n-grams (BoN).

9.3.3 Term Frequency-Inverse document Frequency

Documents will typically comprise a set of words that are commonly used plus others that are rarer. Distinguishing between the two can be a useful way to analyze the information more fully than basic words count. For example, if our corpus comprised a set of sell-side analyst recommendations, words such as increasing, and growth are likely to feature frequently in many of the documents. But other terms such as phenomenal and spectacular probably occur less often and therefore, when they do, we should assign a higher priority to the sentiment they convey. Calculating the term frequency-inverse document frequency, achieves this by assigning higher weights to words that occur frequently in a particular document but rarely in the corpus. On the other hand, if a word occurs very commonly in many documents, whether it appears in document j is of little value in helping us to classify the document.

We define the term frequency, TF of i_j , as the ratio of the number of times a particular term i appears in a document j , W_{ij} , to the total number of terms in document j , $|V_j|$, counting each occurrence separately. So, if a given word appears three times in a particular document, then three is added to the numerator for that word.

$$TF_{ij} = \frac{w_{i,j}}{|v_j|}$$

Artificial Intelligence Risk Certificate

Where $i=1, \dots, |W|$, and there are D documents in total, $j=1, \dots, D$. We normalize the frequency of occurrence of a word in the document by the total number of words in the document, because a particular word will have more importance in a short document than a long one.

The inverse document frequency for term i, IDF_i , is the natural logarithm of the ratio of the total number of documents in the corpus, D, to the number of those documents containing term i.

$$IDF_i = \ln\left(\frac{D}{\sum_j d_{i,j}}\right)$$

Where $d_{i,j}$ is a dummy variable taking the value of 1 if the document j contains word i and zero otherwise.

We can then determine an appropriate weight to apply to each term i in document j, Wi,j

$$(TF - IDF)_{i,j} = TF_{i,j} * IDF_i$$

Although IDF for any word does not vary by document, TF varies from document to document because it is based on how many times a word appears in a particular document. Example:

“Suppose that we are interested in identifying the most relevant words for characterizing a set of messages written by retail investors on a message board and we have the following four posts.

D1= This disappointing stock led me down

D2= Disappointing earnings, down for this stock

Artificial Intelligence Risk Certificate

D3= Earnings down for this month

D4= Fed hikes rates this month.

We first calculate the inverse document frequencies for each word as the log of the number of documents in the corpus, 4, divided by the total number of documents containing that word. Note that these do not vary by document, so we calculate the total number of times each word appears across the entire corpus.

Word	Documents containing word	IDF	Word	Documents containing word	IDF	Word	Documents containing word	IDF
disappointing	2	0.69	for	2	0.69	month	2	0.69
down	3	0.29	hikes	1	1.39	rates	1	1.39
earnings	2	0.69	let	1	1.39	stock	2	0.69
fed	1	1.39	me	1	1.39	this	4	0.00

The term frequencies are the number of times each word appears in each document divided by the number of words in that document.

Document, <i>j</i>		Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	<i>N</i>
<i>d</i> ₁	word	this	disappointing	stock	let	me	down	6
	$TF_{i,1}$	0.17	0.17	0.17	0.17	0.17	0.17	
	$(TFIDF)_{i,j}$	0.00	0.11	0.11	0.22	0.22	0.05	
<i>d</i> ₂	word	disappointing	earnings	down	for	this	stock	6
	$TF_{i,2}$	0.17	0.17	0.17	0.17	0.17	0.17	
	$(TFIDF)_{i,j}$	0.11	0.08	0.05	0.11	0.00	0.11	
<i>d</i> ₃	word	earnings	down	for	this	month		5
	$TF_{i,3}$	0.20	0.20	0.20	0.20	0.20	0.20	
	$(TFIDF)_{i,j}$	0.14	0.06	0.14	0.00	0.14		
<i>d</i> ₄	word	fed	hikes	rates	this	month		5
	$TF_{i,4}$	0.20	0.20	0.20	0.20	0.20	0.20	
	$(TFIDF)_{i,j}$	0.28	0.28	0.28	0.00	0.14		

Artificial Intelligence Risk Certificate

*For instance, this appears across all documents, so TFIDF is 0, in d1 because is TF=0.17, IDF=ln(4/4)=0 and TFIDF=0.17*0=0.*

Although no single word appears more than once in any of the four documents, the TF-IDF for each word shows considerable variation because the total number of times each word is used varies considerably. This means that a particular word, "down", will have different TF for a document of different lengths. Also, "this", has no discriminatory power across documents and therefore has a TF-IDF value of zero. Lower TF-IDF means small discriminatory power and higher TF-IDF means high discriminatory power."

It is worth being aware that there are several common variations on the TF-IDF formula, and different versions are sometimes implemented in software libraries to avoid confusion. For instance, sometimes the frequency in the TF formula is replaced with a binary dummy for whether the word is in the document or not, or W_{ij} in the TF formula is replaced by $\ln(1+W_{ij})$.

9.4 Machine learning approaches

The use of dictionaries to classify documents does not involve any learning. An alternative approach would be to use a sample of documents that have already been classified by a human. For instance, suppose that these were announcements as the example above in italic, labeled as positive, neutral and negative. Then, an algorithm would be used to learn from these documents the words most strongly associated with positive sentiment and the words most strongly associated with negative sentiment. The models developed in this way can then be used to classify other, unlabeled, documents that the machine has not seen.

Such an approach can be designed in a bespoke fashion, tailored to each specific application. But labelling a sufficiently large sample of announcements that the algorithm can learn from could be extremely time consuming. Hence, this collection of techniques is widely implemented in finance than the previous dictionary approach. Such techniques are becoming more popular, now aided by recent advances in computing and advance machine learning algorithms, such as deep learning. Once the documents are prepared and then frequency vectors constructed as above and labeled, the same machine learning classification techniques as previously described before, can be employed, including the naïve Bayes classifiers, support vector machines, sigmoid, or neural networks. Naïve Bayes is sometimes a.k.a a generative classifier because it can create new classifications for unclassified documents, whereas

Artificial Intelligence Risk Certificate

other approaches such as Sigmoid are discriminative classifiers. The latter can select the features, words in NLP applications, that provide the sharpest characterization between the categories, whereas Naïve Bayes assumes that each feature, word, has an equal role in determining the outcome.

Although many of the previously discussed classification algorithms were presented on the context of binary choice, they can usually be extended to cover the situation where there are more than two possible cases, such as positive, negative or neutral sentiment. In the case of logistic regression, this can be extended to a multinomial logit model, and support vector machines can be adapted so that the data are separated into K clusters by K-1 hyperplanes. All the data organization and pre-processing steps would proceed as above and then the ML technique can be applied as discussed previously.

9.4.1 The Naïve Bayes Classifier

Suppose that a retail bank is concerned that its customer service department is providing a poor experience, and it wants to investigate this via feedback forms that clients completed on its website. The bank has a small sample of 8 completed forms, and an employee has read each of them, classifying the feedback as either good, bad or indifferent. The employee has also created a word bank (Vocabulary) of four words that they believe will

Artificial Intelligence Risk Certificate

provide a useful characterization of the service level. For instance, the words might be slow, inefficient, helpful and great. The employee has created a column vector for each costumer's feedback. Based on whether each of the four words appears in their statement.

Table 9.2: Data for customer service naïve Bayes example

	Customer number							
	1	2	3	4	5	6	7	8
Word 1	0	1	0	0	0	0	1	1
Word 2	1	1	0	0	1	0	1	1
Word 3	0	1	1	0	1	1	0	0
Word 4	1	0	1	1	1	1	0	1
Label	Good	Bad	Good	Good	Indifferent	Good	Indifferent	Bad

Suposse also that we have an unlabeled statement, and we wish to use NLP to classify it. This new statement includes words 1 and 2 but not words 3 and 4, and we want to determine how this statement should be classified.

The Naïve bayes approach is based on a straightforward application of Bayes rule, which calculates the probability of an event conditional upon the value of another variable related to that event. It is termed “naïve” because it assumes the words in each document are **independent from one another, or more generally, that the features are independent of one another in machine learning application.** It is popular due to its simplicity and solid classification accuracy in practice.

Artificial Intelligence Risk Certificate

We define C as a set of classes, C1, C2, Ck, and suppose that we have a corpus from which we have labeled by hand a subset of N of the documents, D, d1,d2,dN.

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Here $P(c|d)$ is called the posterior probability that a document belongs to class c, $P(c)$ is known as the prior probability, or the unconditional probability of each class, and $P(d|c)$ is the conditional probability or likelihood of the document. $P(d)$ is the predictor prior probability, or the probability of the predictor variables.

The classification problem is to choose the most likely of the K classes that a specific unlabeled document belongs to. We do this by calculating the probability of the document belonging to each class and then assigning the document to the class with the highest probability, which in technical terms is known as the “maximum a posteriori class”. For convenience, the denominator $P(d)$ can be eliminated because it does not change by class and hence it effectively cancels across classes. We can therefore write an equation for the selected class, \hat{c} :

$$\hat{c} = \operatorname{argmax}_{c \in C} P(d|c)P(c)$$

We can express the document d equivalently using its words, w1 to wN and rewrite the preceding equation.

$$\hat{c} = \operatorname{argmax}_{c \in C} P(w1, \dots, wN|c)P(c)$$

Artificial Intelligence Risk Certificate

The expression $(w1, \dots, wN|c)$ is the conditional joint probability that all these words would have appeared in that class. Given the independence assumption that underpin the naïve bayes approach, we can write joint probability as a product of the marginal probabilities and obtain.

$$\hat{c} = \operatorname{argmax}_{c \in C} P(w1|c)P(wN|c)P(c)$$

$P(c)$ can be calculated easily for each class c_i as the number of documents labeled as belonging to that class divide by the total number of documents:

$$P(c_i) = \frac{n(c_i)}{N}$$

9.4.2 The Naïve Bayes Example

Returning to the example above.

The features are the four words, while the responses or outputs are the three classes for the quality of the bank's customer service, which we can annotate as $C1=$ Good, $C2=$ bad, $C3=$ indifferent. The fist stage is to calculate the unconditional probabilities of each class. There are a total of eight costumers with four of them rating the service as good, two as bad and two as indifferent.

Therefore, the unconditional probabilities are:

$$P(c1) = \frac{4}{8} = 0.5; P(c2) = \frac{2}{8} = 0.25; P(c3) = \frac{2}{8} = 0.25$$

Artificial Intelligence Risk Certificate

Table 9.2: Data for customer service naïve Bayes example

	Customer number							
	1	2	3	4	5	6	7	8
Word 1	0	1	0	0	0	0	1	1
Word 2	1	1	0	0	1	0	1	1
Word 3	0	1	1	0	1	1	0	0
Word 4	1	0	1	1	1	1	0	1
Label	Good	Bad	Good	Good	Indifferent	Good	Indifferent	Bad

Next, we want to calculate the conditional probabilities o words give the class.

Table 9.3 Count of the number of customers using each word given their classification.

	Class		
Word	1 (Good)	2 (Bad)	3 (Indifferent)
1 ($w_1 = 1$)	0/4	2/2	1/2
2 ($w_2 = 1$)	1/4	2/2	2/2
3 ($w_3 = 0$)	2/4	1/2	1/2
4 ($w_4 = 0$)	0/4	1/2	1/2

Focusing first on the reviews classed as good, we need to find out how many statements for class C1, have used the word 1, which is zero. Next, how many people from class 1 used word 2, which is 1 person out of four statements, and so on.

Artificial Intelligence Risk Certificate

$$P(w1 = 1|c1) = \frac{0}{4} = 0$$

$$P(w2 = 1|c1) = \frac{1}{4} = 0.25$$

$$P(w3 = 0|c1) = \frac{2}{4} = 0.5$$

$$P(w4 = 0|c1) = \frac{0}{4} = 0$$

Note that we are interested in $P(w3 = 0|c1)$ rather than $P(w3 = 1|c1)$, because the scenario given in question states that **word 3 is not presented.**

Now we can calculate

$$\begin{aligned} P(w1|c1)P(w2|c1)P(wn|c1)P(c1) &= \\ &= 0 * 0.25 * 0.5 * 0.0 * 0.5 = 0 \end{aligned}$$

We repeat the process for the other two classes

$$P(w1 = 1|c2) = \frac{2}{2} = 1$$

$$P(w2 = 1|c2) = \frac{2}{2} = 1$$

$$P(w3 = 0|c2) = \frac{1}{2} = 0.5$$

$$P(w4 = 0|c2) = \frac{1}{2} = 0.5$$

$$\begin{aligned} P(w1|c2)P(w2|c2)P(wn|c2)P(c2) &= \\ &= 1 * 1 * 0.5 * 0.5 * 0.25 = 0.0625 \end{aligned}$$

Artificial Intelligence Risk Certificate

$$P(w1 = 1|c3) = \frac{1}{2} = 0.5$$

$$P(w2 = 1|c3) = \frac{2}{2} = 1$$

$$P(w3 = 0|c3) = \frac{1}{2} = 0.5$$

$$P(w4 = 0|c3) = \frac{1}{2} = 0.5$$

$$\begin{aligned} P(w1|c3)P(w2|c3)P(wn|c3)P(c3) &= \\ &= 0.5 * 1 * 0.5 * 0.5 * 0.25 = 0.03125 \end{aligned}$$

Note that the three class likelihoods, conditional probabilities, will not sum to one because we have removed the denominator $P(d)$. We can turn these numbers into probabilities by.

$$P(c1|w1 = 1, w2 = 1, w3 = 0, w4 = 0) = \frac{0}{(0 + 0.0625 + 0.03125)} = 0$$

$$\begin{aligned} P(c2|w1 = 1, w2 = 1, w3 = 0, w4 = 0) &= \frac{0.0625}{(0 + 0.0625 + 0.03125)} = \\ &= 0.67 \end{aligned}$$

$$\begin{aligned} P(c3|w1 = 1, w2 = 1, w3 = 0, w4 = 0) &= \frac{0.03125}{(0 + 0.0625 + 0.03125)} = \\ &= 0.33 \end{aligned}$$

Therefore, we would classify this unlabeled statement as most likely belonging to group 2 (“bad”) because it has the highest posterior probability.

We should note two potential issues with the above approach:

Artificial Intelligence Risk Certificate

Firstly, **the probabilities of a specific word appearing in a particular document could be small, and for long and diverse documents we could end up multiplying together several extremely small numbers.** This could cause an underflow problem for the computer, so instead it is common to take the natural logarithm of the probabilities. Because $\ln(AB)=\ln(A)+\ln(B)$, the transformation will mean that instead of multiplying raw probabilities we are summing the log of probabilities, which will be computationally more manageable. Taking logs will not change the results because the class with the highest probability will be the same as the class with the highest log probability.

A further, and often fatal issue, **arises when a specific word does not appear in the labeled sample for a particular category, because then the probability of that category having the word will be calculated exactly as zero.** This will render the probability of the category to be zero for that document irrespective of whether it contains many other words that are usually associated with that category, like our example for c1 word 1, that has 0 so it transforms the overall P to zero. More generally, even with a larger training sample and substantial vocabulary, **if one client uses a positive word in an otherwise negative review, like being sarcastic, and if there were no reviews labeled negative in the training set that contained that word, the probability of that review being set as negative would be set to zero.** The probability of the review being negative would still

Artificial Intelligence Risk Certificate

be zero even if it contained many other words that were usually associated with negative reviews.

To avoid this second situations, an adjustment known as smoothing, is sometimes made to the dataset. This adds a positive integer, lambda, to all the counts in the conditional probability matrix, which will ensure that they are all non-zero and therefore all the probabilities will be non-zero. Laplace smoothing is most used, which sets Lambda=1. We can see the effect of this reconsidering the above example and adding one to the count of each outcome.

Table 9.4: Customer data with one dummy observation added for each outcome for each class

	Customer no							
	1	2	3	4	5	6	7	:
Word 1	0	1	0	0	0	0	1	
Word 2	1	1	0	0	1	0	1	
Word 3	0	1	1	0	1	1	0	
Word 4	1	0	1	1	1	1	0	
	Good	Bad	Good	Good	Indifferent	Good	Indifferent	B

Artificial Intelligence Risk Certificate

8	9	10	11	12	13	14
1	1	0	1	0	1	0
1	1	0	1	0	1	0
0	1	0	1	0	1	0
1	1	0	1	0	1	0
Bad	Good	Good	Bad	Bad	Indifferent	Indifferent

In which 9-14 are just columns full of zeros and 1 to not disrupt the conditional probability, but to ensure that nonzero probabilities.

Table 9.5 Modified count of the number of customers using each word given their classification after Laplace smoothing

Word	Class		
	1 (Good)	2 (Bad)	3 (Indifferent)
1 (=1)	1/6	3/4	2/4
2 (=1)	2/6	3/4	3/4
3 (=0)	3/6	2/4	2/4
4 (=0)	1/6	2/4	2/4

Unconditional probabilities need to be recalculated:

$$P(c1) = \frac{6}{14}; P(c2) = \frac{4}{14}; P(c3) = \frac{4}{14}$$

Then we calculate

Artificial Intelligence Risk Certificate

$$P(w1|c1)P(w2|c1)P(wn|c1)P(c1) = \\ = 1/6 * 2/6 * 3/6 * 1/6 * 6/14 = 0.0020$$

$$P(w1|c2)P(w2|c2)P(wn|c2)P(c2) = \\ = 3/4 * 3/4 * 2/4 * 2/4 * 4/14 = 0.0402$$

$$P(w1|c3)P(w2|c3)P(wn|c3)P(c3) = \\ = 2/4 * 3/4 * 2/4 * 2/4 * 4/14 = 0.0286$$

And finally, the probability by normalizing the values:

$$P(c1|w1 = 1, w2 = 1, w3 = 0, w4 = 0) = \frac{0.0020}{(0.0020 + 0.0402 + 0.0286)} \\ = 0.029$$

$$P(c2|w1 = 1, w2 = 1, w3 = 0, w4 = 0) = \frac{0.0402}{(0.0020 + 0.0402 + 0.0286)} = \\ = 0.583$$

$$P(c3|w1 = 1, w2 = 1, w3 = 0, w4 = 0) = \frac{0.0286}{(0.0020 + 0.0402 + 0.0286)} = \\ = 0.338$$

We can see that the probability that the unlabeled review is good has increased from zero to 2.39%, with the probability of the review being bad falling while the probability of it being indifferent has gone up slightly. Smoothing is a shrinking technique that works similar to LASSO and Ridge regression techniques, and the result is that **the empirically estimated probabilities will push towards a uniform distribution and the larger the value of Lambda, the stronger would be the extent of smoothing. On the other hand, as the sample size increase, the impacted of smooting is reduced.** Smoothing will solve the problem of zero probabilities, but adding one to each count is arbitrary.

Artificial Intelligence Risk Certificate

Another issue is that there will likely be words appearing in the unlabeled, test, dataset that were not labeled, training, set, and therefore about which the model will have no information. Such words would be removed from the test document because they cannot help the algorithm to select an optimal classification for that document.

9.4.3 Words Meanings

The BoW approach, even when used with n-grams, simply counts the occurrence of words or group of words, not imputing any meaning to them. This simplifies the analysis but also represents a severe limitation.

A technique known as word embedding attempts to capture a words meaning by presuming that if two words have similar meaning, they will appear in similar contexts across documents. To represent this numerically, we need to identify not only which words from the vocabulary appear in each document, but also to pay attention to their positions. A solution is to set up a separate vector for each word in the vocabulary and then use one-hot encoding, 0-1 binary variable, for whether the word occurs in a particular position or not. The resulting matrix will be vast in dimension and extremely sparse.

Artificial Intelligence Risk Certificate

Word2vec is an algorithm developed by Google for doing such word embeddings, and it reduces the dimensionality using neural networks.

9.5 NLP Evaluation

The most appropriate way to evaluate an NLP model will depend on the nature of the problem that was specified in the first place. As for the machine learning applications, the dataset can be separated into training, validation and testing sub samples, with evaluation conducted on the latter. If the documents are labeled, this provides a right answer, and in such cases the machine classification can be compared with the labels using a confusion matrix and standard measures such as accuracy, precision, and recall. Simple NLP methods, like BoW are not useful for determining the context in which a word is used to understand the intended meaning. Advanced techniques like n-grams and skip-grams need to be employed for determining the context in which a word is used.

However, in many relevant scenarios the documents will not have a label. For instance, if we were interested in determining how customers were reacting to a new financial product based on the social media posts they were making, there would be no labels, and it would probably be infeasible to go through each message manually to classify it as favorable or unfavorable. In such cases, the models can be evaluated in the same way

Artificial Intelligence Risk Certificate

as unsupervised models. When evaluating NLP applications, it is also essential to consider the speed of the algorithm and how much data is required to train it. As we discussed, textual databases can be vast and sparse word vectors are hard to handle efficiently, which together may make training more sophisticated models such as neural networks very slow.

Appendix 9.A Naïve Bayes application Problem

You work as a data scientist at a sell-side analyst firm. The firm gives way for free the narratives about the companies it provides buy/hold/sell recommendations for, making money by selling the recommendations themselves. The CFO is concerned that it would be possible for a third party to predict the recommendations from a NLP analysis of the narratives. You investigate this by examining a random sample of ten known recommendations and a set of five keywords from their accompanying narratives, observing the information in the following table, where one denotes that the word is present in the narrative and zero that is not. To simplify the analysis, you ignore the hold recommendations and focus on buy and sell.

Artificial Intelligence Risk Certificate

Label	Buy	Buy	Buy	Buy	Buy	Sell	Sell	Sell	Sell	Sell
Word 1	1	0	1	1	1	0	0	0	1	0
Word 2	1	1	1	1	1	0	1	0	0	1
Word 3	0	1	1	0	0	1	0	1	0	1
Word 4	1	0	0	0	0	1	1	0	1	0
Word 5	0	0	1	1	0	1	0	1	1	1

A Suppose that you decide to make a prediction for another set of recommendations based on an analysis of the above table, which can then be compared with the actual recommendations. The first of these new instances includes each of the first three words in its narratives but not words 4 and 5. By making use of the naïve bayes classifier, identify whether the most likely recommendation is a buy or sell.

B If an examination of additional documents from the sample reveals other words not incorporated in your analysis, how would you treat them?

C Explain why smoothing is often required in the context of naïve bayes applications and explain how the technique works.

Artificial Intelligence Risk Certificate

Answer - A

We first calculate the unconditional probabilities, which is simply a count of the number of each outcome over the total amount of outcomes. 5 buys and 5 sells so 50% each, i.e.

$$P(c1) = 0.5; P(c2) = 0.5$$

Then the probability of each word in each outcome type

Word, i	$P(c_1 w_i)$	$P(c_2 w_i)$
1	4/5	1/5
2	5/5	2/5
3	2/5	3/5
4	1/5	3/5
5	2/5	4/5

The question states that the new instance contains words 1-3 and not 4-5, we calculate the conditional probability of each type of recommendation.

$$P(w1 = 1, w2 = 1, w3 = 1, w4 = 0, w5 = 0 | c1)P(c1) \\ = \frac{4}{5} * \frac{5}{5} * \frac{2}{5} * \frac{4}{5} * \frac{3}{5} * 0.5 = 0.0768$$

$$P(w1 = 1, w2 = 1, w3 = 1, w4 = 0, w5 = 0 | c2)P(c2) \\ = \frac{1}{5} * \frac{2}{5} * \frac{3}{5} * \frac{2}{5} * \frac{1}{5} * 0.5 = 0.00192$$

We could then normalize them by dividing each per the sum and obtain $P(c1) = 0.98$ and $P(c2) = 0.02$. Clearly a buy.

Artificial Intelligence Risk Certificate

Answer B

If there are any words in the unlabeled documents that were not in the labeled sample, they cannot be used on the analysis, because we cannot obtain any information from them that will help us to select the most appropriate classifications for unlabeled documents. Unless we are able to extend the labeled sample to include documents containing these words, we would simply have to ignore them.

Answer C

An issue with naïve Bayes is the zero-probability problem. This arises when a specific word does not appear in a document for a particular classification in the labeled sample, which would render the probability of that classification as zero for the new unlabeled data point. Consider that word 4 never appeared in a buy recommendation. Then if a new instance included word 4, then the conditional probability that this is a buy recommendation should be 0. Even if it is unlikely that this narrative corresponds to a buy recommendation, it would be rather draconian to set the probability to zero. Laplace is one of a family of smoothing algorithms that tackles this issue by adding a positive integer to all the counts in the conditional probability matrix, which has the effect of increasing otherwise zero conditional probabilities.

Questions and Answers Module 2 Chapter 9 from GARP – Natural language processing

9.1 Suppose that an analyst wants to determine the sentiment embodied in the prospectus of fifty firms that are undertaking an initial public offering in terms of how bullish they are about the company's prospects. Briefly explain the basic steps involved in doing that via a heuristic approach.

Assuming that the documents are electronic, the following steps should be followed.

Firstly, **establish a dictionary**, as this would comprise three lists of words: positive, negative and neutral.

Secondly, **preprocess the text**, by removing punctuation, symbols, and stop words, as well to modify all words to lower case.

Thirdly, **replace words with their stems and lemmas**

Fourthly, **calculate the proportion of positive words**, the proportion of negative words and the proportion of neutral words.

9.2 Explain how BoW model is used in NLP

Treats each term in a document identically, irrespective of the position of the words and we make a count of how frequently each word is used, with these numbers

Artificial Intelligence Risk Certificate

placed into a vector for each document. The vector can be then analyzed, either by comparing it with a pre-defined list of words or using machine learning models. The words lists are known as dictionaries, for example, to capture whether a particular document embodies a positive or negative statement. This machine learning technique approach would use a labeled sub-set of the documents and treat the words as features with the aim of classifying the remaining unlabeled documents accordingly.

9.3 State and explain two assumptions that underpin the use of naive bayes classifier technique for document grouping

Each word in a document is independent of all other words in the document. This independence assumption is required so that the joint probability can be expressed as a product of the marginal probabilities, $P(A \cap B) = P(A) * P(B)$.

Each word is given equal weighting in constructing the outcome. We could of course remove any stop words prior to analysis, but once that is done, each word is assumed to have equal information value.

Artificial Intelligence Risk Certificate

9.4 Explain why negation words cause problems when classifying documents and suggest an approach to deal with them

Words like Not, isn't, tend to reverse the meaning of a sentence. The solution is to use n-grams, with $n > 1$, where words are considered as blocks rather than individually.

9.5

A- Why some words have zero TF-IDF?

A word will have zero score if it appears in every document, because has zero power to discriminate and will be the $\ln(1)=0$

B- Why does the same word have a higher TF-IDF score in one document than another?

The inverse document frequency does not vary by document, but when it is multiplied by the term frequency within the document, this can lead the same word to have different TF-IDF values across documents, which lead to different time frequency values. In the very short document examples here, each word was only used at most once in each document, but in any practical scenario, some words are likely to be used several times in particular documents, which would lead the TF.IDF for a given word to change further between one document and another.

10. Generative Artificial Intelligence

ChatGPT uses a transformer, a specific large language model. This chapter provides an understanding of the relationship between GenAI, LLMs and the technologies and algorithms that underlie them.

Learning objectives for the chapter:

Describe and distinguish between different generative artificial intelligence technologies.

Describe the role of LLMs in GenAI

Explain how embeddings are used to represent word vectors.

Differentiate between the two Word2Vec architectures.

Differentiate between recurrent neural Networks and transformers for capturing the relationship between words in a sentence.

Describe the basic structure of LLMs at a conceptual level.

Discuss prompt engineering and temperature in the context of LLMs.

Describe applications of GenAI and LLMs.

10.1 Introduction

Technological innovation in generative artificial intelligence, GenAI, has spurred a great deal of interest over the past several years, accentuated by OpenAI's release of ChatGPT. This a type of chatbot based on GenAI technology known as a transformer, which is a specific type of large language model. Due to rapid adoption of this technology, there has been some confusion regarding the distinctions between GenAI and LLMs. Therefore, this chapter begins with an outline of the relationship between the types of GenAI technology and the algorithms that underlie them. We then revisit the word embedding and extend the framework to a more flexible representation in the Word2Vec model. The two forms of Word2vec architecture, namely the continuous bag of words, CBoW, and the skip-gram, are represented as well as the Recurrent neural network model.

10.2 A simple taxonomy for GenAI

The below table describes GenAI by modality of content generated, text, video, etc and by the type of model use, such as transformer, recurrent neural networks.

Regarding **Text**, these are models that will take a prompt and generate textual output. The underlying technology is usually a type of transformer.

Artificial Intelligence Risk Certificate

Regarding **Image**, these are models that will take a prompt and create an image as output. A popular image generator is DALL-E.



Figure 10.1: DALL-E generated image showing the financial markets, and its participants, as a giant roller coaster

Regarding **Audio**, these are models that will generate audio including speech, sound effects and even music.

Regarding **Video**, these are models that will create videos based on a prompt or can be used to modify existing videos.

Multimodal, at last, are models that take textual prompt as their input, and then generate text, images, audio or video. However, significant effort and resources are being deployed to develop multimodal GenAI solutions. These technologies use multiple modalities and formats of data as input and then can generate more robust and realistic outputs in a multimodal format.

Artificial Intelligence Risk Certificate

Table 10.1 Simple Taxonomy for Generative AI

Underlying Model	Content Output Modality				
	Text	Image	Audio	Video	Multi-modal
Large Language Models (LLMs)					
Transformers					
Stable Diffusion					
Variational Auto-Encoder (VAE)					
Generative Adversarial Network (GAN)					
Recurrent Neural Network (RNN)					
Examples	GPT, BERT	DALL-E, StyleGAN, VQ-VAE	Jukebox, WaveNet, Tacotron TTS	Deepfakes, Synthetic Media Tools	Gemini Imagebind GPT4-V Med-Palm2 Kosmos-1 Blip-2

Large Language Models are models developed for natural language processing and are built using vast neural network models with billions of parameters. Instead of utilizing Human-created dictionaries, LLMs develop rules entirely from the enormous corpus of text that they are trained on and are then able to apply this universal knowledge across a wide range of tasks. LLMs are built mainly on transformer architecture and have given rise to the growth of easy-to-use text creation GenAI platforms.

Transformers are a type of deep learning model used for NLP. Popular examples of transformers are BERT, bidirectional encoder representations from transformers and GPT, generative pretrained transformers.

Artificial Intelligence Risk Certificate

Stable diffusion is another powerful image generating technology which is trained on a database of images and uses probabilistic functions to reconstruct an image based on a textual prompt.

Variational Autoencoders, VAE, are a type of neural network with multiple hidden layers and therefore more specifically, would be classified as a type of Deep Learning algorithm or deep neural network. Is trained on a set of images that are compressed on the encoding layer, and the reconstructed in such way here the loss of information in the first layer is replaced using a probabilistic function to create a new image, the decoding layer. A VAE is particularly useful for modifying existing images due to its ability to fill in missing information by probabilistic inference.

Generative Adversarial networks, GANs, are used to create image, audio and video. Employ two neural networks that both compete against and work with one another. The network is generator which creates simulated data. This simulated data is combined with real data- which may be from a curated database, a training set of sorts, or just pulled from the internet, and then fed into another neural network just called the discriminator. The discriminator's job is to figure out which data items are real, and which are fake. It then assigns a probability score that is sent back to the generator, which then uses this updated information to create better generated data. This process continues iteratively until the generated data is virtually indistinguishable from the real data.

Artificial Intelligence Risk Certificate

Recurrent neural networks, at last, are a feed-forward neural network, in which the input features flow from the input to the output layers, generating outputs. However, in a basic RNN, the inputs from the current time step as well as its output from the prior time step are fed back to it. In finance these models are being used due to their ability to represent time series data.

10.3 Word Embedding, Word2Vec and RNNs

The binary BoW sets up a separate vector, V_i , of length $|W|$ of the entire vocabulary for each document i , where $i=1$ to N . The length of $|W|$ could of the order of 100,000 words for a real application. Each element in the vector will take the value 1 if the word appears in the document and 0 otherwise. A slight variant of this is the count BoW, or frequency BoW in which the vector includes the count of how many times the word appears in the document. These two have three major limitations.

The first it that the techniques have nothing to say about the interpretation of a particular word, and therefore an analyst using the bag of words will not be able to identify when two words have a similar meaning, such as amazing and fantastic.

Secondly, this process creates vast vectors that are extremely sparse, containing mostly zeros. This is extremely inefficient, and the information hard to store, let alone analyze.

Artificial Intelligence Risk Certificate

Thirdly, each word is treated independently of all other words in a document so that the context in which they occur is lost. This causes issues when a specific word has more than one meaning, like a call option or a phone call.

Word2Vec addresses the first two problems by using word embedding and RNN are useful for identifying the dependencies between words.

10.3.1 Word2Vec

Created by google in 2013, a team lead by Tomas Mikolov as an alternative to the conventional BoW method for vector representation of words. With Word2Vec, each word in the vocabulary has its own vector that is constructed by examining all the surrounding words in each document where that word occurs. The model can be used to determine the other words that are most like a particular word by determining the degree of co-occurrence of a group of words, i.e., how likely they are to appear together on a sentence. Compared to BoW, W2V, permits a compressed representation by using an embedding, which employs a neural network to effectively reduce the dimensionality by creating a dense vector representation of word storage for analysis.

There are 2 main architectures through which Word2Vec can be implemented, the Continuous bag of words and the continuous skip-gram, both of which roll iteratively through the words in all documents in the corpus.

Artificial Intelligence Risk Certificate

The **CBoW** is a “fill in the blank” technique that proposes a probability-ordered list of words that would fill in the gap between a few words before and after a missing word. If we use a window of length c on each side of the targeted word, W_i , we will use the context word as follows.

$$W_{i-c}, W_{i-c+1}, W_{i-1}, W_{i+1}, W_{i+c-1}, W_{i+c}$$

We thus have $2c$ context words and the goal is to select the targeted word with the maximum probability, i.e. the most likely target words given by the surrounding words, which we could write as :

$$P(W_i | W_{i-c}, \dots, W_{i+c})$$

The **skip-gram**, on the other hand, is like a word association technique that uses a particular word to predict the words that surround it within a certain number of places before and after that word. Here, we would try to predict the context words $W_{i-c}, W_{i-c+1}, W_{i-1}, W_{i+1}, W_{i+c-1}, W_{i+c}$ based on knowledge of only the word W_i . Now, the goal is to maximize the joint probability of all context words given one input word.

$$P(W_{i-c}, \dots, W_{i+c} | W_i)$$

Word2Vec can capture semantic similarity and uses a neural network that is somewhat like an autoencoder, with a single hidden layer, known as embedding layer. This is illustrated assuming a context of three words before W_i and after. For CBoW there is one output that

Artificial Intelligence Risk Certificate

is the target word and six inputs, and for skip gram one input and six outputs.

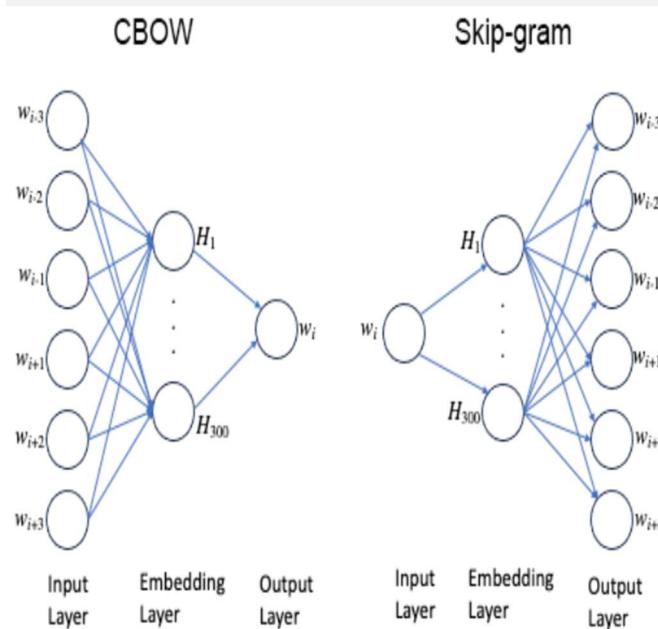


Figure 10.2: The two architectures available for the Word2Vec vector representations

Typically, the embedding layer includes 300 neurons. This implies that we can measure how close every word is to every other word using a weight space, of dimension $|W|*300$, from the input layer to the hidden layer, rather than $|W|*|W|$ or the conventional bag of words, which implies a considerable reduction in dimensionality. The input layer comprises a set of one hot encoded vectors that take the value 1 in the position reflecting the index of the word, and 0 everywhere else,

Artificial Intelligence Risk Certificate

while the output layer contains the probabilities associated with the predictions for the targeted word, on CBoW or for the context words on Skip-gram. Estimation of the weights usually takes place via gradient descent method like the one employed for training feedforward neural networks.

To find word similarities, a measure such as the cosine similarity of two words in a word vector space is typically used. Cosine similarities measure the distance between words where similar words will be close together and dissimilar words would be far apart in the vector space. The model learns the relationship between words by looking for co-occurrences across sentences. For example, for the word risk is likely that words such as loss and management would be close to it, but university would probably be much further away. In addition to analyzing how close the words are to each other. The embeddings can also reveal other relationships between words. One can perform quasi-mathematical operations with the word vectors, which capture word associations in different dimension of word vector space. For example, the embeddings can be used to identify analogous words. It is possible to perform the following operation.

$$\begin{aligned} \text{vector}(x) = & \text{vector(USA)} - \text{vector(NY)} \\ & + \text{vector(London)} \end{aligned}$$

And find that the word UK that is closets to the resulting vector using the cosine similarity measuer. This example demonstrates that embeddings have geographical, and

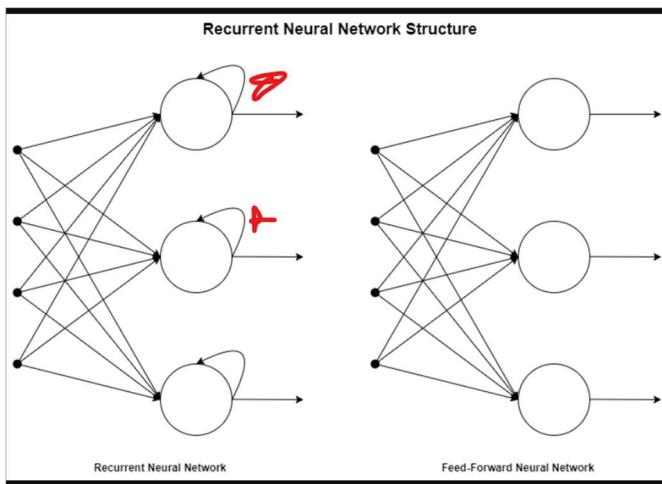
Artificial Intelligence Risk Certificate

other relationships embedded in them. Words embeddings are useful because they can capture pseudo meanings through similarities measured by distances between vector representations. However, although word embedding techniques such as Word2Vec represented a major step forward in NLP, they are nonetheless limited by their inability to capture sentence-level contexts, like call option and phone call. Word embeddings ignore more distant words that fall outside of the c range, and do not consider word order within the range. Word2Vec usefulness is also constrained by its requirement that the context window length is fixed and must be specified a priori by the analyst.

10.3.2 RNNs

Alternative architecture that has been used to capture the relationship between distant words is a RNN. In a basic version, RNN, a memory of hidden cell is present. The hidden cell's value at time t is a function of the current inputs, as well as its value at time step t-1. At any time step, the RNN combines the inputs from the current time step as well as the values stored in memory to generate the output.

Artificial Intelligence Risk Certificate



RNNs are trained like a feed-forward neural networks, by unrolling the network in the time domain. In other words, the same RNN Cell can be thought of as a separate unit in a feed forward network with the inputs moving through the cells in each time step. Similar to the training of FNNs, backpropagation is used to train the RNNs with the cost function being evaluated only at the final time step.

RNNs do not treat a word in a sentence as if they are independent of one another. Rather, they are designed to handle ordered sequential data and they operate iteratively, i.e. in a loop, by combining the current inputs with the values stored in the hidden units to generate the output for a time period. The output values are stored in the hidden units, which then become part of the output calculations in the next time period. For this reason, RNNs are sometimes also known as

Artificial Intelligence Risk Certificate

autoregressive neural networks. They can capture the order of the words in a sentence and potentially, they can also handle dependencies between one word and another word any number of places away in a sentence. RNNs use data sequentially, which means that they process each observation one by one.

Word2Vec requires that each vector to be the same length, RNNs can handle variable length vectors.

However, generic RNN models face the issue of vanishing gradient, limiting their ability to learn long-range dependencies. This limitation means that the further away a word is in a sentence, the lower the impact of the current value. The gradients can also explode the activation functions with derivatives that can increase in value from step to step, also known as the exploding gradient problem.

With RNNs, words a long way from the current one can have a very small connection with the current word, and hence only a weak link with it in the model. Yet it might be the case that a word a long way from the current one has the most important association with it. RNNs are not ideally to capture long-range dependencies between words. A particular class of RNNs, known as LSTM model, was developed to mitigate this issue by using gates to regulate information flows.

10.4 Transformers and LLMs

Transformers were the next major development in models for natural language processing. They are a class of feedforward neural network models based on the concept of attention. The original ones consisted on the concept of encoder and decoder. The encoder is the unit that processes the text, converts it to vectors and understands the text. The decoder is the unit that generates text based on the input of the prompt it receives. This type of structure with both an encoder and a decoder are called a sequence-to-sequence structure. The transformer architecture has evolved from its beginnings with some later transformers specializing only in encoding and other focusing on decoding. Transformers attempt to create a semantic understanding of a sentence by learning the relationships between all the words that it contains, which allows them to capture long-range dependencies in sentence structures. Each sentence is encoded into a set of vectors with one word each. Consider the sentence “The analyst stopped using MGARCH because is too hard to estimate”. To analyze correctly, it is essential to link the word it with MGARCH rather than the “analyst”.

The transformer does this by calculating, for each word, a similarity score between a query vector, which captures the information that the model seeks, and a key vector, which captures the usefulness of each word for the query. The similarity scores are then standardized

Artificial Intelligence Risk Certificate

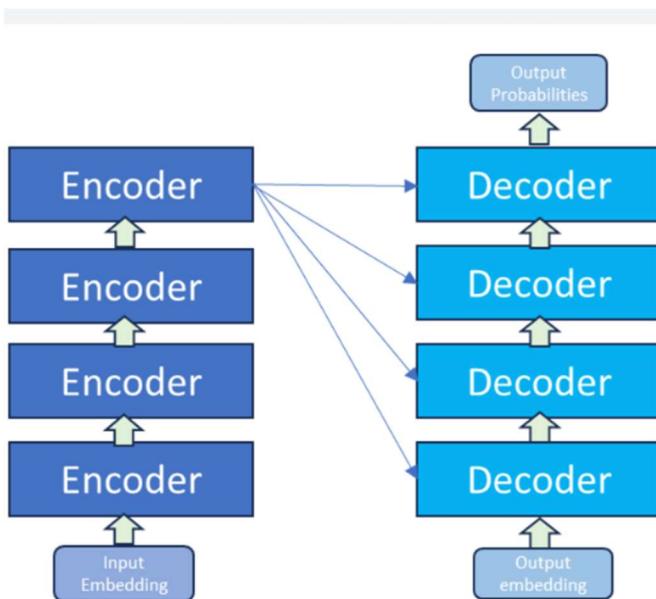
and used as weights for each word. Here, the similarity score for “it” would be highest for “MGARCH”.

The attention mechanism is a key feature of transformers and other deep learning models. It is applied to each set of words repeatedly to encapsulate different relationships, resulting in it emphasizing some parts of the text being processed that are determined to be important. In this way, the model calculates the joint probability distributions across sequences of words, encapsulating the word combinations that are most likely to occur. Transformers have proven much better at most NLP tasks than RNNs. They are deep learning neural networks, that have many layers. They can use parallel processing capabilities and can be trained much faster than RNNs. Consequently, they have rendered Word2Vec as well as RNNs outdated. The structure of transformers enables them to construct universal world representations, i.e. makes them better at applying their training to a variety of tasks. Additionally, transformers make use of contextualized embeddings, which means that words have different vector embeddings when they are used in different contexts, allowing the models to differentiate successfully between instances of a specific word that may have more than one meaning.

Transformers operate on whole sequences of text inputs simultaneously. This makes them better at capturing long-range dependencies between words and allows them to have more efficient parameter estimations when compared to RNNs. On the downside, transformers are computationally more intensive than RNNs. Also, when

Artificial Intelligence Risk Certificate

compared to other types of neural networks, transformers are more difficult to interpret due to their complex, multi-layered structured. Transformer's relative opaqueness can make the identification of the sources of any issues, errors or problems quite challenging.



10.4.1 Large Language Models

LLMs are predominantly based on transformers that are pre trained on a range of vast textual datasets. At the heart of LLMs is a large neural network with billions of parameters that is trained on a very large corpus of text. The weights of the parameters are determined using gradient descent or other methods similar to the process used for training a basic neural network. The LLMs can be either autoencoding or autoregressive, or a combination of both.

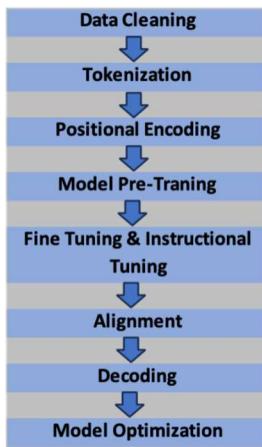
An **autoencoding LLM** is trained to encode sentences by masking some of the input and predict the masked words using the other parts of the input. An **autoregressive LLM**, on the other hand, predicts the next word or part of a word. This makes them useful for generating text. A combination of both is more versatile than the individual approaches, as they have both the encoder and the decoder components. Although BERT is an example of an autoencoding LLM, the GPT family of LLMs are examples of autoregressive LLMs.

The training of an LLM starts with cleaning the training data, corpus, which includes filtering, removal of duplicates, noisy data and punctuation marks as well as resolving or removing ambiguous data. The data also should be balanced, which entails adjusting the class distributions to have a fair representation and avoid any biases. The data is separated into small parts or tokens to enable efficient processing. This is followed by encoding to store the positional information of

Artificial Intelligence Risk Certificate

sequences of text. The model is then pre-trained on the textual data usually without any labels. Through pre-training, the LLM understands the relationship between the words in the corpus it is trained on.

Once pre-training is completed, the models may require fine-tuning to transfer the knowledge gained during pre-training to conduct specific tasks. To accomplish this, the model may be fed labeled data relevant to the specific task being handled. The model is then modified in a process called alignment to conform with human preferences, beliefs and principles. During this process, feedback from user may be used to improve the model. The decoding module employs different strategies to generate the output based on the input text. Finally, the model is optimized for production use by turning off some lower-level neural network layers, dropping unimportant weights and reducing the precision of weights.



Artificial Intelligence Risk Certificate

BERT is an autoencoding LLM, with 100M parameters and was trained on a corpus of books and Wikipedia pages. New versions allow for hidden parts of the sentence to be hidden in each training cycle, which ensures that the model will be more robust to diversity on the training data. FinBERT was trained on Reuters data including corporate statement and analyst reports. BERT only include an encoder side, which converts the words into embeddings and a positional encoder that captures the location of the words within sentences. There is no decoder as there would be in an autoencoder model, for instance. These encoders are bi-directional, which means that for determining the meaning of a sentence, they can examine words appearing both before and after a given word. A set of hidden layers in a transformer model employ self-attention to capture the relationship between words. Several self-attention “heads” are used in parallel, with each capturing the linkages between different sets of words in the sentence. Because BERT includes only an encoder, it is not generative, but it can be used for applications such as classification.

BERT involves training the model on a large body of unlabeled text. The second stage of training involves two aspects.

Masking words randomly within sentences and getting the model to predict the hidden words, usually 15% of words are masked. This is known as self-supervision as the hidden words are known and therefore act as a supervised technique.

Artificial Intelligence Risk Certificate

Next sentence prediction, where sentences are extracted in pairs from documents and the model is tasked to predict an entire sentence from the preceding one rather than just a single word.

BART is another class of **autoregressive transformer model** incorporating both encoder and decoder layers. Incorporating a decoder makes it able to generate output text rather than merely being able to analyze inputs and encode them in vectors. This means that BART can be used for a wider range of tasks such as document summarization, language translation, and creative writing, although it requires greater computational resources than encoder-only models. BART is trained by deliberately distorting blocks of text and getting the algorithm to learn to generate the original forms. The autoregressive decoder allows it to generate sentences one word at a time while remembering the previous words in those sentences. BART has been found to have provided superior results in many standardize test applications, and responding to questions, in comparison to alternative models.

10.4.2 Cloud Based LLMs

Because the corpus is huge for LLMs, and the estimated weights matrix is in the millions, billions or even trillions, training is only performed once. It would be infeasible to train separately on this volume of data for each task. Most LLMs are so large that it is also infeasible to store them locally. Hence, they are known as cloud based LLMs. The best known one is ChatGPT which is one of a family of generative pre-trained Transformer, GPT. First stage was like BERT in which the model had around 100M parameters trained on unlabeled corpus of unpublished books. Then at a second stage, the model was fine-tuned for specific applications using a smaller, labeled datasets. These applications could be, for example, sentiment analysis, responding to questions, summarizing a document or classification.

Two important terms are Low-shot task transfers and no-shot fine-tuning. The first term refers to tuning a model that has been already trained with a large data set using only a small set of new examples. The second term is a fine-tuning technique wherein we simply rely on the prior training of the model without any fine-tuning with new examples. Each generation of GPTs has become better at low-shot task transfers, implying that the number of labeled examples required to fine-tune the model for a specific application has declined. GPT-2 removed the need for the second stage fine-tuning, so that “no-shot fine-tuning” became possible, meaning that the model can adapt to new situations without any supervised examples at all. GPT-3 comprised 175 billion

Artificial Intelligence Risk Certificate

parameters, which further enhanced the model's ability to adapt to different kinds of tasks, and gave it generative capacity, whereby it could create new instances of text from human instructions and examples contained within the training set. For instance, it could create synthetic news articles that were comparable to the ones written by humans. This is a form of generalization, which is different from simply memorizing sentences and repeating them verbatim when tasked to create material on a specific topic.

ChatGPT 3.5 is an improved version of the previous ones, which allow the model to show labeled examples of the required output, instructing the model to create additional instances, then allowing the user to provide ratings for the generated outputs, also known as reinforcement learning from human feedback (RLHF). Through this process the LLM was fine-tuned to be able to follow instructions more closely, providing informative responses to queries while refusing to perform tasks deemed inappropriate.

GPT4 is the most recent one, and as well as further increasing the training corpus and number of parameters, the specifications include a significant new capability to be able to accept images as an input rather than purely text.

BARD is a fine-tune version of LaMDA, language model for dialog applications model. BART was trained using infiniset, a vast corpus including Wikipedia, books and scraped webpages.

Artificial Intelligence Risk Certificate

Gemini a multimodal LLM that is an updated version of BARD, was trained on text, video and audio. According to Google, it is the first LLM to outperform human experts in massive multi-task language understanding.

10.4.3 Chatbots

Are conversational interfaces that interact with a human user. It can be either web based or through automated speech recognition and text-to speech synthesis. Chatbots represent an important application of NLP, and they have been increasingly used by organizations to reduce costs and improve response times to queries. For instance, banks often use chatbots as automated costumer service representatives that can deal with queries and either provide the required information directly or put the costumer in contact with the most appropriate department in the bank. The most basic form of chatbots simply provide responses to standard lists of frequently asked questions (FAQs). They have no real conversational ability and are of limited use in costumer service, extracting and quoting information that is probably already on the website. More complex chatbots are flow-based, meaning that they can have an interactive, multi-step conversation with a costumer by asking more detailed follow-up questions in response to a costumer question or comment. For instance, if a costumer states that they are unhappy with the service that they had received, the chatbot might ask whether it

Artificial Intelligence Risk Certificate

relates to a mortgage or savings product, or whether they transacted in a branch or on-line.

The quality of chatbots is being improved significantly using LLMs, which can give a deeper and more accurate understanding of customer's request than earlier NLP models, therefore providing more informed answers. They can also draw on a wider range of material than simply what is on the web page. We are fast approaching the point where chatbot responses will be almost indistinguishable from those given by a human customer service agent.

10.4.4 Using LLMs – Prompt engineering and Temperature

Using LLMs poses challenges. This is where prompt engineering, the art and science of designing prompt for effective use of LLMs comes in. **Prompt engineering is an interactive process for getting the best results for a given task from an LLM.**

Prompts should be direct and concise, provide the context to the model by providing reference material and output from other tools, split tasks into smaller and more manageable tasks for better responses, provide sufficient information to obtain the best results, be flexible for use in different models and should allow the model time to find good results.

Artificial Intelligence Risk Certificate

During the prompt design process, users may supply the model with a few relevant examples, so that the model can generate more context-specific responses. This technique is known as few-shot learning. It is also a good practice to ask for output in a structured format so that it can be easily used elsewhere. Prompts with different personas, or styles may be used to test how the model responds to a query in different ways. The model developers can use the output from these tests to fine-tune the model to avoid undesirable responses. Users can also control how the LLM Behaves by choosing an hyperparameter called temperature.

Temperature hyperparameter is used to change the shape of the probability of the distribution of tokens being predicted by the model. If temperate is set below one, the probability distribution is narrowed such that only the most likely tokens are selected. If it is above one, the probability distribution is widened and flattened, resulting in outlying tokens also being selected. If one, then the probability distribution is not changed. **The lower the temperature parameter is, the more predictable the output is.**

There are other parameters to adjust and control the behavior of LLM. Maximum number of tokens generated the penalty level for the tokens that repeat themselves and so one.

10.5 Applications of Generative AI and LLMs

Generative AI and LLMs are one of the most recent and exciting developments in AI. Users in academia and industry are currently experimenting with these tools.

Textual processing, the area that has seen the most usage so far. Analyst can use GenAI to assist with writing reports as well as for generating summaries of documents. This is particularly useful in legal analysis and litigation support.

Gen-AI based tools can be used to monitor news feed, customer opinions and social media posts on specific companies to identify any potential risk.

LLMs can be used for analyzing incoming market information, economic data, and new flows as well as for predicting short term movement prices. SEC filings and other reports by companies can be analyzed quickly using GenAI based tools for detecting any emerging risk. Compliance Monitoring is another area where LLMs can be effective. LLMs can be deployed for translating text written in foreign languages. Can play an important role in credit underwriting as they are useful for processing both structured and unstructured data and generating reports for credit analysts. LLMs will be useful in conducting an in-depth examination of all available materials on borrowers, as well in reducing the time taken to arrive at credit decisions.

Popular GenAI tools like Gemini, ChatGPT and Llama have been used successfully in textual processing.

Artificial Intelligence Risk Certificate

Bloomberg GPT and FinGPT are examples of LLMs developed for analyzing financial news flows and for generating trading signals. Several investment firms are using these to analyze news.

Code generation, in which developers can use LLMs to generate and debug code for their projects, which could reduce software development costs. LLMs can also be useful for generating documentation.

Chatbots and virtual assistants, can be developed for applications such as online customer support, providing information to risk professionals on existing and upcoming regulations, helping medical professionals analyze symptoms, assisting investors and investment advisors in identifying potential investments for their portfolio. Most LLM providers have tools for developing custom chatbots and virtual assistants.

Fraud and anomaly detection, in which GenAI's capacity to automate the monitoring and processing of a company's own internal data could make an even more significant contribution. Equipped with the latest statistical tools and algorithms, risk managers could leverage AI to compile and scrutinize transaction data across various departments for anomalies or outliers linked to specific suppliers or operations.

Cyber security, in which LLMs can be used to constantly monitor emails and other data streams for potential cyber threats, cross-referencing them with an institution's system profile to pinpoint specific vulnerabilities such as malware. After identifying a

Artificial Intelligence Risk Certificate

threat, the system could automatically alert risk managers and other relevant individuals in real time. The true power of AI system would be to then proactively source patches for these threats directly from approved software vendors for system engineers to implement.

Claim processing in which GenAi can be used for developing applications that can automate the interviews with costumers that are filling claims.

Image generation and image processing, which can be very helpful and useful in providing rapid prototype development of image as a starting point for artistic projections such as marketing materials, book covers and illustrations to be sued in slide decks.

Text, speech and video conversion, Multimodal LLMs, will be useful for converting text to speech and for generating video from the text. They are also useful for transcribing text from audio.

Autonomous driving vehicles, which is a technology under development since 1970. While there are currently no systems that offer completely autonomous driving, semi-autonomous driving systems have been deployed by various companies, such as Tesla. Artificial intelligence is a crucial element of autonomous driving technology. Although there is considerable interest in this technology, there are some concerns about it.

Although text generation capabilities have been used to produce good first drafts, others like images and videos are still at an experimental stage, with their usage

Artificial Intelligence Risk Certificate

currently limited to generating logos, mockups and basic videos. As these technologies are adopted and used widely, the quality of the text generation function is likely to converge to near final draft quality by the end of this decade. Similarly, code generation can also progress from first drafts that need to be modified by experienced coders before use to production level code quality in a few years.

	Pre-2020	2020	2022	2023 ?	2025 ?	2030 ?
Text	Spam detection Translation Basic Q&A	Basic copywriting First drafts	Longer form Second drafts	Vertical fine-tuning gets good (scientific papers, etc.)	Final drafts better than the human average	Final drafts better than professional writers
Code	1-line auto-complete	Multi-line generation	Longer form Better accuracy	More languages More verticals	Text to product (draft)	Text to product (final), better than full-time developers
Images			Art Logos Photography	Mock-ups (product design, architecture, etc.)	Final drafts (product design, architecture, etc.)	Final drafts better than professional artists, designers, photographers
Video/ 3D/ Gaming			First attempts at 3D/video models	Basic/first draft videos and 3D files	Second drafts	AI Roblox Video games and movies are personalized dreams

Large model availability: ● First attempts ● Almost there ● Ready for prime time

10.6 Chapter Summary

Although it is still too early to make a prediction, it seems fair to say that they will be playing a big role in the future as they are adopted by business and others to improve processes. This is likely to result in productivity gains and fundamental changes in how workers perform their work and the nature of their tasks.

Appendix 10.A Operations with word Embeddings

Word embeddings are vector representation of words used in NLP applications. They provide a dense representation of words, as opposed to the sparse representation provided by other approaches like one-hot encoding. Synonyms and words that belong to a particular category such as geography have embeddings that are close to each other. The purpose of this exhibit is to illustrate how embeddings can be used to perform quasi-mathematical operations. We use GloVe embedding vectors and compute the cosine similarity score, which is the dot product of two vectors, for word pairs. If a pair of words are close to each other, the similarity score would be close to 1. If they are not close to each other the score would be closer to 0.

Table 10.A.1 Similarity Scores for Similar Words

Word 1	Word 2	Similarity Score	Category
China	Japan	0.84	Countries
King	Queen	0.78	Royalty
Gold	Silver	0.95	Metals
Europe	Asia	0.83	Continents

We can also expect that dissimilar word pairs to have low similarity scores. For example, Africa and Tokyo belong to different categories.

Artificial Intelligence Risk Certificate

Table 10.A.2 Similarity Scores for Dissimilar Words

Word 1	Word 2	Similarity Score	Category
Australia	Tokyo	0.40	Country, City
King	Man	0.53	Royalty, Gender
Gold	Dollar	0.50	Metal, Currency
Banana	Rose	0.30	Fruit, Flower

Another example of operations that can be performed using word embeddings is the testing of word analogies. A word analogy is of the form “a is to be” as “c is to d. For example, consider the word pairs Europe-Spain and Asia-China. If we retrieve Spain from Europe and add China, the resulting embedding vector is close to Asia in embedding space. The embedding Images below were plotted after they were reduced to 2 dimensions using the t-SNE technique. It can be seen from this figure that Europe and Asia are close to each other, similarly Spain and China are also close to each other in the embedding space.

Table 10.A.3 Word Analogies

Operation	Closest Word	Similarity Score
America - Washington + London	UK	0.77
King-Man + Woman	Queen	0.86
Gold-Dollar + Pound	Silver	0.80
Europe-Spain + China	Asia	0.79

Artificial Intelligence Risk Certificate

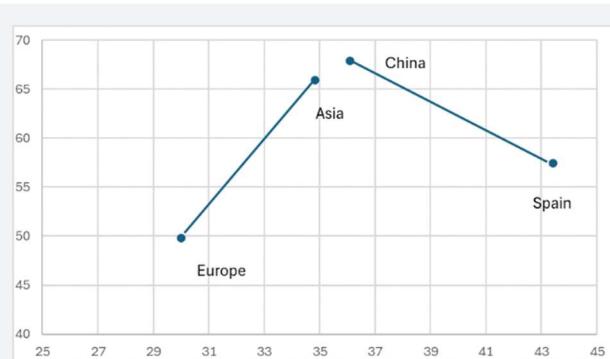


Figure 10.A.1 Word Pairs in Embedding Space After Projecting to 2 Dimensions

Questions and Answers Module 2 Chapter 10 from GARP - Generative Artificial Intelligence

10.1 State whether each of the following is True or false and explain why

A- *Word2Vec can generate blocks of text, such as article summaries and response to questions*

False. A large language model such as GPT could generate long pieces of text, but Word2Vec was developed to represent text as numerical vectors and to be able to analyze that text, for example looking for similar words or group of words.

Artificial Intelligence Risk Certificate

B- Recurrent neural Networks use a static embedding that cannot allow for word ordering in a sentence.

False. RNNs operate in a dynamic fashion, iterating through sentences so that they can capture the ordering of words through their autoregressive structures.

C- Transformers models can be trained to understand different contexts that the same word is used in

True. This is one of the benefits of such models. The attention heads in the model can link together words in different parts of a sentence and so they can capture different ways that the same word will be used in other contexts.

D- Cloud base large language models can be accessed remotely or alternatively, the training corpuses can be downloaded and stored locally.

False. Cloud base only exist on the cloud. The training datasets and weight matrices are huge, which cannot be stored locally. Also, these details are usually proprietary to the model developer.

E- BART includes both encode and decoder layers.

True. BART includes both a bidirectional encoder of the sort incorporated into BERT and an auto-regressive decoder like the one incorporated into the GPT Family.

F- LLMs can only analyze text and not images

False. This was true until recently, but both Gemini and the latest version of GPT can handle images as well.

Artificial Intelligence Risk Certificate

10.2 Explain the difference between the continuous bag of word and skip-gram model architectures.

CBoW is tasked to predict a masked word base on a few context words before and after that, whereas skip-gram tries to predict the context words around a particular word.

10.3 What are the disadvantages of the traditional BoW approach to representing documents as vectors of numbers and how does Word2Vec overcome these?

Traditional BoW treats each word as independent from all other words in a document, the only aspects it captures are whether a particular word appears in a document and how many times. The limitation implies that BoW cannot impute any meaning to the words and nor can it propose synonyms for a particular word. The word embedding is also very inefficient because the vectors created will each be of the same length as the entire vocabulary but very sparse, containing predominantly zeros. Word2Vec uses a variable length embedding that can reduce the dimensionality by creating a dense representation and encodes the position of words within a document so that their meaning can be captured.

10.4 What is the Key difference between RNN and Feedforward neural networks, FNN?

At any time step, the RNN combines the input form the current time step as well as the values stored in memory to generate the output.

Artificial Intelligence Risk Certificate

10.5 Describe the Training process of LLMs

First step is cleaning the Data, corpus, which includes filtering, removal of duplicates, noisy data and punctuation marks as well as resolving or removing ambiguous data. Then, it should be separated into small parts or tokens to enable efficient processing. This should be followed by encoding to store the positional information of sequences of text. The model is then pre-trained on the textual data usually without any labels. Then the model is finetuned to transfer the knowledge gained during pre-training to conduct specific tasks. Then it is modified by the process called alignment to conform with human preferences. Then generate outcome by decoding the model. Finally, the model is optimized for production use by turning off some lower-level neural network layers, dropping unimportant weights and reducing the precision of weights.

10.6

A- What are the guidelines for drafting good prompts?

Prompts should allow the model time to find good results, i.e. asking that the analysis be finished before answering, asking the mode to see if anything is missing in its answer and so on.

B- How to control the behavior of LLMs.

Users can also control how an LLM behaves by setting a parameter called temperature which changes the probability distribution of tokens being predicted.

Module 3 – Risk and Risk Factors

Learning Objectives

Describe and differentiate between the concepts of individual and group fairness. Describe the various measure of group fairness. Discuss trade-offs associated with different concepts and measures of fairness. Describe source of algorithmic bias and unfairness. Describe explainability, interpretability and transparency. Describe techniques for making algorithms more explainable, discuss risk posed by AI to human autonomy, safety and well-being. Describe sources of AI-related reputational risk and strategies for mitigating those risks and finally, discuss global challenges and risk associate with AI.

1.0 Introduction

Technologies pose risk and we are only on the beginning to understand the various risks associated with the use of AI technology. The evolving nature of these risks requires us to adopt a dynamic and informed approach both to ensure beneficial use and to mitigate unintended consequences. In considering the risk implications of AI, it is important to recognize the following.

AI is hugely pervasive in both range and variety of use. This poses challenges to risk prediction, as risks can take on different forms depending on the use case.

Artificial Intelligence Risk Certificate

AI technologies also have increasingly large capabilities, which enables us to use them in increasingly high-stakes settings. Doing so implies that any technological shortcomings can have significant consequences for people. For example, an automatically generated outcome of a loan application can profoundly affect the applicant. Similarly, a clinical diagnosis generated by a medical algorithm can have a long-lasting impact on the patient's life. AI development is rapid, which often poses challenges for impact prediction, risk measurement and risk mitigation. A further consequence of this rapid development is the difficulty of establishing robust governance mechanisms that effectively address risks. This is especially a challenge for regulatory responses, which are often slow to develop and difficult to change once established. A slowness to regulate, on the other hand, may allow for potentially harmful technologies to persist, and can lead to the normalization and entrenchment of harmful practices.

It is important that those developing and using AI technologies recognize and manage the risks. As will become clear, risk from AI are often interconnected. Lack of transparency can affect fairness and safety, algorithmic bias can lead to reputational risk and so on.

2.0 Algorithmic Bias and Fairness

AI systems are being deployed across myriad important domains, executing tasks that can have profound effects on people's lives. These include medical diagnostics, bail decisions, policing, loan approvals, college admissions and recruitment. Although algorithms are often hailed for providing more consistent and objective decision making than humans, they too, can exhibit bias and create unfair results. By now, there are countless accounts of algorithmic bias and discrimination, some with severe consequences for the people subject to algorithm decision-making. Although consistent decision making is desirable for many reasons, the downside of such consistency is that bias, if it occurs, become systematic. As a result, it is extremely important to ensure that algorithms do not exhibit such unwanted biases. This has been recognized by regulators across the globe, who increasingly demand that AI systems be fair and non-discriminatory.

2.1 What is Bias?

The term Bias is used differently across different disciplines and contexts, and it is useful to keep this in mind when referring to algorithmic bias. Data scientists and statisticians use bias to refer to a systematic error in the measurement or prediction process, which in turn leads to a discrepancy between the ground truth and the measurement prediction. Within psychology and

Artificial Intelligence Risk Certificate

cognitive science, bias is a systematic error in judgement, or a systematic deviation from rationality. For example, a user putting more trust into a computational system than would rationally be justified is said to suffer from automation bias. When used in the context of algorithms, bias typically not only include systematic errors of various kinds, but also the resulting unfair outcome of certain groups of people. Here, bias has become a value-laden concept that explicitly refers to the effects of algorithms on humans.

Algorithm bias is a systematic deviation of an algorithm's output, performance, or impact relative to some norm, aim, standard or baseline.

We can make an additional distinction between explicit and implicit bias. In the former case, the algorithm's developer intentionally projects its own biases into the algorithm, leading to biased outputs. In the latter case, bias creeps in undetected through the algorithms design or via the data used to train the algorithm. In those cases, bias is implicit and unintended.

2.2 What is fair?

Algorithmic fairness is a critical yet complex aspect of AI development. Crucially, it requires us to first answer the question, what is fair? The two main approaches to fairness in the context of algorithmic design are individual and group fairness.

Artificial Intelligence Risk Certificate

Individual fairness relates to the Aristotelian doctrine of treating like cases alike, in other words, it demands similarly situated individuals be treated equally.

Group fairness, on the other hand, does not consider individuals, but instead looks at statistical differences between groups.

An example is on college admissions, a college wants to ensure its admissions process is fair and consistent. To aid with the decision-making process the college uses college AD, an algorithm that analyzes test score, school records and writing samples of applicants, and outputs a hierarchy of the most promising students.

2.2.1 Individual Fairness

In college admissions, the prevailing belief is that admission should be based on merit with only the best students receiving offers. We often consider grades as a good indicator of merit, and so the natural consequence would be to offer places only to applicants with top grades. This is an application of the individual fairness doctrine. Because the relevant metric for college admissions is scholarly achievement and nothing else, we avoid direct discrimination of the biases of protected characteristics, such as race, gender or religion. If one applicant has higher grades than another, is selected, no matter what other characteristics it might possess. Here, treating like cases alike means treating applicants that are alike in all relevant aspects alike.

Artificial Intelligence Risk Certificate

However, we quickly encounter limits to this seemingly straightforward principle of individual fairness. Not all students start out equally. Some might have had to overcome significant hurdles to achieve certain grades. If we truly are interested in merit, should we take this fact into account when making decisions? And if so, how should we weigh this fact against other relevant criteria? Is it fair to reject an applicant with top grades but admit a student with slightly lower grades whom we know to have displayed extraordinary resilience, motivation and self-discipline to overcome socioeconomic hurdles or physical difficulties? Moreover, how would we determine what those hurdles were and translate them into numbers that allow us to compare applicants?

These questions have no easy answer, different individuals will hold different views as to whether and how socioeconomic factors should influence college admissions. This is a limitation of the treating like cases alike doctrine. Intuitive as it may be, it requires us first to agree on what counts as “alike”, which is often contentious. This issue becomes especially pronounced in the case of algorithmic design, where the designers need to be explicit about the relevant factors that feature in the decision-making process. These may, furthermore, vary with context.

2.2.2 Group Fairness

As opposed to individual fairness, group fairness does not consider the treatment of individuals, but instead looks at the statistical differences between groups. For example, we may ask whether there are statistical differences between the admission rates of male and female college applicants. This would be a matter of group fairness, as we are interested in whether there are a certain group that are disadvantaged by the algorithm and that share a protected characteristic. There are many different notions of group fairness, and which one is appropriate depends heavily on the context in which the algorithm is deployed. Broadly, these notions fall into two categories. The first category focuses on performance equality and compares algorithmic performance across different demographics. The second category focuses on output distributions without regard to the correctness or accuracy of those inputs. The main concern here is how a given outcome is distributed across different demographics.

Demographic parity

Is one of the most important fairness notions regarding this second category. Requires that the distribution of predictions to be equal across subpopulations. In the case of college admissions, this means that the admission rate is the same across groups. In the case of lending, it means that the proportion of granted loans is

Artificial Intelligence Risk Certificate

the same across groups. Mathematically, demographic parity can be expressed as

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$$

Where,

$\hat{Y} \in \{0,1\}$, a binary classifier with predicted outcomes 0, negative and 1, positive.

$A \in \{0,1\}$, presence or absence of an attribute A, indicating group membership.

$P(\hat{Y} = 1|A = 1)$, the probability of the algorithm making a positive prediction $\hat{Y} = 1$, conditional on the membership to a particular group ($A=1$). Can be immediately seen, demographic parity is entirely independent of the true value, Y. In other words, it does not consider if the predictions are correct, only if the predictions are equally distributed. In some cases, this can be useful, such as when we don't have direct access to the true value of the prediction, the target variable. In practice, this is often the case. In the case of credit scoring, for example, there is a significant time lapse between the time of decision making, and the time of the actual outcome observed. In fraud detection, the algorithm is trained on historical data, but fraud is rare, and fraud patterns evolve over time. Often, direct access to true fraud cases is limited, especially when it involves new types of fraud. Demographic parity ensures that the outcomes are equally distributed, and this is often a desirable property. However, because **it does not consider the quality of predictions**, using demographic

Artificial Intelligence Risk Certificate

parity as a fairness measure can have drawbacks. For example, demographic parity can be satisfied even if all predictions about a particular subpopulation are wrong. In the college admission case, the algorithm might perform worse on a particular subpopulation and thus label many unqualified candidates as qualified candidates. Demographic parity would still be satisfied in the proportion of positive predictions remains the same across subpopulations. Other times, using a demographic parity is simply inappropriate for a given task, such as when base rates significant differ between groups, and those differences are relevant to the decision at hand. In a medical context, for example, **Demographic parity might not be the best or only fairness measure because the primary concern is the quality of prediction.**

Confusion Matrix

As opposed to Demographic Parity, fairness measures that aim to achieve performance equality across subpopulations require us to examine the true value of the target variable. Illustrates how algorithmic predictions can be divided into four categories. In the case of fraud detection, for example, the confusion matrix labels transactions that were correctly identified as fraudulent (TP), transactions that were correctly identified as non-fraudulent (TN), transactions that were erroneously identified as fraudulent (FP) and transactions that were erroneously identified as non-

Artificial Intelligence Risk Certificate

fraudulent (FN). The first thing would be to calculate accuracy, which once again, is the number of correct predictions over the total number of predictions.

It is tempting to consider accuracy a good indicator for the quality of a classification algorithm, but this isn't always the case. Consider the following example, there are 100 job applicants, 10 of whom are suitable, and 90 of whom are not. Imagine now the algorithm correctly classifies the 90 applicants as unsuitable (TN) but also classifies the 10 applicants as unsuitable (FN). In this case, the algorithm would have an accuracy of 90%, even though provided only one outcome. This is notoriously in cases where the data set is heavily unbalanced.

Predictive rate Parity

In the case of some contexts, the cost of false positives is high, and thus accuracy of the predictions for the positive or negative class is important. In other words, we want to maximize the positive predictive value, Precision, which is the $TP/(TP+FP)$. It denotes the proportion of correctly predicted positive outcomes. In the case of fraud detection, it indicates how many transactions that were marked as fraudulent were indeed fraudulent. In the case of credit scoring, it indicates the proportion of applicants predicted to default that actually default. A high PPV in this case means that the algorithm can effectively identify individuals who are likely to default, whereas a low PPV indicates that there are many individuals who would

Artificial Intelligence Risk Certificate

have repaid their loans but were erroneously labeled as high-risk.

We can now define another fairness measure, predictive rate parity, which is satisfied when PPV is equal across subpopulations. In other words, predictive rate parity demands that those predicted as positive by the model, the proportion of individuals correctly identified as positive is the same across groups.

$$P(Y = 1 | \hat{Y} = 1, A = 0) = P(Y = 1 | \hat{Y} = 1, A = 1)$$

Where Y is the target variable, Y hat is the predictor, and A a protected characteristic. This can be extended to multiple groups. Predictive rate parity has an intuitive appeal as a fairness measure. We often want to avoid stark differences in false positives, across subpopulations. For example, predictive rate parity ensures that the proportion of accepted college applicants who succeed at college is the same across groups. This would prevent a situation where the academic potential of applicants is consistently overestimated in one group and underestimated in another.

A downside of predictive rate parity is that it does not account for differences in base rates across groups. If base rates differ significantly, then achieving the predictive rate parity becomes extremely challenging and can have adverse effects on algorithmic performance.

Artificial Intelligence Risk Certificate

Equal Opportunity

In some cases, it is important to ensure that positive cases are recognized as such, as in the case of medical diagnostics, where failing to diagnose a disease can have life changing consequences. A measure that fits this task is sensitivity, also known as true positive rate or recall. It quantifies the proportion of true positives that are correctly identified. i.e., the number of True positives, over the sum of True positives and false negatives, i.e. all that were in actual terms positive. As a performance measure, sensitivity is useful in situation where it is important not to miss out on a positive. Maximizing sensitivity requires us to minimize false negatives. If sensitivity is the same across groups, we have achieved equal opportunity, which is the last important fairness measure we discuss here.

Equal opportunity ensures that individuals of each group have an equal chance of being correctly identified as positive. In medical diagnostics, for example, it ensures that patients who belong to different demographic groups and who have a given disease are equally likely to be diagnosed correctly.

$$P(\hat{Y} = 1 | Y = 1, A = 0) = P(\hat{Y} = 1 | Y = 1, A = 1)$$

The definition can be easily extended to multiple groups. In the case of equal opportunities, the equality constrain is only applied to the true positive rate, so that each group has the same opportunity of being granted a

Artificial Intelligence Risk Certificate

positive outcome. A more-restrictive version of equal opportunity is equalized odds, where we not only require equality of sensitivity across groups, but also equality of specificity. Equal opportunity is a useful fairness measure that ensures, say, that the same proportion of qualified candidates is admitted to college. However, just as with predictive rate parity, equal opportunity requires access to the true value of the target and can be difficult to achieve when there are stark differences in base rates across groups. In those cases, it is difficult to achieve fairness without compromising on other aspects of the model's performance.

Impossibility and Trade-offs

Different notions of fairness highlight different requirements for fair decision making. Each of them has some drawbacks, so could we not simply combine them and demand that all fairness measures hold to ensure the decision is fair? Unfortunately, a mathematical theorem demonstrates that doing so is impossible. Specifically, when base rates between populations are different, which is almost always the case, then it is impossible to satisfy demographic parity, predictive rate parity and equal opportunity simultaneously. The very fact that we cannot achieve fairness on all dimensions demonstrates how important it is to deliberate on the kind of group fairness one wants to achieve when designing an algorithm, because there are always trade-

Artificial Intelligence Risk Certificate

offs to consider. These trade-offs not only apply to fairness measures, but to algorithmic performance optimization more generally. For example, a good fraud-detection algorithm ideally would have high sensitivity and low false-positive rate. That is, we want the system to be effective at correctly identifying fraudulent transactions while minimizing the likelihood of mistakenly flagging a legitimate transaction as fraudulent. To maximize sensitivity, then, we could simply block all transactions, and achieve the maximal sensitivity, since there are no negatives. However, the false positive rate would shoot up in this case. On the other hand, to minimize the false positive rate, we might go to the other extreme and label all transactions as non-fraudulent. Again, because there would be no positives, the FP rates drop to 0. However, because there are now many false negatives, sensitivity would massively decrease. The example demonstrates that balancing different performance measures is by no means trivial task. Instead, it forces us to consider the trade-offs that are involved and pick the balance appropriate to the task at hand.

Artificial Intelligence Risk Certificate

Table 1: Group Fairness Measures and the Questions they Answer

Accuracy	"Of all cases, how many did we correctly identify as either positive or negative?"
Positive predictive value/Precision	"Of all positively predicted cases, how many were actually positive?"
Sensitivity	"Of all actual positive cases, how many did we correctly identify as such?"
Demographic parity	"Is the likelihood of a positive prediction the same across groups?"
Predictive rate parity	"Is the likelihood of a true positive prediction the same across groups?"
Equal opportunity	"Is the likelihood of being positive when predicted to be positive the same across groups?"

2.3 Source of Unfairness

The reasons why algorithms might lead to unfairness, and it is not always straightforward to identify them, we approach the topic in a more-or-less chronological order that leads us through the AI development process. Algorithmic bias arising in problem specification and feature selection, data collection and data composition, model development and model deployment.

2.3.1 Problem Specification and Feature selection

Technology carries instrumental Value, and it helps us solve a particular problem by carrying out a particular task. To build useful technology, however, we first need to specify what this problem is and how it could be solved. During problem specification, purpose and design of the algorithm are explicitly stated. Unsurprisingly, how a problem is specified can have large effects on both algorithmic performance and algorithm fairness.

Consider the example of an Hiring algorithm, in which your company wants to improve and speed up the recruitment process by using a hiring algorithm. You choose to try HireMe, an algorithm that scans applicant CVs and outputs a shortlist of the best applicants. HireMe is trained on data provided by current employees of the firm. For this to work, we first need to specify what is a good applicant. Doing so requires expertise to determine what is important. An editor is better placed to recognize a promising author than a scientist, whereas the scientist is better placed to determine which PhD student should be receiving a scholarship. Problem specification always requires us to have some degree of familiarity with the domain and needs in question. In other words, domain expertise is a necessary condition for good problem specification. In the case of HireMe let us assume that the company is looking for someone who is productive, has a strong

Artificial Intelligence Risk Certificate

work ethic, is a good team player, and is likely to remain in the company for substantial time.

As any hiring manager would know, the choice of job criteria will affect who will apply for the job and who will get hired. Within the field of econometrics, this phenomenon is also called endogeneity, which refers to situations in which the error term model is correlated with the predictor variable. In simple terms, it refers to situations in which there are hidden or overlooked correlations that are not captured by the model in question. In our hiring case, there are selection effects in the applicant pool, because the job classification will affect who ends up applying for a particular position.

Once we have decided on the relevant criteria, they need to be translated into measurable and quantifiable features. In some cases, this translation is straightforward. Most of the time, it requires to use proxies, which we judge to be strongly correlated with the criterion in question. For instance, we cannot easily measure productivity, but we can measure how many sales a given sales agent has made within one year of hiring. Using proxies always risks the introduction of bias as they reduce complex concepts to a much less-complex, measurable construct. In our case, using total sales as an indicator of productivity, might miss that female workers are more likely to work part time and thus having a lower sales rate. Being a team player is quite hard/impossible to quantify. Yet, simply skipping this feature may lead to predictions that can hurt

Artificial Intelligence Risk Certificate

candidates, for example, teamwork was an area where women generally outperformed men.

It is tempting to think that because supervised ML algorithms learn on past data, we don't need to be explicit about the specific criteria we are looking for in a job applicant. We could simply provide our algorithm with previous applicants' CVs and a simple binary target variable that indicates whether they were hired for a given role. This is another way of specifying the problem. Now, a good candidate, by definition is a candidate that reassembles previous hires. This method bears its own shortcomings. A prominent example of the difficulties associated with this method is Amazon's attempt to create an unbiased recruiting tool.

Finally, when choosing which features to incorporate into your algorithm, sometimes less is more.

Fairness through awareness?

Intuitively, we don't want protected characteristics to influence algorithmic decision making in most contexts. For example, most people will consider it inadmissible for an algorithm to explicitly take into account the ethnicity of a criminal defendant when determining whether the defendant should receive bail. The demand that protected characteristics play no role in the decision making process is equally embedded in most legal doctrines. As a result, these frameworks currently

Artificial Intelligence Risk Certificate

prohibit the explicit use of protected characteristics by an algorithm.

Consider the following example, FTA prediction, in which JusticeGuard is a fictional algorithm that helps judges to decide whether to grant bail. It outputs the likelihood that a defendant awaiting trial will fail to appear, FTA, in court. This could be correlated to belonging to a give ethnical group. Protected characteristics should not be used as inputs. Making decisions that are explicitly based on the use of protected characteristics leas to charge on direct discrimination or disparate treatment.

Imagine we estimate the following linear relationship between FTA (Y) having a prior convition and being member of a protected group.

$$Y = \beta_0 + \beta_1 \cdot x^{prior} + \beta_2 \cdot x^{prot} + \epsilon_0$$

Where Beta 0 is constan and Epsilon is the error term. We assume that there is a correlation between belonging to the protected group and FTA. We also assume that there is a correlation between belonging to the protected group and having a prior offense. If we consider all our variables to take binary values 0 or 1 and run an OLS regression, we obtain the following:

$$\hat{Y} = 0.034 + 0.543 \cdot x^{prior} + 0.332 \cdot x^{prot} + \epsilon_0$$

This means that our algorithm predicts a 54% increase in risk if an individual had committed a prior offense, and a 33% increase if they belong to a protected group. We don't want our algorithm to make predictions on the basis of group membership so we could.

Artificial Intelligence Risk Certificate

$$Y = \gamma_0 + \gamma_1 \cdot x^{prior} + \xi_0$$

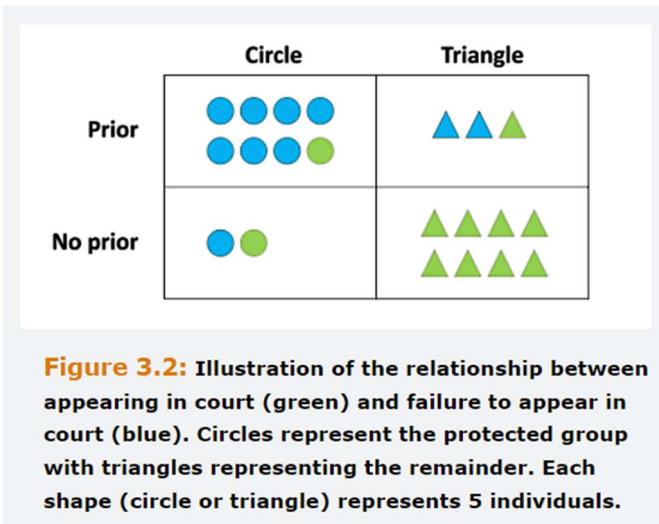


Figure 3.2: Illustration of the relationship between appearing in court (green) and failure to appear in court (blue). Circles represent the protected group with triangles representing the remainder. Each shape (circle or triangle) represents 5 individuals.

Notably, gamma 0 is not equal to beta 1. We already mentioned that X prior and X prot are correlated, which means that we can express the protected variable in terms of the correlated variable X prior

$$x^{prot} = \alpha_0 + \alpha_{corr} \cdot x^{prior} + v_0$$

Where alpha corr quantifies the correlation between the protected variable and having a prior. When we plug both equations, the coefficients have changed

$$Y = (\beta_0 + \beta_2 \alpha_{corr}) + (\beta_2 + \beta_3 \alpha_{corr}) \cdot x^{prior}$$

The coefficient of X prior is contaminated, due to the correlation between xs. Plugging the number confirms this.

$$\hat{Y} = 0.10 + 0.72 \cdot x^{prior} + \xi_0$$

Artificial Intelligence Risk Certificate

Instead of 54% we calculated having a prior now is 72%. In other words, now that we removed the protected variable from our model, the model counts having a prior more heavily for risk predictions. What it shows us is that simply removing the protected variable from our input features does not mean that the protected variable won't play any role anymore. Instead, the correlated variables are weighted more strongly in the predictions. We could remove all correlated variables, but much of the time, these variables are highly relevant for the prediction, and the algorithm would perform way worse. Other examples would be to use the average of the Protected variable, however care needs to be taken to ensure legality.

2.3.2 Data Collection and Data Composition

Both supervised and unsupervised learning rely on large amounts of data. These algorithms make evidence-based decisions, and as with any evidence-based decision, the decision can only be as good as the evidence it is based on.

Historical Bias and Feedback loops

Supervised learning algorithms rely on historical data. Even if this data is representative, it can reflect historical biases. For example, if a hiring algorithm is trained on past hiring data, and women were historically less likely to be selected for certain roles than men, this algorithm will not only repeat the same pattern but exacerbate them. This is an example of a feedback loop.

Artificial Intelligence Risk Certificate

Data Collection

One of the biggest challenges when developing AI tools, it is obtaining a training set that adequately reflects the actual statistical distribution of features within the target population. One of the most common forms of data bias is representation bias. Here, the collected sample is not representative of the target population and can result in problematic model specifications. This has been an issue in many settings, including the training of facial recognition algorithms, medical testing and others where the training was dominated by one ethical group. In those cases supervised AI algorithms will perform better on the dominant group and perform less well on underrepresented groups. Performance can be particularly low for groups that are the intersection of different, underrepresented subgroups.

Another form of bias that can occur during the process of data collection is **measurement bias**, when measurement methods are not consistent across the sample.

Data Composition

Even if measurement and sampling methods have been carefully monitored, algorithms can end up biased as a result of data composition. Algorithms deliver the statistically ideal output. Lack of data points for a certain subgroup, could lead to underperformance. It is quite important that data sets are balanced, meaning that the

Artificial Intelligence Risk Certificate

relevant minority subgroups should occupy a similar proportion of the space as majority groups. This ensures that the algorithm considers the subgroups as statistically relevant. This is different from having proportional representation. Instead, if we want the dataset to be balanced, we will increase the proportion of triangles in the data set. As becomes clear, balancing a dataset can mean that the dataset becomes less representative of the actual population.

Bootstrapping is quite useful, in which datapoints are “resampled” from an existing dataset and added back into the dataset. There are 3 methods, over-sampling data points from the minority group, under-sampling datapoints from the majority group or a combination of both.

Balancing datasets is quite harsh, especially in domains where the relevant cases are rare, such as fraud, disease prevention and so on. This issue can also be addressed via model design, for instance by ensemble methods, using multiple models, which will ensure that all the minority data would be used to train the model. Also, a cost function or a penalty for misclassification of minority class cases than for misclassification of majority class cases.

There is no one-fits-all solution to the problem of data composition, and it will be informed by legal considerations as well as the ethical and practical concerns of those involved with the algorithm's construction.

2.3.3 Model development

Some of the processing techniques can introduce bias. Word embeddings, transform words into high dimensional vectors. These ones capture the meaning of a word through the way they relate to other word vectors. Words embeddings may contain stereotypes and as a result, NLP algorithms that are based on word embeddings that contain stereotypes may incorporate them. The choice of the objective function introduces value judgement into the development of algorithms. What the algorithm optimizes for can make an ethically significant difference. Is the algorithm trying to minimize prediction errors, or does it focus on how errors are distributed.

Consider the following example, for cancer detection, algorithm A takes as input an image of a skin discoloration or mole and outputs a risk score that predicts whether the mole is cancerous. Algorithm B takes as input a brain scan and outputs a diagnosis as to whether the patient has brain cancer.

Algorithm A it seems reasonable to err on the side of caution and minimize the number of time cancer goes undetected and more erroneously characterized as harmless. In the case of algorithm B, treatment can be highly invasives and might require brain surgery. In this case, there are additional considerations, such as risks that are introduced through surgery. Another source of bias can enter in the testing stage. It is currently common practice to test algorithms against publicly available

Artificial Intelligence Risk Certificate

benchmark datasets. These datasets themselves may not be representative of the intended use population and are not necessarily a good indicator of a model quality.

2.3.3 Model deployment

There are numerous sources of harm that might arise during the deployment of algorithms. For example, algorithms that are used within different contexts than what they were designed for can lead to a loss of performance or other challenges. This applies both to different contexts as well as to the use of historical data might be outdated. To give just an example, in the context of law enforcement, an algorithm to predict recidivism may not be adequate when trying to determine the sentence length. Related, if the values of the user do not align with the values that were built into the algorithm, this might lead to harm.

Often companies will deploy their algorithms in house, instead of 3rd parties' algorithms. If their understanding of a good employee differs from the developers understanding, the algorithm will not deliver the required results. Ensuring safe and fair algorithms will require human oversight of the data collection process, the development of algorithm and its deployment. As algorithms are deployed in more consequential domains, regular auditing will become important.

3.0 Explainability, Interpretability and Transparency

As AI systems are increasingly deployed to perform tasks that are of consequence to people's lives, the twin concepts of explainability and interpretability emerge as a central concern when evaluating risks associated with AI systems.

Explainability refers to the capacity to explain in understandable terms how an AI system makes decisions or predictions, often after the fact.

Interpretability refers to the degree to which a human can comprehend and predict the model's behavior with built-in mechanisms to understand inherently how the inputs in the model affect the outputs.

For Risk professionals, understanding these concepts is essential to ensuring transparency, accountability and trust in AI systems. In environments in which AI-aided decision-making is of significant consequence, the ability to both explain and interpret algorithmic decision making becomes a cornerstone for risk mitigation as it helps us align AI systems to ethical standards, organizational values and regulatory requirements.

3.1 Black Box problem

At the core of this challenge is the difficulty of understanding how certain algorithms make decisions. Here, there are vast differences among models. Some, such as decision trees, can be interpreted so the algorithmic reasoning becomes clear to humans, but others are much opaquer. For example, neural networks become more intricate with numerous layers of and large amounts of data, understanding the decision-making process becomes an arduous if not impossible task. In some domains of application, the black box problem can pose significant challenges. When decisions made by AI systems have substantial impact on individuals or groups of individuals, the inability to explain the algorithmic decisions can lead to accountability gaps or the erosion of trust among users and stakeholders.

This problem can be divided into a technical and philosophical problem. The latter part consists of determining what counts as a good explanation. The technical part consists of finding ways to obtain good explanations.

3.2 Opaqueness

Majority of the times the problem is not that the algorithm is opaque itself, but that modes of deployment used to implement the algorithm are opaque.

Consider the following example in which a recommendation algorithm extracts personal data from social media posts, clicks, buys and visit websites to extrapolate a profile associated with a given individual. Based on this profile the algorithm makes recommendations as to what advertisement is shown to the individual.

A first form of opaqueness is **secrecy**, in which an individual subject to algorithmic profiling is not aware of having been subject to algorithmic decision making in the first place. An even more radical form of ignorance is when individuals are unaware of the very algorithm in question.

A second form of opaqueness is **confidentiality**, which refers to the opaqueness of the algorithmic decision-making process itself. Affected parties may be aware that they are subject to algorithmic decision-making, but they do not have access to the workings of the algorithms or to the reasons for why certain outcome was reached.

There are sometimes valid reason for why the inner workings of an algorithm are not disclosed. **Security** is one of them as releasing too much detail about an algorithm therefore isn't always desirable, as it might

Artificial Intelligence Risk Certificate

increase risks of fraud, scams and cyber-attacks. A second reason is **proprietary interests and trade secrets**. Keeping their algorithms secret helps companies retain their competitive advantages. For companies using third party algorithms, confidentiality can become a big challenge if it prevents them from assessing the impacts of the algorithm on their customer base. In fact, companies are recommended to ensure some access to the algorithm in question to enable them to perform their own risk assessments and provide accountability.

Finally, a third hurdle is that interrogating an algorithm often requires specialized knowledge. Even if subjects or users have access to source codes and training data, they may not be able to make sense of them, unless they have received specialized training.

3.3 Explainable AI (XAI)

Explainable AI aims to make system more understandable. AI systems are becoming increasingly complex and therefore increasingly inscrutable for human interlocutors.

Feature importance scores are a prominent technique used in XAI. It ranks the importance of input features for outcomes. These ones can not only help to scrutinize how a model makes decisions, but can help identify potential biases or errors in the model. Other techniques similarly analyze the impact of input features on the models output. Shapley Values, for example, by calculating the contribution of each input feature to the models output by calculating the marginal effect of including

Artificial Intelligence Risk Certificate

each feature and doing so across all possible feature combinations.

LUCID, locating unfairness through canonical inverse design, on the other hand, works backwards from a particular algorithm output to create a distribution over the input space conditional on the particular output, thereby revealing a model's internal logic.

Surrogate models are what their names suggest, simpler, more interpretable models that approximate the behavior of a complex AI system. **LIME**, local interpretable model-agnostic explanations, for example, approximates a complex models prediction locally with an interpretable model, such as a linear regression or decision tree.

There are several challenges associated with XAI.

Firstly, an oversimplification of a complex decision-making process distort our undesting of the decision to the point of misinterpretation.

Secondly, several XAI techniques are computationally intensive, which makes developing and deploying them more costly.

Thirdly, the above techniques are all ex-post techniques, which means that they are used to understand the models after they have been trained. Often, complex models achieve higher accuracy than simpler, explainable models. In these cases, we are necessarily faced with an accuracy interpretability trade-off that requires weighing both factors in the context of the algorithm's application.

Artificial Intelligence Risk Certificate

4.0 Autonomy and Manipulation

Ai Systems increasingly inform or even dictate choices in various consequential domains, such as medical diagnostics, policing or finance. This raises questions about the impact of AI systems have and will have on the human ability to make meaningful choices. Human autonomy is the ability to hold beliefs, make and execute decisions of our own.

4.1 Autonomy

Understanding how AI systems might affect autonomy requires us to understand what autonomy is. There are at least two dimensions to the way we use and value autonomy. The first dimension refers to our ability to have values, hold beliefs and make decisions that are in some important sense our own and not the product of distorting external factors, such as manipulation. The second dimension refers to our ability to execute these decisions and requires us to have the freedom and opportunity to do so.

Fist misconception to address is that by delegating tasks to AI systems, we automatically lose our autonomy. We tendentiously outsource tasks to other people and entities, but we are not less autonomous by doing so. As long as users and subjects have adequate control over this outsourcing, as long as there is meaningful human oversight, we remain autonomous in these respects.

Artificial Intelligence Risk Certificate

4.2 Manipulation

AI systems shape our online experience. Search algorithms determine what search engine entries show up first, and recommendations algorithms determine the ads we see. Simultaneously, these algorithms draw on large amounts of information about us and our past behavior, allowing them to make predictions about future actions and influence behavior.

Cambridge Analytica collected data from numerous Facebook users who had completed personality tests and extracted information about alleged correlations. Based on these ones, Cambridge Analytica claimed to be able to infer a Facebook user's personality profile from their online behavior, allowing them in turn to send psychologically tailored political advertising to influence vote behavior. This incident demonstrated how big data could be used by political interests to manipulate vote behavior.

Another prime example of manipulation is the Facebook emotional contagion experiment. Researchers tried to test whether emotional contagion could take place online. Without the knowledge of the users, researchers changed users' timelines to display predominantly positive or negative posts of their peers. The experiment caused public outcry because users did not consent to participate in it. In any case, it demonstrates the power of recommendation algorithms to shape our experience and, in this case, to manipulate us.

These examples demonstrate the need for robust data and security measures, but also underscore the potential to AI manipulate public opinion. The potential for misuse of AI systems must be balanced against the benefits of AI in personalization and targeting. To do that, adequate human oversight throughout the development process as well as

Artificial Intelligence Risk Certificate

some degree of regulatory oversight may be necessary. Current efforts to prevent manipulation through AI systems are mostly located in the policy domain, where regulators strive to address the issue. The European Union's AI act, addresses the risk of manipulation by imposing transparency obligations on certain systems and by prohibiting systems that use subliminal techniques on people. However, the ambiguous notion of manipulation and the difficulty of operationalizing it into a quantifiable, measurable quantity continues to hamper progress in developing and enforcing mitigation techniques.

5.0 Safety and Well-Being

Understanding the risks of AI becomes important to risk professionals. The phenomenon of automation bias and overreliance, the dynamics of human-machine interactions, specifically in mental health support systems will be covered here.

The integration of AI in healthcare and public safety comes with significant risks. For example, the critical system failure of an Uber Driverless vehicle that fatally injured a pedestrian. The vehicle systems failed to recognize the pedestrian in time, some of the systems safety mechanisms were deactivated and the human safety driver was distracted and did not react quickly enough, because of Automation bias. The incident underscores the risk of over-dependence on AI systems in critical scenarios, but also highlight the need for robust testing and validation processes. In these case, there needs to be comprehensive testing protocols that assess the systems' performance in real world scenarios.

Artificial Intelligence Risk Certificate

Additionally, automation bias is a real problem. It refers to the tendency of humans to over-rely on automated systems, leading to complacency and reduced vigilance. Another risk is skill atrophy as prolonged reliance on automation can lead to decay of essential skills, as operators move from being active decision makers to becoming passive observers. There are some ways we can mitigate them. The first is to inform users and operators of the risk of automation bias, and to ensure that they understand the limitations of AI. The second is to implement design mechanisms that require periodic human input and verification. This keeps operators and users engaged and alert. For example, Uber could have helped the driver keep her eyes on the road by using eye tracking technology in combination with alerts to ensure meaningful human control of the AI technology.

Consider another example, related to mental health, in which we have an AI-driven mental health support system, such as chatbots and virtual assistants. Although they offer patients accessibility and anonymity, we need to ask ourselves about the adequacy of these systems in handling complex emotional and psychological issues. It has been shown that AI systems can help patients with seemingly empathetic responses to their worries, better than the average human. However, if patients are told that the responses were AI generated, positive effects typically dwindle. AI does not have empathy, a crucial part of human behavior.

There is a golden rule in AI development, that is: *Just because we can, does not mean we should.*

This rule serves as a reminder that we need to reflect carefully on which issues lend themselves to automation and which issues should stay in human hands until we have found ways to overcome existing inadequacies or challenges.

Artificial Intelligence Risk Certificate

Other key rule is to involve multiple stakeholders, from different backgrounds and also policy makers for AI in critical situations. This ideally ensures that the technology is ethical, safe and appropriate. Finally, AI in critical systems requires both developers and users to be transparent and accountable. There needs to be clear lines of responsibility in case of failures or adverse outcomes. Such failures, finally, can be avoided by ensuring AI systems are regularly audited and updated.

6.0 Reputational Risk

As companies begin to deploy more AI systems they simultaneously expose themselves to novel risks and controversies that are associated with their use. When the use of the outputs of an algorithm goes against prevalent norms and values, companies face reputational risks.

6.1 Causes for AI- related reputational Damage

It has been shown that three main causes for AI related reputational damage are privacy breaches, algorithmic bias and lack of explainability. As mentioned previously on the scandal of Cambridge Analytica serves as a cautionary tale about **mishandling user data**. Incidents involving **automation bias** and unfair algorithms can equally have large consequences for the perpetrators. In 2012, it became famous that a fraud detection algorithm deployed by Dutch government to detect fraudulent child benefits claims had falsely denied and reclaimed child benefits of thousands of Dutch citizens. Finally, **lack of transparency and explainability** can equally lead to reputational losses. This relates to the

Artificial Intelligence Risk Certificate

phenomenon that customers are often not informed about their being subject to algorithmic decision making or about the reason behind that decision making. Customers might reasonably be unhappy if they are automatically denied a loan but have no way to find out why.

6.2 Types of AI-related criticism companies face

We can divide criticism of companies broadly into two categories, competence and value alignment.

Competence is criticized when there is a perception that the company lack the relevant expertise and capacities to ensure the algorithms, they deploy are safe and fair, and the data they collect, or use is stored safely. Technical failings of algorithms, unfairness or lack of explainability often fall under this category.

Value alignment is questioned when it becomes clear that the company has been fully aware of the risks and nevertheless went ahead. In the emotional contagion experiment, the lead scientist was a researcher at Facebook, therefore the company was fully informed on the experiment and the fact that users were not asked to consent was a violation of users trust and expectations. Value alignment can also be challenged if there is an obvious lack of due diligence to prevent incidents that can harm the stakeholders. Often, the distinction between the criticisms of competence and value alignment is not clear-cut. A data breach can be the result of incompetence, but it can also be the result of a company not investing enough resources and therefore not valuing the privacy of their stakeholders. In this case, the perception may be both one of incompetence and of misaligned company values.

6.3 Management and Mitigation Strategies

Continuous monitoring and evaluation is the First and most obvious strategy to prevent reputational loss from AI. **Identifying the risks and address them proactively.** This can be achieved by putting in place organizational governance mechanisms that ensures risks are regularly monitored and evaluated. There are a number of measures that can be taken to address risks from algorithmic bias and explainability. They can be used to ensure a company lives up to the high standards of its costumers base and the wider stakeholder community. They should including at least the following.

Ensuring privacy and security - Data governance policies should comply with regulations such as GDPR and include secure data storage, handling and processing practices, along with transparent data usage policies.

Ensuring Algorithmic fairness – Algorithms should be trained on diverse datasets and regularly audited for potential discriminatory patterns.

Promoting Transparency – Be as transparent as possible to your costumer base. This includes ensuring that algorithms remain explainable using XAI techniques, but also ensuring that costumers are aware of the use of algorithms. This builds both trust and accountability.

Demonstrate the will to change – Once an incident has happened it is important to limit reputational damage. This is best achieved by being as transparent as possible about what has gone wrong and why. In cases of incompetence, companies can demonstrate that they are working to fix the issue. In case of value misalignment, the problem may be rooted more deeply in governance structures or company culture. In these case, stakeholders need to be reassured that

Artificial Intelligence Risk Certificate

the underlying problems will be fixed via organizational means. Again, transparency is key to rebuilding trust.

Ethical AI frameworks – Ethical guidelines can help to prevent incidents by providing both guidance and overview of the risks. A word of warning, many AI frameworks are too vague to be helpful in cases of ambiguity. Developing ethical AI frameworks in collaboration with experts may help ensure they provide the guidance need to prevent and address risks from AI.

Stakeholder engagement and communication – Regularly engage with stakeholders, customers, employees and regulators about AI initiatives and their impacts. Clear communication prevents misunderstandings and builds trust.

Risk analysis and crisis- management plan – Companies Can perform a risk analysis, the basis of which can inform the development of a crisis management plan specifically tailored to AI-related incidents. This can includes communication strategies as well potential steps to rectify the issue.

7.0 Existential Risks

Existential risks from AI relate to scenarios in which the advent of advance AI could pose severe or even catastrophic threats to human existence. A central theme in discussions on existential risks from AI is the notion of superintelligence, which refers to an AI system that operates beyond human-level intelligence. Such a Superintelligence AI system could, in theory, act in ways that threaten human society or existence. The concern here is not just about malevolent AI, but equally about well-intentioned AI systems whose objective misaligns with human values or priorities. There are several reasons for

Artificial Intelligence Risk Certificate

one to be concerned about AI existential Risk. The first is the rapid pace of AI development. By now, a large language model can easily fool us into believing that a human has written a text. Soon, some scholars predict, AI will reach a point of **singularity**, where AI surpasses human intelligence, which in turn might lead to unpredictable outcomes. Second, in the case of a superintelligent AI, we should concern about AI alignment, which refers to the issue of ensuring that AI System's means and goals are aligned with human values. An often cited example is that of a superintelligent AI system that is programmed to make paperclips, and wipes out humanity because this would help it to make as many paperclips as possible. Finally, several scholars have drawn attention to the possibility of an AI arms race, which describes the risk emerging from different stakeholder creating powerful AI systems without adequate safety measures. This could lead to catastrophic, if not existential, global conflicts.

Just as there are people advocating for existential risk research, there are others who believe the above concerns are overstated. A key argument that is often made refers to the current state of AI development. It is true that AI has surpassed humans in some tasks, such as chess or large-scale calculations, yet the current AI is narrow, meaning is only good in a specific test and not in general terms. As AI systems become more capable and awareness of their risks grows, we will surely be able to develop the right governance framework to contain those risks. Finally, some scholars draw parallels between historical events and the initial reactions to those events, which often included hype and unfounded fears. With the introduction of the railway, for example, some thought that humans would be melt at speeds faster than 50 miles per hour. Generally speaking, practical and ethical considerations

Artificial Intelligence Risk Certificate

have historically guided the responsible integration of technology into society.

8.0 Global Challenges and Risks

These include risks associated with the potential displacement of jobs by AI, as well as the exacerbation of existing socio-economic inequalities brought on by a widening gap between those who have the skills to work alongside AI systems and those who lack them. In addition to inequality experienced at the individual level, there is also the risk of a widening gap between nations based on their relative ability to contribute to or take advantage of advancements in AI. The ability to leverage AI to mislead or influence public opinion through misinformation is an ongoing concern for governments worldwide, as it can affect democratic elections or lead to a loss of trust in institutions or the reliability of information itself.

Finally, the growth in the data economy and the widespread collection of data for the purpose of training AI models has caused data privacy to be a top concern globally.

8.1 Economic risks from AI

One of the most immediate concerns about AI is the displacement of jobs by AI. Responses are both alarmistic and skeptical. With the release of OpenAI's ChatGPT, however, the discussion about job displacement has been reignited due to the perceived potential of LLM to affect work. All big studies present alarmistic numbers for job destruction / replacement.

Artificial Intelligence Risk Certificate

When looking at past technological breakthroughs, such as the ones that lead to the industrial revolution, automation traditionally affect low-skill job sectors. This might not be the case with AI, as high-skill jobs, such as accounting equally seem to be at stake. Nevertheless, the gap between those who have the skills to work alongside AI systems and those without them could widen, further exacerbating existing socio-economic inequalities. For risk professionals, this underscores the need for strategies that address workforce re-skilling, and policies that mitigate the unequal distribution of AI's economic benefits and burdens. Predictions about job loss should be taken with care. It is difficult to predict how quickly AI will advance, and how quickly it will be ready to take over tasks. It is also too early to say how much expert human supervision will be required in practice and by law for many tasks. An it is difficult to forecast the potential for new job creation resulting from automation.

8.2 Global Inequality

Currently, most large AI firms are based in the US, and technological adoption of AI systems varies widely between countries. At this point, there is the risk that AI advancement will exacerbate global inequalities with nations that are the center of AI development leading ahead in terms of economic growth and political influence, leaving behind less technologically developed countries. This presents ethical considerations as well as risk of triggering economic and geopolitical tensions.

8.3 Misinformation Campaigns

AI technology enable sophisticated misinformation campaigns. Ranging from bots that create misleading harmful social media posts to AI generated imagery and videos, AI has the potential to mislead and influence public opinion. This poses a serious risk to political institutions and democratic elections. Managing risks from misinformation campaigns requires strict governance measures that ensure information authenticity and reliability, as well as the deployment of AI detection tools to expose AI generated content.

8.4 Privacy and Surveillance

Many AI models require large amounts of data for training. As a result, there is a strong motivation for companies to collect such data. Risks to privacy have therefore exploded, alongside the growth in the data economy.

9.0 Conclusion

The topics presented in this model underscore the profound impact AI technologies can have on individuals, companies, governments and society as a whole, and the importance of managing these risks effectively. In the next chapter, we build on the risks we've examined and explore how ethical principles and governance can guide the development and deployment of AI technologies in a way that promotes trust, safety and fairness.

Questions and Answers Module 3 from GARP – Risk and Risk Factors

1. What is algorithmic Bias?

Is a systematic deviation of an algorithm's output, performance, or impact relative to some norm, aim, standard or baseline.

2. Name three group fairness measures and provide either their informal or formal definition.

Demographic parity in which the distributions of predictions is identical across subgroups.

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$$

Predictive rate parity, the proportion of individuals correctly identified as positive is equal across subgroups (precision)

$$P(Y = 1|\hat{Y} = 1, A = 0) = P(Y = 1|\hat{Y} = 1, A = 1)$$

Equal opportunity is the proportion of positive individuals that are correctly identified as positive (Sensitivity).

$$P(\hat{Y} = 1|Y = 1, A = 0) = P(\hat{Y} = 1|Y = 1, A = 1)$$

Artificial Intelligence Risk Certificate

3. Name some possible sources of algorithmic bias

Historical bias.

Data collection (representation bias and measurement bias)

Data composition (imbalanced datasets and statistical minorities)

Model development (preprocessing techniques and choice of Objective function)

Deployment (context)

4. What is the difference between explainability and interpretability?

Explainability refers to the capacity to explain in understandable terms how an AI system makes decisions or predictions, often after the fact, ex-post.

Interpretability on the other hand, refers to the degree to which a human can comprehend and predict the model's behavior with built in mechanisms to understand inherently how the inputs in the model affect the outputs.

Artificial Intelligence Risk Certificate

5. Name three distinct ways in which the mode of deployment of algorithms can be opaque

Secrecy- individuals are not aware of the existence of the algorithm.

Confidentiality- individuals lack access to the workings of the algorithm.

Specialized Knowledge – understanding the algorithm would require specialized training.

6. What are surrogated models, why do we use them and what are the challenges associated with them?

Surrogate models are simpler models that approximate the behavior of a complex model. They are used to enhance interpretability of an AI system. Challenges are oversimplification, potential computational costs and the fact that such models are ex post techniques.

7. What is automation bias and how could it be mitigated?

Is the tendency of humans to over rely on automated systems. Mitigation techniques include informing users about the bias itself as well as to implement design mechanisms that require periodic human input and verification.

Artificial Intelligence Risk Certificate

8. True or False: Regarding bias, if a sample is representative of the population, algorithms will always perform equally well or equally bad across all subgroups.

False

9. How does group fairness differs from Individual fairness?

As opposed to individual fairness, group fairness does not consider the treatment of individuals, but instead looks at the statistical differences between groups. For example, we may ask whether there are statistical differences between the admission rate of male and female college applicants. This would be a matter of group fairness, as we are interested in whether there are certain groups that are disadvantaged by the algorithm and that share a protected characteristics.

10. Identify and describe two common XAI techniques aimed at making algorithmic decision making more explainable.

Feature Importance scores is a prominent technique used in XAI. It ranks the importance of input features for outcomes. For example, consider the case of a credit score AI system. Feature scores can reveal which factors have the most importance on the credit score. Feature importance scores can not only help to

Artificial Intelligence Risk Certificate

scrutinize how a model makes decisions but can help identify potential biases or errors in the model.

Surrogate Models are what their name suggests, simpler, more interpretable models that approximate the behavior of complex AI systems. LIME, Local interpretable model-agnostic explanations, for example, approximates a complex model's prediction locally with an interpretable model.

11. What are some challenges associated with XAI?

An oversimplification of a complex decision making process can distort our understanding of the decision to the point of misinterpretation.

XAI techniques may be computationally intensive, which makes developing and deploying them more costly.

XAI techniques are typically ex post techniques, which means that they are used to understand the models after they have been trained.

12. What are some common causes for AI-related reputational damage

Privacy breaches, algorithmic bias and lack of explainability.

Artificial Intelligence Risk Certificate

13. What are examples of economic concerns associated with AI?

The potential for displacement of jobs by AI and the potential for widening gap between those with the skills to work alongside AI and those without, further exacerbating socio economic inequalities.

Module 4 – Responsible and Ethical AI

Learning Objectives

Discuss potential benefits of implementing a practical ethics framework.

Compare and contrast consequentialism, deontology and virtue ethics.

Discuss the principles of nonmaleficence, beneficence , justice, autonomy and explainability.

Discuss sources of and strategies to address algorithmic bias and unfairness.

Describe important ethical principles related to privacy.

Discuss the current regulatory landscape and governance challenges associated with AI.

Artificial Intelligence Risk Certificate

1.0 Introduction

Responsible AI refers to the Ethical and responsible development, deployment and use of Artificial Intelligence systems. It aims to ensure that AI benefits holistically while mitigating potential risks. Responsible AI is about aligning AI systems in a variety of ways, including ethical standards, industry standards, regulation, legislation, global principles, and more to ensure that they operate in a fair, transparent and accountable manner.

Ethical frameworks help us decide how to design AI, and how to deploy it in a way that minimizes risks, harms, and wrongs. Ethical frameworks shape AI by helping us define principles to abide by, identify unacceptable biases, assess what kind of transparency is desirable, identify stakeholders and their interests, promote accountability, and make justifiable decisions about the design and implementation of AI. Different ethical frameworks help us make justifiable decisions in pluralistic societies.

2.0 Practical ethics

Is the application of ethical theory, and the development of good practices to solve real world moral dilemmas. The goal of practical ethics is to provide a concrete guidance for moral decision making and problem-solving. Some key aspects of practical ethics include.

Applied focus, in which practical ethics aims to address actual ethical dilemmas that arise in domains like medicine, law, politics, business and now AI. It can offer recommendations, not just theories.

Artificial Intelligence Risk Certificate

Multidisciplinary, in which practical ethics draws on moral philosophy, social sciences like psychology and sociology, domain expertise, law, computer science and other fields to inform the analyses of complex issues.

Context sensitivity, in which practical ethics emphasizes that situational nuances matter in moral decision making.

Pluralistic approach, which considers different ethical frameworks like consequentialism, deontology, and virtue ethics can provide insights for evaluating issues. Practical ethics may blend these approaches.

Typical decisions about moral dilemmas involve making a choice in which someone stands to lose out like, for example, when the task is to distribute scarce resources or design an algorithm that selects job candidates.

Practical ethics aims to make decisions in a way that can be justifiable to the whole society.

In short, practical ethics aims to offer actionable advice for ethical issues by applying philosophical theories, principles and methods in context, using an interdisciplinary lens, and focusing on practical guidance over theoretical debates.

2.1 Why might a firm consider a Practical Ethics Framework.

Proponents of practical ethics would argue that it can provide a foundation for decision-making that may help firms when facing difficult challenges. Furthermore, they would assert that it can also make a good business sense by providing benefits for the company, such as:

Building trust and Reputation – Will improve brand image.

Avoiding scandals – Unethical behavior, like fraud, can damage a company's reputation and bottom line.

Attracting and retaining Talent – Due to workers greater awareness lately for such topics.

Strengthening culture – Ethics programs can nurture teamwork, accountability and integrity which will boost morale and productivity.

Supporting Risk Management – Having an ethical framework undergirding policies and procedures can help companies to reduce their risk of safety failures, data breaches and regulatory non-compliance.

Encouraging and guiding innovation – An ethical framework can help guide innovation in a responsible direction that considers impacts on people and society.

Promoting long term thinking – Managing in an environment in which there are ethical considerations can help promote a long-term view.

Artificial Intelligence Risk Certificate

Summing up, proponents of practical ethics would argue that they are good for business. Regardless, ethical frameworks underlay many of the responsible AI approaches that are advanced by regulatory, quasi-regulatory, legislative bodies, consulting firms, and businesses around the world. As such, it is helpful to understand the major ethical frameworks that have developed over the years.

3.0 Ethical Frameworks

Consequentialism, which is an ethical theory that judges morality of an action based on the consequences of the action.

Deontology, which is an ethical framework that judges the morality of actions based on adherence to ethical duties and rules.

Virtue Ethics, which emphasizes virtuous character traits and living a good life.

3.1 Consequentialism

Judges the morality of the action based on the consequences of the same.

At its core, consequentialism focuses on the outcomes or results of an action to determine whether it is right or wrong. The most common form is **utilitarianism**, which aims to maximize the overall utility. Utility is often defined in terms of pleasure, happiness, or the satisfaction of desires.

Under utilitarianism, the morally correct action in any situation is the one that produces the greatest net utility for all effected.

Actions are not intrinsically moral, but they derive their moral value solely from their results. Utilitarianism is forward-looking, circumstantially relative, and focused on end consequences.

A key advantage of Consequentialism is that it provides a single, quantifiable metric for determining moral value. One shortcoming is that the choice of metric can be highly subjective, and quantification of value can be challenging. Utilitarian calculations aim to *be impartial, objective and amenable to scientific measurement* of utility. However, consequences are almost always unpredictable, and it is unclear where the calculation should stop.

Critics of utilitarianism and other forms of consequentialism argue that always maximizing utility

Artificial Intelligence Risk Certificate

can lead to actions that many consider immoral, like severely violating individual rights for the greater good. Utilitarianism struggles with situations in which utility is maximized by something most would consider unethical. The classic example is one in which a doctor has the chance to kill an unfriendly patient to use his organs to save five other patients. AI systems that violate rights with the justification of it being more efficient and therefore saving resources, could be implemented.

In response, some consequentialists grant moral weight to following general moral rules, as opposed to acts, that tend to maximize utility.

Rule consequentialists judge acts by whether they adhere to utility-maximizing rules, not only by their specific-case outcomes. This workaround address some issues with utilitarianism by ensuring rules against murder, lying and the like are upheld even when breaking them may increase utility in isolated cases. Some theorists believe that rule consequentialism collapses into deontology.

3.2 Deontology

Judges the morality of actions based on ethical duties and rules.

One of deontology's most important proponents is Immanuel Kant. According to Kant, individuals have a moral obligation to act in a way that is universally applicable and treats others as ends in themselves, rather than only means to an end. In other words, we shouldn't treat people like things. People are ends in themselves, in that they have their own values and goals as a part of being autonomous.

The categorical imperative, a fundamental principle of Kantian ethics, asserts that one should act only according to maxims that could be willed as universal law without contradiction. Kant emphasizes the importance of moral principles, rationality and a sense of duty in guiding ethical decision making, irrespective of the consequences.

Morality, in Kant's view, is grounded in reason and the intrinsic value of individuals, providing a principal foundation for ethical behavior. In his view, lying is always wrong, irrespective of potential consequences.

Most contemporary deontologists care about consequences and grant them moral weight. For deontologists, however, consequences are not the only moral consideration worth taking into account and, for most of them, there will be red lines that should not be

Artificial Intelligence Risk Certificate

crossed even when it seems like the consequences might be beneficial overall.

Rights, rules and ethical principles are all deontological in nature. They provide ethical guidance of what to do and what not to do that goes beyond consequences.

3.3 Virtue Ethics

Emphasizes virtuous character traits and living a good life, rather than rules or consequences. It has roots in ancient Greek Philosophers like Aristotle, who taught happiness comes from living a life guided by virtues.

The key question in Virtue Ethics is, *What kind of person should I be?* Rather than focusing on universal duties or maximizing utility, virtue ethicists ask what character traits we should cultivate to live well. These virtues include wisdom, courage, humanity, justice, temperance and generosity. Acting virtuously means exercising practical wisdom to moderate our emotions, appetites and behavior appropriately in each situation.

Virtue ethicists believe we should aspire to ideals of human excellence. Virtues are nurtured through practice, habit, and modeling virtuous exemplars. One way to approach ethical dilemmas from a virtue ethics point of view is to ask what virtuous agent would do in a particular situation. The agent can be thought of in the abstract or concrete example.

Artificial Intelligence Risk Certificate

Virtue Ethics sees morality as a matter of character built over a lifetime, not just discrete acts.

Critics argue that virtue ethics lacks clear guidance for moral decisions compared to duty-based or consequentialist approaches. Because different virtues can conflict, how to weigh them is unclear. Virtue Ethicists counter that practical wisdom helps navigate hard cases, and that they are in no disadvantage with respect to other theories. Moral duties can also conflict and consequences are not always comparable. Virtue Ethics also integrates well with common morality, given that most people seem to learn about morality through habituation in the context of socialization.

Modern developments in virtue ethics expand its scope beyond individual character. For organizations and societies, virtues might include justice, accountability, environmental stewardship and responsible innovation. *In the context of AI, some scholars have suggested that to build AI that is ethical, we must build in a virtue ethics way, which would imply it learning from experience and habit, like children do.* Otherwise, morality is so complex that we might never be able to code it in a top-down approach.

As mentioned before, consequentialism, deontology and virtue ethics are not mutually exclusive, and in the context of practical ethics complement one another. *The best kind of moral decision is one that accords with all three theories,* an act that maximizes the good

Artificial Intelligence Risk Certificate

consequences, respect rights, complies with ethical principles and embodies virtues.

4.0 What can AI ethics learn from medical Ethics?

Ethical concerns have a long history in the field of medicine, given its direct involvement in matters of life or death. The Hippocratic Oath, thought to have been first written in Greece between the fifth and Third centuries BC, emphasizes the importance of physicians doing no harm to patients. Since that time, there have been various attempts at formalizing a code of medical ethics and exploring issues related to medical ethics.

There was an acceleration of, and an increase attention to issues surrounding medical ethics in the 20th century, which saw the creation of several important documents, including the Nuremberg code, the declaration of Geneva, the Declaration of Helsinki and Belmont Report. Medical ethics became more fully involved in the 1970s, driven by factors including the increased concentration of medical cares and other depersonalized settings, the rising cost of medical care and increased role of government in health insurance funding, the development of patient right as an outgrowth of broader efforts around civil rights, public outrage at medical scandals such as the Tuskegee Syphilis experiment, and rapid advances in technology.

Artificial Intelligence Risk Certificate

Technological Advances posed new ethical challenges for doctors that need solutions. The advent of the mechanical ventilator, for instance, prompted a reconsideration of the concept of death and led to the development of ethics surrounding organ transplantation. Physicians were not confronted with the dilemma of warm, heart-beating bodies with non-functioning brains, who presented an opportunity for organ procurement for transplantation. Whether to take the organs of these bodies is a moral question, not a medical one. Practical needs therefor were an impetus behind the establishment of ethical frameworks, emphasizing that the responsibility of resolving ethical dilemmas should not rest solely on healthcare professionals, whose expertise lies in maintaining health rather than navigating ethical complexities.

Furthermore, with rapid advances in technology related to collection, analysis, and utilizations of personal data, along with the design of new applications, platforms and tools such as autonomous cars, novel ethical dilemmas arisen. All key stakeholders are facing these challenges and their training and experience may not be fully equip to undertake them.

5.0 Principles of AI ethics

Although most AI ethics codes contain long lists of principles, the following principles of nonmaleficence, beneficence, justice, autonomy and explainability are both relevant and common.

5.1 Nonmaleficence

This principle asserts an obligation to avoid harming other or inflicting injuries. Part of what it means to avoid harming other is a prohibition on imposing risks of harm that are not justified or that outweigh potential benefits. In other words, not only should you not go around hurting others, but you should also not impose unnecessary or unjustified risks on others.

Nonmaleficence does not prohibit all types of harm unconditionally. Some level of risk is permissible if it enables benefits that justify that risk, or if no alternatives are available. For instance, medical procedures inherently incur some risk but may still be justified by their necessity and benefits.

Importantly, it matters who is making the decision, and who will bear the brunt of the harm if things go badly. It is more ethically acceptable to impose risks on people who stand to benefit from whatever the proposed action is. For example, very risky clinical research may be morally acceptable if the research subjects suffer from a sufficiently serious ailment and stand to benefit from the

Artificial Intelligence Risk Certificate

research if it goes well. The same risky research may very well be considered morally unacceptable on healthy research subjects, or on research subjects who have an ailment that does not stand to be cured by the research. An analogous situation in AI would be considering it unacceptable for people who are at no risk of harm to impose algorithmic risks on people who do not stand to gain from those risks.

5.2 Beneficence

This principle refers to the moral obligation to act for the benefit of others. Beneficence requires taking positive steps to help others, rather than simply refraining from harm. It moves beyond nonmaleficence, which tells us not to injure others, and commands us to advance the welfare and legitimate interests of people in need. Beneficence could include acts like donating to charity, volunteering, and providing resources or assistance to improve people's lives.

A common misunderstanding is that beneficence is solely an outcome-focused, consequentialism concept. However, a duty-based deontological framework includes beneficence as an obligation we must fulfill above and beyond what may maximize utility.

There are limits to the duty of beneficence, however. No one individually can alleviate all the suffering in the world, so reasonable constraints apply. Considerations like scarce resources, competing obligations,

Artificial Intelligence Risk Certificate

reasonableness and demandingness should factor into determining the extent of our duty of beneficence. Additionally, the recipient's right to autonomy may preclude unwanted benefits that disrespect personal agency and choice. People have a right to decide what is best for them, and with some exceptions, that right usually trumps unwanted offers of beneficence. Nonetheless, within these bounds, actively pursuing the welfare and legitimate interests of others remains a key deontological duty.

In the context of AI, one way to think about beneficence is a duty that AI systems benefit humanity in some way. At a minimum, an AI system should offer solutions to problems and be designed to improve the lives of those who interact with it.

5.3 Justice

This principle refers to the moral obligation to act in accordance with principles of fairness, equality, impartiality and proportionality. In ethics, justice requires giving each person his or her proper due while upholding duties toward fairness and equality.

There are different concepts of justice. Procedural justice demands fair processes and impartiality. Distributive justice focuses on equitable allocation of benefits and burdens in society. Restorative justice aims to repair harms through reconciling victims and offenders. Interactional justice concerns respect and

Artificial Intelligence Risk Certificate

fairness between individuals. Social justice refers to just institutions in society that provide for basic rights and needs.

Justice is concerned with ensuring human rights are respected, resources are distributed equitably, opportunities are available to all, the law is applied impartially, and no one is discriminated against unfairly. Violations of justice may lead to human rights abuses, discrimination, corruption, inequality, and exploitation of vulnerable groups.

However, there are debates around what constitutes a just distribution of goods or a fair process. Different principles of justice, like egalitarianism, utilitarianism, meritocracy, or need-based allocation can conflict. There are also disagreements around what goods justice should be concerned with distributing, like resources, opportunities, power and welfare.

Despite these debates, there is a broad agreement that justice is a vital moral principle and remains a cornerstone of ethics. To enjoy legitimacy, moral decisions must be justifiable to all and align with what is fair. Another point on which there is a broad agreement is that, as a matter of justice, people should not be discriminated against for characteristics that are morally irrelevant.

5.4 Autonomy

This principle refers to the capacity of people to make their own informed, un-coerced decisions about their lives and actions. As an Ethical principle, autonomy commands respecting and supporting others' abilities to determine their own course in life.

In cases in which autonomy may be constrained because a person lacks the capacity to make rational decisions, a surrogate decision maker, may need to act in the individual's best interest.

Autonomy has roots in humanistic and existentialist traditions. It depends on capacities for self-awareness, independent decision making, critical reflection, and personal freedom. Infringing on someone's autonomy contravenes their right to life.

In healthcare, respect for autonomy is crucial. Patients have the right to voluntary informed, consent or refusal treatment. The doctor's duty is to inform the patient appropriately, and it is up to the patient to decide what, if any treatment to pursue. Coercion, deception, manipulation, and undue influence all undermine autonomy.

An ethical AI system respects people's autonomy by not using coercive or manipulative tactics to get people to act in a particular way. Technology should help people further their own life goals, as opposed to trying to further the goals of 3rd parties.

5.5 Explainability

Also called explicability, refers to the idea that AI systems, especially those with decision making capabilities, should provide transparent and understandable explanations for their actions or decisions.

One reason explainability is thought to be important is for the purposes of accountability. Decisions made by AIs can have a profound impact on individual lives. Ensuring that AI systems can explain their decisions is essential for being able to hold accountable the companies and the people that design and implement them. It can also help identify and rectify errors, biases or unfair practices.

Explainability is also thought to further trust. Trust is a fundamental component of the adoption of acceptance of AI technologies. If users and stakeholders cannot understand how a system reaches its conclusions, they are less likely to trust it. Explainability fosters trust by making AI systems more transparent and predictable.

Without the ability to explain why an AI system made a particular decision, it becomes harder to ensure that it adheres ethical guidelines and respect individuals' rights.

There are various levels of explainability in AI.

Local Explainability focus on explaining the decisions of a specific AI model on a single instance or prediction.

Artificial Intelligence Risk Certificate

Local Explanations provide insights into why a particular decision was made for a particular case.

Global explainability looks at an AI model's overall behavior and decision-making processes. It provides a more comprehensive understanding of how the model operates across various inputs.

Model-specific explainability refers to the fact that some AI models have specific explainability techniques tailored to their architecture. For example, decision trees have intuitive rules for explaining their decisions, whereas deep neural networks may require different methods. In contrast, model-agnostic methods are designed to work with any AI model, making them more versatile. They don't rely on the specific architecture or algorithms used in the model.

There is considerable debate about what exactly counts as an explanation and to whom an explanation is owed. The former partially depends on the latter because the explanations intended for experts will likely differ from the kinds of explanations that are intended for regulators or ordinary citizens.

What counts as a good explanation will likely vary depending on the kind of AI, but one popular approach is to develop counterfactual explanations. Consider a case in which an algorithm decides whether to grant loans. A counterfactual explanation might involve presenting a hypothetical scenario that contrasts with the actual decision. For instance, suppose the AI denies a loan to an individual based on a certain criterion such as

Artificial Intelligence Risk Certificate

having too little money in the bank, or earning a low salary. A counterfactual explanation could be constructed by presenting an alternative scenario, stating the conditions under which the loan would have been approved. This Contrafactual scenario helps the individual to understand the specific factors that led to the denial and provides actionable insights, such as increasing their salary or their bank savings. Contrafactual explanations contribute to transparency and help users comprehend the influence of different variables on AI Decisions.

6.0 Bias, Discrimination and Fairness

Bias within the realm of AI refers to a systematic deviation in the output or impact of an algorithm compared to a desired norm or standard. Essentially, an algorithm is considered biased when it deviates from its intended function. Suppose an algorithm is designed to identify the best job candidates for a position as an executive. If the algorithm tends to recommend for or against candidate based, not on their qualifications, but on an unrelated feature such as sex or race, then its biased because it is deviating from its intended function.

A well designed AI should align with its stated purpose, optimizing performance according to established standards. The aim is not to achieve an entirely objective algorithm, as every algorithm inherently reflects values embedded in its design. These values are shaped by the

Artificial Intelligence Risk Certificate

perspective that certain aspects are deemed valuable or important, as the algorithm strives to excel based on specific metrics. For instance, an algorithm assessing loan eligibility may prioritize a person's bank account balance, considering it relevant to optimizing loan repayment.

Not all biases are inherently problematic from an ethical standpoint. Justifiable biases can form part of a well-designed AI. Conversely, not all AIs that are statistically or legally unbiased are necessarily ethically acceptable. Even statistically unbiased algorithms can inflict unwarranted harm, such as implementing a service that charges exorbitant fees to everyone. The ethical concern when it comes to AI biases arises when biases result in unfairness, disadvantaging individuals for unjustifiable reasons in comparison to others.

6.1 Problematic Biases

Problematic Biases in algorithms often occurs unintentionally. Due to their complexity, algorithms can inadvertently incorporate ethically problematic biases. Four primary sources to contribute to biases are:

Problem Specification

Data

Model, Validation and Design

Deployment

6.1.1 Biases in Problem Specification

An algorithm may exhibit biases from its inception if the goals it is designed to achieve contain inherent problems. Operationalizing complex goals is a nuanced task, and often, the selected target variables may fail to capture real-world objectives accurately. For instance, If a bank aims to eliminate all risks and lends only to individuals that it is certain will repay, it may inadvertently end up catering exclusively to affluent individuals. In such cases, modifying the target goals to accommodate a reasonable level of risk may be necessary. Another example is an algorithm that was meant to identify patients who are sicker to assist health professionals with triage, but that was using health care expenditure as a proxy, thereby favoring not the sickest patients, but the richest ones who tend to spend more on healthcare.

6.1.2 Biases in data

If they rely on historical data, ML algorithms may tend to perpetuate biases from the past. This propensity is commonly known as historical bias. In addition, it can broadly lead to inaccurate or malfunctioning algorithms, especially when lab data does not align with real-world trends.

Consider a theoretical data frame containing all the loans that were made. That data would likely show that successful loans, have mostly been given to men, as

Artificial Intelligence Risk Certificate

women have been excluded from active participation in the banking system until recent times. If an algorithm used that full historical data set as an input, it would likely favor men, even if there is no valid reason for such a preference.

A different but related data challenge is that historical data rarely show counterfactual outcomes. This problem is called selective labels problem. For example, a company probably doesn't track the career progression of those it didn't hire. Therefore, it will never know whether it indeed hired the best candidate. A bank has data on the people to whom provide loans, but not for the rejected applicants. These ones could even turn out to be better clients than the ones who did receive a loan, but in this case the model will continue to select based on the same categories as before.

Another related but distinct kind of bias stemming from data is sampling bias. They are bias well known in science. Sampling bias arises when the data sample is not random. If the data sampled is not random, the trends shown by the population under study many do not generalize to another population.

6.1.3 Biases Modeling, Validation and Algorithm Design

Even the problem specification is sound, and the data is unbiased, biases can emerge during the modelling, validation and design phases of algorithms. Choices related to optimization functions, the application of different regression models, consideration of subgroups, and how information is presented can all introduce biases. For example, a search engine designed to help in selecting financial products may inadvertently favor popular items, perpetuating a cycle where popularity leads to increased exposure, irrespective of product quality. These biases can affect the overall fairness and effectiveness of algorithms, highlighting the importance of careful choices throughout the development process.

6.1.4 Biases in Deployment

Even if all the above are meticulously addressed, biases can still emerge when an unbiased algorithm is deployed in the real world. Consider an algorithm designed to address the risk of a person to default on a loan. Suppose that the algorithm is still undergoing testing, and its limitations are well known, prompting a cautionary advisory against relying solely on it for decision making. However, in practice, some bank employees may defer entirely to the algorithm suggestions. Some studies indicate that when human beings receive a suggestion from a computer, they often opt to defer to the

Artificial Intelligence Risk Certificate

automated system. There are a few hypotheses as to why people tend to defer to this. Automatized systems may appear to be more objective, and people know their own fallibilities and that might create self-doubt. Perhaps deferring to an algorithm shields people from responsibility. At best, responsibility seems shared, whereas going against the recommendation of an algorithm might expose people who make mistakes to harsher judgements and blame. This tendency to do as we are told by an algorithm creates unintended incentive effects, where individuals might relinquish responsibility, allowing them to attribute blame to the algorithm in case of any issue. The implementation of algorithms can thus inadvertently shape behavior and decision-making processes in unanticipated ways.

6.2 When does bias count as discrimination?

Whether bias results in discrimination in a legal sense may vary depending on jurisdiction. In general, algorithmic bias is likely to lead to discrimination when it results in disfavoring people based on their race, sex, ethnicity or any other classification protected by law. Such disadvantages typically violate legal protections, and designers and developers and deployers and risk managers and such, have an interest in taking proactive and continuous measures to protect against it.

6.3 Fairness

Fairness entails the absence of bias or preferred towards an individual or group based on irrelevant characteristics. An algorithm is considered fair when it does not exhibit problematic biases. There are two primary types of fairness, individual and group.

Group fairness involves statistical criteria, where for example, in loan distribution, statistical parity would require the demographics of approved individuals mirror the overall population, i.e, if 51% of the population are women, 51% of the approvals should be women.

Individual Fairness emphasizes treating similar individuals, similarly, even though defining similarity can pose challenges.

A significant challenge to ensure fairness in AI relates to the mathematical impossibility of automating fairness when base rates are unequal. When base rates between populations are different, it is impossible to satisfy demographic parity, predictive rate parity and equal opportunity simultaneously. Automating fairness becomes feasible only when base rates are equal, which is seldom the case. Fairness can ultimately be considered a moral or ethical judgement, not a mathematical one, and it can involve making imperfect compromises and trade-offs that might need to change in response to changing circumstances.

Fairness is not only about the outcome, but about procedure. Procedural fairness provides reassurance,

Artificial Intelligence Risk Certificate

not only that a fair outcome will be sought, but that will be sought through impartial and just processes. Considering the Justice system as an analogy, procedural fairness involves having the right structures in place to have rule of law. Outcome fairness involves making sure guilty people receive an appropriate punishment and innocent people go free. Sometimes there are mistakes, but when there is a fair process in place, those mistakes can be justifiable and there are ways to right some wrongs. In the context of AI, the challenge is to create corporate structures that can carry out both procedural and outcome fairness. For instance, having an ethics committee, that can weight consequentialism, deontological and virtue ethics considerations do develop and implement best practices can help achieve both outcome and procedural fairness.

6.4 Avoiding problematic Biases and Unfairness

There is no such thing as an objective algorithm. Given that there is a mathematical impossibility to satisfy all definitions of fairness simultaneously, the task is not to avoid biases in general, but to avoid problematic ones. There is a trade-off between accuracy and sensitivity to some extent. The most important is to be aware of trade-offs and to make decisions that are justifiable to the population at large, the stockholders, the stakeholders, regulators and those who lose out. Consequentialist considerations to be considered in the self-driving car example would include calculating the potential risk of

Artificial Intelligence Risk Certificate

accidents using different versions of AI. Deontological considerations would include taking care that an AI doesn't disfavor people within protected categories, or include safety minimums below which we would not be willing to make compromises. Virtue Ethics considerations would include putting in place processes for ethical decision-making that would result in responsible professionals and a responsible company.

Technological Solutions are being developed to assess the amount of fairness in a system. Internal Audit structures or for 3d party models demand bias report. Although this can help to identify and avoid bias, technology rarely solves bias problems on its own. If fairness cannot be automated, then algorithmic auditing for fairness cannot be automated either, or AI cannot solve the problem itself creates.

Data quality, by ensuring that the data is diverse, updated, accurate, representative and free from past discriminatory tendencies goes a long way toward avoiding biases, but again, is not a panacea on its own. In some cases is possible to use synthetic data, which does not entail privacy risk, given that the data does not come from actual individuals. The disadvantage is that sometimes isn't as precise as real data and that might differ from real data in ways that are not obvious.

Auditing is likely to be the best way to identify and correct biases. There are private companies that offer this type of service. Auditing will include using

Artificial Intelligence Risk Certificate

technological tools and statistical analysis, but also fresh and diverse look at the systems.

Ethical Committees or similar structures, by instituting forums in which possible problematic biases can be discussed, and in which decisions about trade-offs can be made considering consequentialist, deontological and virtue ethics considerations can help ensure procedural fairness and can contribute to outcome fairness.

Training for Board and C-Suite, since given the wide range of risks to which a firm might be exposed if AI models are designed and implemented in a manner that is not responsible or ethical, firm leadership should be educated to the risks associated with AI and the importance of ethical/responsible AI practices to help mitigate those risks.

7.0 Privacy and Cybersecurity

In the context of AI, someone has privacy with respect to some person or institution and in reference to some personal data point if that person or institution has no access to it. In other words, we have privacy to the extent that others don't have access to our personal information. Privacy is important because, among other things, it protect us from possible abuses of power. The more someone knows about you, the easier it is for them to interfere with your life.

Artificial Intelligence Risk Certificate

Data security is doing what is necessary to prevent unauthorized access to data. Data security includes protecting data from attacks such as ransomware, which can encrypt or destroy the data, as well as from theft, and from attacks that can corrupt the data. It covers physical security of hardware and storage devices, administrative and access controls, organizational policies and procedures and software applications.

7.1 Why is Privacy an Ethical Issue?

Because the lack of it can lead to wrongs, harms, and risks for individuals, institutions and society at large. In ethics, some wrongs are sometimes distinguished from harms. Wrongs can lead to harms, but they are immoral even when they don't. Privacy losses can harm citizens in a variety of ways. Individuals can suffer discrimination, blackmail, exposure and public shaming, identity theft and more. Privacy losses are also a potential liability to the institution. Every personal data point is a potential lawsuit, a potential fine. Finally, privacy losses can result in harm to society. The extent of personal data collection has made it relatively easy for anyone to learn of sensitive information about military personnel or politicians and blackmail them, which can endanger national security and democracy in various ways.

The more personal data is gathered, the longer is stored, and the more is analyzed the higher the risk of harm down the line. Personal data suffers from the very

Artificial Intelligence Risk Certificate

dangerous combination of being cheap to mine, very valuable, very sensitive and prone to being abused, and very hard to keep safe.

In the cyberspace, defenders are in disadvantage with respect to attackers. Whereas the attacker can choose the moment and method of attack, defenders must always protect themselves against any type of attack. If there is an attacker with enough resources and motivation, it's a matter of time before they get to the data they want. Arguably, the imposition of a risk amounts to a wrong.

Companies that collect more personal data than is needed are creating their own risk. One way to think about it is that personal data is toxic, albeit potentially highly valuable asset. It might be cheap, easy to mine and profitable, but it also exposes the company to risk of hacks, leaks, lawsuits and more. Personal data can also be expensive to manage. Given its sensitivity, it needs expensive infrastructure as well as legal teams to ensure compliance. Companies, therefore, have an incentive to consider the risks and rewards related to the collection, storage and use of personal data when they design products and services.

7.2 Principles and Good practices

Making decisions about what personal data, if any, to collect how to keep it safe, how to use it, and how and when to delete it will include a combination of technical practical capabilities as well as ethical reflection. Consequentialist considerations will include, what are the best and worst case scenarios? How can we minimize the risk that the worst happens? Just how sensitive is the data? How confident are we that we can keep it safe? Deontological considerations will include taking care to respect the moral and legal right to privacy, as well as following ethical principles like data minimization. Virtue ethics considerations will include asking we what a responsible company would do given the circumstances.

The following are the most important ethical principles related to privacy and cybersecurity to ensure best practice.

Right to privacy is generally considered a moral right. That is, for ethical reasons, we have a claim against others that, other things being equal, they do not access our personal data, whether the law in a particular country or historical moment recognizes that right or not. The right is also enshrined in constitutions around the world and the universal declaration of human's rights.

Data minimization is the most effective way to protect privacy, collecting only the data is necessary to fulfil a specific purpose. Given that one is imposing a risk on

Artificial Intelligence Risk Certificate

data subjects when one collects their data, personal data should only be collected when the benefit to the data subjects outweighs the possible disadvantages.

Contextual integrity, the use of personal data should adhere to contextual norms of privacy. When people give up their data in a particular context, they have certain expectations about how that data will be used. When we transfer the data to a different context, privacy norms are violated. For example, if a person gives personal data to a bank to carry out a transaction, that data should not be sold to a marketing company or used for another purpose. If a patient gives their data to their doctor for the purposes of receiving a diagnosis and treatment, that data should not end up in the hands of a data broker.

Data deletion, personal data should be deleted as soon as it is not necessary. Routine data deletion is a way to protect individuals, and it is also a way to keep data accurate. Personal data should not be collected within the intention of being kept forever. Having an expiry date is an element of good data-security practices.

Data security is part of complying with due diligence. It's good practice to use all technical tools available to keep data safe, from strong encryption to thorough anonymization of data and use of cryptographic methods such as differential policy. If an organization cannot keep data safe, it puts itself at risk by collecting personal data in the first place.

8.0 Governance Challenges

AI can be difficult and challenging to govern. Some of the difficulties typically considered include power asymmetries, institutional opacity, algorithmic opacity, lack of AI ethics structures, lack of national regulation, lack of international regulation, unpredictability of how these systems might change or how people might interact with them, generative AI not being truth-tracking, and worries about privacy and copyright.

8.1 Power Asymmetries

Many of the tech companies that are at the cutting edge of AI are often more powerful than some national governments and regulatory agencies. These asymmetries can lead to challenges in terms of legislative and regulatory bodies effectively responding to potential risks posed by AI and its uses.

8.2 Institutional opacity

Most AI systems have been developed by private companies that may not be subject to the same transparency requirements as public institutions or universities. As a result, the public, academic, journalist, policymakers, and regulatory agencies may have little detailed information about the practices that went into

Artificial Intelligence Risk Certificate

designing and training large language models, for example, from the datasets used to details about whether and how the systems were tested and tuned for safety.

8.3 Algorithm Opaqueness

Even with greater knowledge regarding the companies building AI, there is still a challenge related to the opaqueness of these systems. Neural networks are sometimes called “black boxes”, because often not even computer scientists can be sure of exactly the model is doing what is doing. One major reason for this lack of transparency is that AI systems like LLM are trained using a method called backpropagation, which adjusts the weights of the Neural network to minimize the error between the model’s output and the desired output. Although this method can be effective at improving the model’s performance, it does not provide any insight into how the model arrived at its decision.

The use of proxies is a common technique used by AI systems to simplify the training process by representing complex or difficult-to-measure objectives with a simpler, easier to measure metric. For example, an AI system designated to generate new articles might use word count as proxy for article quality, rather than trying to measure the quality of the content itself. By using proxies, AI systems can make progress toward a goal without needing to optimize directly for the goal itself,

Artificial Intelligence Risk Certificate

which can be a challenging and computationally expensive task. However, this approach can also introduce potential issues, such as optimizing for a goal that is not truly aligned with the system's overall objective, and if outside observers don't know what proxies the model is using, it is difficult to govern that model. Additionally, the sheer size and complexity of the models, with millions or even billions of parameters, make it difficult to trace back the processes behind the models' outputs. Different methods to audit the outputs of the systems are being developed to try to get around the difficulty of looking into the black box.

8.4 lack of AI ethic structures

If we compare AI ethics with medical ethics, the lack of structure seems evident. Medical ethics is supported by bioethicists, ethical codes, and ethics committees, among others. Every doctor must take a bioethics class and is a licensed professional. Every hospital adheres to international ethic codes. Every clinical research is overseen by an Ethics Committee. There is nothing similar when it comes to AI ethics.

Even though AI ethics is gradually becoming mainstream, a computer scientist can still go through an education without ever taking a course on AI ethics or being a licensed or certified professional. Although boards are increasingly worried about AI risks, it is still rare to see AI ethicists as board members.

8.5 Lack of National and International Regulation

At the time of writing, there are still not national laws regulating AI, except for China. There are likewise no specialized agencies overseeing AI, and no government-mandated auditing of large language models or other kinds of AI. Just as regulatory agencies have arisen for other kinds of regulation, the future may bring regulatory agencies that have the specialization needed to understand and govern AI. AI is an international technology, from its sources of data and talents to its implementation and use. It is therefore not unreasonable to expect some movement toward international regulation or agreed best practices.

8.6 Unpredictability Issues

One of the characteristics of some kinds of AI like neural networks is that they present emergent behavior and properties that were not explicitly coded into the system. These systems can therefore surprise human beings not only in their capabilities, but also in the kind of mistakes that they make. AIs can also be surprising in the way people interact with them. It can be difficult to predict the uses and misuses to which they can be subject. One potential option to minimize the risk of unpredictability is to subject AIs to randomized controlled trials, to test for their safety. Another option,

Artificial Intelligence Risk Certificate

complementary to RCTs, is to audit AIs periodically for safety and accuracy.

8.7 Lack of Truth Tracking Abilities

One kind of AI that has become most popular, LLMs are not based on an understanding of truth or knowledge of the world. Rather, they make statistical inferences and probabilistic guesses to construct responses. Given the input and their training, they are designed to give plausible responses. But plausible responses are not necessarily truthful. Even when responses are false, they can still appear plausible and convincing. This can create risks including the creation of plausible misinformation, physical safety risks and libel and other speech-related risks. Speech is a continuous area of governance, but speech created by a machine that is in turn created by a company with data that is unclear whether it was acquired lawfully can make for a governance nightmare.

8.8 Privacy and Copyright

When companies scrape data off the internet, or collect data from their users, questions related to whether they have a claim to that data arise. There is the worry that the privacy of data subjects has been violated by collecting the personal data of millions of unsuspecting internet users. It is unclear, for example, whether LLMs can comply with European's General Data Protection

Artificial Intelligence Risk Certificate

Regulation, as European Citizens are supposed to have the right to ask companies what data they have on them, to modify that data, and delete that data. It is far from clear that the companies that develop and sell access to these LLMs can comply with such data requests. Finally, there is the concern that copyright has been violated with LLMs ingesting material like books. At the time of writing there are various lawsuits in process related to these matters.

9.0 Regulatory Landscape

There is a broad consensus that AI should be regulated and the passage of AI related laws in countries around the world bears this out. There is not, however, a commonly held view of what form such regulation should take. The remainder of this module takes a high level look at the regulatory approaches and initiatives of three governmental players likely to have an impact on how AI models are developed and deployed worldwide.

9.1 The relationship between Ethics and Law

Ethics is considered a complement to the law and necessary to ground, inform and shape law. Societies tend to regulate behavior according to what they deem morally acceptable. Ethics helps one distinguish between just and unjust laws. Laws, however, are narrow in scope. They typically establish a minimal

Artificial Intelligence Risk Certificate

requirement of behavior for social institutions to function well. Ethics goes beyond that. Ethics, therefore, can be considered more ambitious than the law.

Laws allow us to have orderly interactions with one another within a framework of basic fairness. Ethics allow us to strive toward ways of life that will be most conducive to our own and others wellbeing. Even though AI ethics is gaining importance, there are some who feel that laws are also needed to govern AI. Some Laws, like European's General Data Protection Regulation, were not designed to legislate AI, but are relevant for the design and implementation of AI.

9.2 Europe

GDPR – General Data Protection Regulation 05/2018

It was designed to regulate personal data. Under GDPR, data subject have a right to receive concise and transparent information about their data, access their personal data upon request, request erasure of their personal data and object to processing of personal data from marketing or other purposes unrelated to the service being offered. The law also states that data controllers are under a legal obligation to notify within 72 hours the supervisory authority of any data breach.

Part of what was significant about the GDPR is that it applies not only if the data controllers or processor is located inside Europa, but also if it is located outside the

Artificial Intelligence Risk Certificate

European Economic Area and offers services within EEA. The extraterritorial jurisdiction has made the GDPR hugely effective across the world. It has made some international corporations improve their standards everywhere, because it is too complicated to have one system for European Residents and a different one for the rest of the world. It also looks bad to have better standards for some user than for others. The GDP also inspired some countries to come up with their own privacy legislation.

Digital Markets Act 2022

It targets large online platforms with gatekeeper status, which are defined as companies that hold a dominant position in a market and have the ability to distort competition. It includes,

Ban on Self-preferencing as gatekeepers are prohibited from favoring their own services or products over those of their competitors.

Open Access to Data as gatekeepers must provide access to their data to third-party developers and business, allowing them to create innovative services that compete with the gatekeepers' offering.

Transparency obligations as gatekeepers must be transparent about their algorithms and practices, allowing users and regulators to understand how they

Artificial Intelligence Risk Certificate

operate and identify potential anti-competitive behavior.

No Tying and Bundling as gatekeepers cannot require user to purchase additional services or features that they don't want or need to access their core services.

Interoperability of messaging services as gatekeepers must ensure that their messaging services are interoperable with other messaging services, allowing users to communicate seamlessly across platforms.

Digital Services Act 2023

Applies to online intermediaries and platforms, including marketplaces, social networks, content sharing platforms, app stores and more. It sets out obligations for these platforms.

Prevent the dissemination of illegal content as platforms must proactively identify and remove illegal content, such as hate speech, child sexual abuse, and counterfeit goods. They must also have clear and effective reporting mechanisms for users to flag illegal content.

Be more transparent about their content-moderation practices as platforms must publish clear information about their policies and procedures for identifying and removing illegal content. They must also provide users with access to their content-moderation data, allowing them to see how their content has been handled.

Artificial Intelligence Risk Certificate

Address the use of disinformation as platforms must take measures to prevent the spread of disinformation, such as providing clear labels on political advertising and promoting fact-checking initiatives.

Protect users from algorithmic bias as platforms must ensure that their algorithms are not biased against certain groups of users. They must also be transparent about how their algorithms work and how they are used to personalize user experiences.

Allow users to opt out of receiving personalized content as this very large platforms are required to allow their users to opt out of receiving personalized content.

The EU AI ACT 2024

Aims to establish a common regulatory and legal framework for AI. Has been designed to ensure proportionate risk mitigation over a range of AI functions. For instance, a company offering an AI service to screen job applicants would have to take steps to prevent their systems from unduly hurting individuals' access to opportunities. The regulation also imposes a legally binding requirement to notify people when they are interacting with a chatbot, biometric systems, or emotion recognition. Companies will also need to label deepfakes and content generated by AI, as well as design systems to make AI-generated media detectable.

Artificial Intelligence Risk Certificate

Organizations like banks and insurance companies that offer essential services, and companies that deploy AI classified as high risk, are obligated to do an Impact assessment on how AI will affect people's rights. Providers of High-risk AI must also keep thorough records of the datasets used, programming and training methodologies, and measures taken for oversight. The following systems are expected to be prohibited with just six months for companies to ensure compliance:

Biometrics categorization systems that use sensitive characteristics.

Untargeted scrapping of facial images from the internet or CCTV footage to create facial recognition databases.

Emotional recognition in the workplace and educational institutions.

Social Scoring based on social behavior or personal characteristics.

AI systems that manipulate human behavior to circumvent their free will and AI used to exploit the vulnerability of people.

Non-compliance can lead to substantial fines, ranging from 35 million or 7% of global turnover.

The AI Act established the European AI office that will oversee compliance, implementation and enforcement. It is the first body in the world to enforce bidding rules on AI. Much like the GDPR, this law likely will set new global standards.

Artificial Intelligence Risk Certificate

9.3 United States

Privacy

The US does not have a federal privacy law. However, it does have some laws that are relevant for privacy.

Health insurance portability and accountability act

Dated 1996 protects the privacy and security of health information. It applies to health care providers, health plans and other organizations that use or store electronic health information. HIPAA requires these organizations to implement safeguards to protect EHI from unauthorized access, use, disclosure, alteration or destruction.

State-level Privacy Regulations

Has a patchwork of state privacy regulations that govern how businesses can collect, use and share personal information about consumers. These regulations vary in scope and enforcement, but they are all designed to protect consumer privacy.

Cybersecurity 2021

The Executive order is a comprehensive and ambitious plan to strengthen the cybersecurity of the US against threats.

Artificial Intelligence Risk Certificate

Requires the Federal government to adopt and implement a zero-trust architecture for all federal networks and systems. Increase the security of critical infrastructure, such as energy, transportation and healthcare. Increase the security of software supply chains and promote the use of open-source software. Expand public-private partnerships to improve cybersecurity and invest in education and training for cybersecurity professionals.

AI

The US does not have any federal laws related to AI, however both the white house and several federal agencies have been actively working on the development of guidelines for the development, deployment and use of AI systems.

On 2023, was issues an EO on the safe, secure, and trustworthy development and use of Artificial Intelligence. Is a comprehensive plan to address the national security, economic, and ethical challenges posed by AI:

Promote the development and use Trustworthy and reliable AI that is aligned with US values

Address the risk of AI enabled harms, such as biases, algorithmic decision making and cybersecurity threats.

Enhance international cooperation on AI governance.

Artificial Intelligence Risk Certificate

Invest in research and development to advance AI safety, security and reliability.

Accelerate the hiring of AI professionals.

Require that developers of powerful AI systems share their safety test results and other critical information with the US government.

National Institute of Standards and Technology and AI risk management framework were developed in 2023. Its overachieving goal is to promote the responsible and trustworthy use of AI while mitigating potential harm. Focus on 4 key components, namely Govern, Map, measure and manage. Through these, organizations can build governance structures. Identify and map AI risks, measure and assess those risks and implement appropriate mitigation strategies. The framework emphasizes flexibility and adaptability, catering to organizations of various sizes and across different sectors. Its non-prescriptive approach that allows customization based on specific AI use cases and Risk profiles.

AI safety and security board, was established 2024 by the US department of Homeland Security and advises the Secretary, the critical community, other private stakeholders and the broader public on the safe, secure and responsible development and deployment of AI. The AI Board will develop recommendations to help critical infrastructure stakeholders, such as transportation service providers, and other on how to leverage AI.

Artificial Intelligence Risk Certificate

9.4 China

China has been rapidly developing and implementing regulations related to AI, covering technology, cyber security, privacy and intellectual property. One significant regulation is the Provisional Administrative measure of generative Artificial Intelligence services, which were published by the Cyberspace Administration of China in 2023. These measures apply to the use of GenAI for generating text, video, sounds and others within China territory. They impose various obligations on generative AI service providers, including the prohibition of generating illegal content, taking measures to prevent the generation of discriminatory content, and not infringing on other's rights, including privacy rights and personal information rights.

China has enacted several comprehensive Laws aimed at protecting personal information, like the Personal informational protection law and the Internet Information service Algorithmic Recommendation Management Provisions. These regulations mandate data minimization, user consent, and transparency in algorithm decision-making.

Questions and Answers Module 4 from GARP – Responsible and Ethical AI

1. Consequentialism, deontology and virtue ethics are mutually exclusive.

False, in pluralistic societies it is beneficial to take into account different ethical theories.

2.What are the five principles common to many AI ethics Codes?

Nonmaleficence, Beneficence, justice, Autonomy and explainability.

3. What is the difference between Consequentialism, deontology and virtue ethics?

Consequentialism is an ethical theory that judges the morality of an action based on the consequence of that action.

Deontology is an ethical framework that judges the morality of actions based on adherence to ethical duties and rules rather than focusing on consequences.

Virtue ethics emphasizes virtuous character traits and living a good life, rather than rules or consequences. It has roots in ancient Greek Philosophers like Aristotle,

Artificial Intelligence Risk Certificate

who taught happiness comes from living a life guided by virtues.

4. What are some best practices surrounding privacy?

Data minimization

Implementing the right to be forgotten

Giving control to data subjects

Respecting contextual integrity

Routine personal data deletion

Having strong data security

Following the provisions of GDPR.

5. What is practical Ethics.

The application of ethical theory and the development of good practices to solve real world moral dilemmas.

Artificial Intelligence Risk Certificate

6.What are some reasons for a company to consider a practical ethics framework?

- Building Trust and reputation
- Avoiding Scandals
- Attracting and retaining talent
- Strengthening culture
- Supporting risk management
- Encouraging and guiding innovation
- Promoting long-term thinking.

7.Can Fairness be automated?

False, because its mathematically Impossible to satisfy operationalizations of fairness when rates are unequal.

8. When does Bias count as discrimination.

Whether bias results in discrimination in a legal sense may very depending on jurisdiction. In general, algorithmic bias is likely to lead to discrimination when it results in disfavoring people based on their race, sex, ethnicity, age, or any other classification protected by law. Such disadvantages typically violate legal protections. All key stakeholders have an interest in taking proactive and continuous measures to protect against it.

Artificial Intelligence Risk Certificate

9. What are some factors than can make AI difficult or challenging to govern?

- Power asymmetries
- Institutional and algorithmic opaqueness
- Lack of ethical structures
- Lack of AI ethics structures
- Lack of regulation
- Unpredictability
- Lack of truth tracking abilities
- Privacy and copywriting.

10. What are some actions a company can take to avoid problematic issues?

- Use technological solutions
- Ensure data quality
- Audit algorithms
- Be inclusive
- Establish Ethics structures.

Artificial Intelligence Risk Certificate

11. What are four common sources of algorithmic bias?

Biases in problem specification

Biases in data

Biases in modeling, validation and algorithmic design

Biases in deployment

12. Differentiate between nonmaleficence and beneficence.

The principle of Nonmaleficence asserts as obligation to avoid harming others or inflicting injuries. The principle of beneficence refers to the moral obligation to act for the benefit of others. Beneficence requires taking positive steps to help others, rather than simply refraining from harm. It moves beyond nonmaleficence, which tells us not to injure others, and commands us to advance the welfare and legitimate interests of people in need.

13. Describe Explainability in the context of AI

The ethical principle of Explainability, a.k.a. explicability, has gained significant attention in the context of AI. It refers to the idea that AI systems, especially those with decision-making capabilities, should provide transparent and understandable explanations for their actions or decisions.

Artificial Intelligence Risk Certificate

14. Discuss Fairness in the context of an AI algorithm.

Fairness entails the absence of bias or preference toward an individual or group based on irrelevant characteristics, such as their race. An algorithm is considered fair when it does not exhibit problematic biases. There are two primary types of fairness, group and individual. Group involves statistical criteria, such as statistical parity. On the other hand, individual fairness emphasizes treating similar individuals similarly, even though defining similarity can pose challenges.

15. Discuss Autonomy in the context of AI

Autonomy refers to the capacity of people to make their own informed, un-coerced decisions about their lives and actions. As an ethical principle, autonomy commands respecting and supporting others' abilities to determine their own course in life. An ethical AI system respects peoples autonomy by not using coercive or manipulative tactics, to get people to act in a particular way.

Module 5 – Data and AI governance

Learning Objectives

Describe Elements of data Governance Framework

Describe Elements of a model Governance Framework

Describe steps in the model development and testing process

Discuss model validation and its importance

Discuss policies and procedures related to model governance

Describe factors to be considered when registering AI/ML applications in a model inventory

Discuss roles and responsibilities associated with model risk management.

Describe how the model review framework differs from AI/ML models

Describe the steps involved in model implementation and adaptation

Discuss Potential sources of misinterpretation of model results.

Artificial Intelligence Risk Certificate

1.0 Introduction

This module discusses data and model governance and provides a starting point to establish a firm-specific model validation framework for artificial intelligence and machine learning. In what follows, the primary focus is on the financial sector, specifically quantitative risk models, due to established formal regulatory framework around QRMIs. Nevertheless, the overarching principles presented should prove valuable and applicable to a wide range of industries.

Sound governance of models is essential for effective and prudent risk management, regardless of the types of models employed. Without it, a range of outcomes and decisions may occur that can prove detrimental. Therefore, model governance should exist across the entire model lifecycle, from model development, through performance monitoring and ultimately decommissioning.

Compared with more traditional models, sound governance of AI/ML may be even more important as it helps to ensure that technological advances do not overshadow the need for fundamental and necessary challenges to applicability, theory, benchmarking and outcome analysis. The opacity of AI/ML models makes the proper governance of the data used to train these models particularly relevant, and this topic is explored in the next section.

Artificial Intelligence Risk Certificate

To make this model more accessible to a general audience, the basic principles of model governance are studied together with some relevant issues concerning AI/ML techniques.

2.0 Data Governance

It could be defined as a system of decision rights and accountabilities for informed related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods. Broadly speaking, it encompasses the people, processes, and technologies requires to manage and protect data assets.

2.1 Data strategy: Developing Vision, and Setting goals and priorities

Data procurement and use requires a clear and approved data governance policy that is strategic and outlines operational practices. This should be include a level of authorization for non-publicly available personal data, including the specification of whether the data is for single or recurring use, and the source of the information.

The use of alternative data may require additional controls and testing due to concerns about privacy, data

Artificial Intelligence Risk Certificate

quality and accuracy, and other potential risks, as well as the general nature of unbounded data streams. Data applicability warrants an evaluation and approval process. Data treatment and relevant controls should also be specified, documented and approved by the data governance authority, such as a data governance committee. A data governance committee is typically responsible for a range of approved data practices and policies, including data collection, reconciliations, treatment, access, monitoring, and management of data stewards. Complete, relevant, transparent, and accurate data should be the primary objective.

2.2 Data Quality: Accuracy, Consistency and Integrity.

Model Risk management requirements around data quality do not change with AI/ML algorithms. However, the use of alternative data sources, while increasingly common with AI/ML implementations, warrants additional scrutiny. Verifying the accuracy, consistency and integrity of alternative data sources may be quite difficult, as the data providers may change field definitions in the historic time series or after deployment. Even well-established data sources are subject to this risk. New data sources introduce additional risk.

The pandemic demonstrated that data sources can suddenly disappear , have reporting biases, or exhibit

Artificial Intelligence Risk Certificate

previously unimagined shocks. Therefore, a key part of model monitoring is input data monitoring to alert model owners of flawed or otherwise problematic inputs. Data flaws may not be apparent from the data series alone, and outlier detection may be insufficient to stop anomalies, particularly when definitional changes are only disclosed in the footnotes of the data reported.

2.3 Data Provenance: verifying the legal right to use data, especially alternative data

The word provenance, refers to the sequence of ownership and handling of items of value. Works of fine art, wine, a long list of collectible objects usually will not be purchased without proof of the object's provenance. This term now applies to the data used to create the models.

The use of alternative data has results in scrutiny of whether the vendors of such data and the builders of models based thereon have a legal right to use this data. Copyright and intellectual property lawsuits against vendors of Generative AI modes are indicative of this issue.

The US federal trade commission has stated that any model proven to be built on data that was not properly obtained must be destroyed.

The first such legal action has already been settled with the destruction of such models. One stated hypothetical

Artificial Intelligence Risk Certificate

example would be if a company built a credit risk model using scraped data from websites without necessary permission.

The consequence of model destruction means that no model can be approved for use under model risk management guidelines without proving ownership of the input data. In the case of vendor models, the vendor will need to provide such proof and have contracts revised to indemnify model licensees from such violations. This is not a problem with some data, like that from established credit bureaus, but any use of alternative data will require scrutiny. Until the copyrights lawsuits are settled, this is a risk that must be considered for the use of GenAI models in any situation where model risk management principles apply.

2.4 Data Classification: Structure and Confidentiality

Data will be either quantitative or qualitative, such that is either measured or determined numerically, or determined non-numerically. Furthermore, these may be observed, projected or based purely on judgment. What is critical for good governance is to distinguish the classification of data being used, and ensure access and use is permissible. When data is either confidential or personally identifiable, controls must be present, operational and effective.

2.5 Metadata Management: Collection, Documentation and management

Metadata gives basic information about data, including file type, time of creation, data author, data source, file size and other relevant information. There are several distinct types of metadata, including descriptive, structural, administrative, reference and statistical metadata. Each of them provide an unique information about the data used for modeling.

A robust metadata management strategy should aim to ensure data is high-quality, consistent and accurate across various systems. Data documentation, data mapping, data dictionary, data definitions, data process flows, data relationships to other data and data structures are essential for robust metadata management. The use of comprehensive metadata management strategy should enable better informed business decisions, which is an important objective of any data governance initiative.

2.6 Data protection, security and compliance: Regulatory Aspects and Overview.

Data Security is often used interchangeably with data protection and data integrity. There are several regulatory requirements in safeguarding and protecting the data. The strategic security process, data protection plan, and procedural steps identified to safeguard the

Artificial Intelligence Risk Certificate

availability of the data, access to data, and data privacy are all part of data security. A solid data-protection strategy should be in place for safeguarding important information from corruption, malicious or accidental damage, compromise or loss. The importance of data protection increases with the amount of data created and stored. A data retention policy should also be in place and adhered to.

2.7 Data Access: Permission and Secure and Effective Data Sharing

Ensuring that data is readily accessible and shareable to those who have the requisite permission and need to access it is the most important aspect of data access. Data security, data protection, data permissions and data access controls are all part of the data-governance process.

Multiple regulations exist requiring organizations to implement appropriate security measures, including encryption, to safeguard data. By encrypting data, organizations can ensure its confidentiality and integrity, thereby helping to maintain compliance. The following are US examples:

Safeguards Rule requires institutions under FTC jurisdiction to have measures in place to keep consumer information secure.

Artificial Intelligence Risk Certificate

Gramm-Leach-Bliley Act requires financial institutions to provide customers with information about the institutions' privacy practices and about their opt-out rights, and to implement security safeguards to costumer information.

Privacy Act of 1974 governs how federal agencies can collect and use data about individuals in its systems of records. The act prohibits agencies from disclosing personal information without written consent from the individual, subject to limited exceptions including the census bureau for statistical purposes.

2.8 Compliance: Ensuring Compliance with legal and regulatory requirements when handling data

Although there are many rules within current regulations, the majority can be boiled down to three basic principles: Obtaining consent, minimizing the amount of data you hold and ensuring the rights of data subjects.

2.9 Roles and Responsibilities.

A company's board of directors plays a critical role on overseeing a firm's data governance framework and ensuring that the framework aligns with the organization's strategic objectives, risk management policies, and compliance requirements. Among other responsibilities, the board provides approval of the overall data governance framework and policies. The policies should include clear guidelines on data quality, privacy, security and compliance with relevant regulations. The data governance framework serves as the foundation for how data is collected, stored, used and shared within the organization.

The board is responsible for oversight of compliance and risk management, ensuring that the organization's data-governance practices comply with legal, regulatory, and ethical standards. This includes overseeing compliance with data protection laws, industry standards and internal policies. The board also assesses and manages risks related to data breaches, data quality issues and misuse of data. The board further ensures that the data-governance framework is regularly reviewed and updated to adapt to changing business need, technologies and regulatory requirements.

Ultimately, the board's involvement in data governance is critical to ensure that data is treated as a strategic asset, managed responsibly, and used to enhance decision-making, operational efficiency, and compliance. By providing oversight and setting the tone at the top,,

Artificial Intelligence Risk Certificate

the board can help foster a strong data-governance culture that support the organization's long term success.

3.0 Model Governance

It is important to note that the concept of risk contains subjective elements in its perception and modeling. Quantitative risk modeling encompasses three main elements.

Quantity of interest, deemed as a numerical object whose future value, referring to a specific point in time ,risk horizon, is uncertain. Examples include the value of a portfolio in ten business days, the revenue projection for a business over the next five years or the expected number of new clients by the end of the current calendar year.

Potential future scenarios, which are the scenarios that represent possible values of quantity of interest. They depict potential future outcomes, such as the value of a portfolio in ten business days conditioned on a specific investment decision, the revenue projection for a business over the next five years after implementing changes to the business model and so on. To facilitate quantitative analysis, each potential scenario is assigned to a weight, probability, indicating its relative importance compared to other scenarios.

Artificial Intelligence Risk Certificate

Risk measure, which summarizes the essential information derived from analyzing the potential future scenarios. Examples of this include the V@R and Expected shortfall and so on. Even the most basic statistical risk measures can be useful within the context of QRM. Often these statistical measures are then brought back to quantity of interest. Risk measures are frequently presented and monitored via dashboards or discussed periodically at stakeholder meetings and are crucial for evaluating potential losses, model performance, making informed decisions , managing risks effectively and complying with regulatory requirements.

In summary, risk models offer a structured approach to envision the future through scenario analysis. However, the effectiveness of this approach depends on the quality of each element. Some important considerations include:

Completeness of scenario sets, as it is challenging to anticipate every potential future scenario, especially regarding rare events. Historical data may not fully reflect these events and capturing them through expert judgement can be difficult.

Feedback Effects, as the presence of feedback can complicate matters as scenarios and subsequent decisions may influence the behavior of other market participants. As a result, scenario sets and their weights may need ongoing updates.

Artificial Intelligence Risk Certificate

Communication of results, as reports on QMs should provide a summary of the main assumptions used, ensuring transparency and avoiding complacency, and should reflect perceived risk based on perceived uncertainty and exposure.

SR11-7 by the Federal Reserve Board and the Office of the comptroller of the currency published in 2021 a Handbook on model risk management.

3.1 Model Development and testing

Institutions may decide to develop their models in house, rely on vendor models, or a mix of the two approaches. A general overview of the typical steps involved in the model development process are presented below.

Define objectives and scope by clearly defining the objective and scope of the model. Determine the specific problem or risk being addressed, the applicable products, the required data, desired outcomes and the target audience for the models' outputs.

Data collection and preprocessing by gather relevant data needed for model development. This may involve identifying and retrieving appropriate data sources, cleaning and transforming the data, handling missing values and outliers and ensuring data quality and consistency. Whether the models are AI/ML or traditional, document any data excluded along with the

Artificial Intelligence Risk Certificate

rationale. With AI/ML models, automation may lead to exclusion of data that may contain explanatory power and should be examined by a human before proceeding.

Exploratory data analysis by conducting exploratory data analysis to gain insights into the data, identify patterns or relationships, and understand any limitations or biases present. This is a key step that helps inform subsequent modelling approach and feature selection. Graphical packages for visualization, summary statistics and so on need to be considered up front. Certain models require data scaling and normalization, so any data transformation should be tested and documented.

Feature engineering by selecting and engineering the appropriate features or variables that will be used as inputs for the model. This could involve transforming or combining variables, creating new derived features or performing dimensionality reduction techniques. Sometimes the algorithm will identify key features that a human may not have detected. In this case, the features recommended by the algorithm should be carefully inspected rather than pre imposing features on the model. The human user can then review the feature importance ranking and use his domain knowledge to decide which features should be included. In addition to experience, there are several quantitative tools to support feature selection.

Model Selection allows to choose the most suitable modeling technique based on the objective, data characteristics and available resources. This may involve

Artificial Intelligence Risk Certificate

selecting from various statistical models or other techniques like time series analysis or ML algorithms. If AI/ML models are used for econometric or time series modeling, always run a traditional model alongside to benchmark results and ensure the expected parameter signs, weights, and other key model components are as expected. Be sure to test a range of outcomes to determine the boundaries/limits of the model and document any assumptions. Appropriate model performance metrics should also be defined at this point, whether this is RMSE, MSE, MA or other. Back testing procedures and frequency must also be defined and all associated assumptions should be documented.

Model training/model calibration in which we should train the selected model or calibrate the parameters using the preprocessed data. This typically involves optimizing the models parameters or hyperparameters to best fit the data and minimize errors or loss functions. Special emphasis should also be placed on training with data leakage. Use cross validation.

3.1.1 who is responsible for testing QRMs?

Naturally, programmers bear the responsibility of debugging and resolving known bugs, many of which may have already been identified during the development phase. Testers, who possess expertise and qualifications in software testing, are entrusted with the task of discovering and remaining significant bugs. Given

Artificial Intelligence Risk Certificate

the need for quantitative skills, it is common to rely partially on external testers to offer a fresh perspective and ensure comprehensive testing. However, it is crucial to avoid exploiting end user of a QRM for testing purposes, as their frustration with buggy software might impede their ability to provide useful bug reports. In many companies, though, the software developer must also bear responsibility for developing use and test case.
Good code design and documentation are key.

Testing should commence as early as possible, as identifying bugs early usually facilitates easier resolutions. Traditionally, testing follows a chronological order encompassing specific stages.

Unit tests validate that each piece of the code performs as designed.

Component tests verify the functionality of individual code sections.

Integration tests verify the interfaces between components.

System tests ensure that a completely integrated modeling system meets its requirements. This process may be repeated on a higher level as a system integration test.

Performance tests asses the speed, responsiveness and stability of the model or application under various conditions.

Artificial Intelligence Risk Certificate

Regression tests verify that the system continues to function after modifications to some of its components.

Acceptance tests determine whether the model or application meets the agreed upon requirements and is ready for deployment.

Overall, testing is not a continuous activity but is conducted on demand. However, due to changes in the QRM, its implementation or testing requirements, the demand for testing will persist. In addition, because it is unlikely that a model will anticipate every action the user would take, it is important to release in beta versions or parallel to production for some time to allow any additional and obvious bugs to surface.

3.1.2 How are White Box and Black Box tests conducted?

White box testing involves testers having access to internal data structures, algorithms, and the actual code. It may include line-by-line proofreading of the code.

Black-box testing treats the software as closed box, without any knowledge of its internal implementation.

Employing both approaches can be beneficial as white box testing is considered more effective, whereas black box testing reduces the likelihood of bias. Note that the types of test that can be conducted may depend on whether the model was developed in-house or

Artificial Intelligence Risk Certificate

externally. Many proprietary vendor models are black box.

In general, tests need to be appropriately configured, and their usefulness hinges on the adequacy and sufficiency of these configurations. Testers should not limit their configurations solely to artificial cases, as this may render they testing irrelevant in practice. Sometimes, the most valuable test configurations emerge from real-word situations. For instance, if the model implementation has reached a prototypical state where parameters and input data can be fed into it, establishing a preliminary process that automatically generates test results from available data is recommended. The less realistic the parameters and the input data the better. Documenting the experienced gained from these tests is invaluable.

3.1.3 What is the use test?

The use test examines the QRM within its context, considering human interaction, actual usage, acceptance of the model, interpretation of results, and the application of those results. The use test is closely intertwined with user testing, but its implications go beyond that. It serves as an ongoing validation tool, which may not be initiated until the QRM has been in use for a considerable period of time. The use test is qualitative in natura, and it is difficult to follow a schematic treatment. It represents a validation ideal

Artificial Intelligence Risk Certificate

rather than a specific tool. In essence, the use test evaluates adherence to a foundational principle.

Unlike other quantitatively oriented validation activities the use test revolves around people rather than numbers. It focuses on recording and conveying credible stories of acceptance, interpretation, and application of the model, whether they are stories of success or failure. Technical and process related challenges usually take a back seat. The results of QRMs are present during decision making processes, but the crucial questions are whether they are utilized appropriately, whether they prove useful, and if not, why. The use test should explore factors such as acceptance, trust, comprehensibility of results, and communication of feedback to modelers.

The results of the use test are typically present to senior management rather than documented as technical reports or spreadsheets. Consequently, it is often easier to communicate the results of a use test, minimizing the risk of it being regarded as an arduous requirement. Furthermore, the use test can contribute to improving the culture of risk modelling and risk management.

Overall, testing QRMs requires a collaborative effort between programmers and dedicated testers, early initiation of testing, a combination of white box and black box testing approaches, appropriate test configurations and a focus on user-oriented testing. Additionally, the use test offers valuable insight into the acceptance, interpretation, and application of the

Artificial Intelligence Risk Certificate

model, contributing to an improved risk modelling and risk management culture.

3.2.1 Model Validation: What is model Validation and when it should be performed?

Model validation is a critical component in identifying model risk, involving qualitative and quantitative aspects. Model validation follows a lifecycle starting with the identification of the model. Once a model is identified, it is inventoried and scheduled for an initial validation, which occurs prior to implementation and usage. Once the model is in production, routine periodic validations occur to ensure that the model continues to perform as expected. These can include annual review and more in-depth periodic baseline revalidations. A change based validation will be triggered if the model owner makes a material change. Ultimately, the model may be retired, in which case it should be stored in a retired model inventory and then decommissioned. Back testing and performance monitoring occur throughout models production usage. The model validation effort culminates in a validation report and a rating of whether the model has passed validation. Finding or issues that need to be addressed by the model developers and or owners may also result from the validation.

The frequency and intensity of validation should be determined based in the risk rating of the model. For example, a bank's high-risk models, may have a periodic

Artificial Intelligence Risk Certificate

baseline revalidation every two years, whereas lower risk models may be on a three or four year frequency, and so on.

Model validation includes:

- Assessment of appropriateness of data sources and data analysis.
- Assessment of any model assumptions and parameters.
- Identification of model limitations, strengths and weaknesses.
- Identification of approved product and users.
- Review of appropriateness of implementation platform.
- Review of use of expert judgement and overlays.
- Review of the model conceptual soundness as well as assessment of any benchmarks and best practices.
- Analysis and appropriateness of ongoing performance monitoring plans including frequency, key risk indicators, and escalation plans in the event of breaches.
- Appropriateness and adequacy of key risk indicators.
- Review the quality of model documentation.
- Review the documentation to ensure that backup plan exists for both internal and external models.
- Review controls around model implementation and usage

Artificial Intelligence Risk Certificate

- Review model implementation and access control.
- Review outcome analysis, including the appropriateness of the dependent variable and how the model supports the business objectives.
- Review the stress and scenario tests of model boundaries.
- Review of parameter stability and so forth.

Back testing need to be performed at an appropriate cadence with the model usage. For example, a VaR model required daily monitoring along with frequent backtesting to detect potential issues early on. A credit risk Model may undergo testing monthly and an asset liability management model may be tested quarterly.

It is useful to differentiate between initial validation and ongoing validation, performed periodically or on demand throughout the model's lifespan.~

Initial validation serves 3 goals:

- 1-** To ensure that the models operational feasibility has been checked, and that the model can run as intended without technical malfunctions.
- 2-** Ensuring that the model is properly documented, adheres to firm-wide standards, and includes executive summaries for essential documents.
- 3-** Ensuring that model users receive sufficient training to interpret and utilize the model's effectively.

Artificial Intelligence Risk Certificate

Proper documentation is essential for validating activities, especially when engaging with external parties such as auditors or regulators. The results of initial validation from the basis of the model official approval, sign off, emphasizing the importance of conducting thorough initial validation. Assumption need to be justified and limitations and weakness of the model must be documented.

The main goal of Ongoing performance Monitoring, is to observe whether the model remains aligned with its intended purpose, the assumptions remain valid, the data are still appropriate, and performance monitoring indicates that the model continues to perform as expected. In addition, model methodology should be reassessed periodically, to ensure that it is still in line with best practices and reflects the real world. OPM should assess if initial assumptions are still valid. Ongoing validation involves an iterative process between the modeling and validation cycles, adapting to changes in the model and repeating successful validation activity when necessary. Should use benchmarks or challenger models.

3.2.2 Should there be limits on Model Use?

Every QRM is based on underlying assumptions that should be clearly stated in the documentation and communicated to model users. However, it is important to identify model limitations and weaknesses and to establish limits on the usage of models, to prevent misuse. This can include product applicability, model access restrictions, and so forth. Rigorous assessment of the model usage through the model inventory and model landscape processes can help identify instances in which a model is being used beyond its intended domain. For example, a bank has approved a model from pricing standard IRS, it should not be used for a Bermudan Swaption without additional validation and approval.

Imposing portfolio size limits can also be valuable, particularly when new markets or instruments lack sufficient data or experience within the institution. Limits on model use should be mentioned in the model inventory. This is to ensure that the model risk ranking does not drift above the understood ranking.

The Fed SR11-7 established regulatory guidance for model risk management. Nevertheless, each institution needs to adapt these general guidelines to its specific situation. Activities in this domain depend on specific models, the size and structure of the financial institution, the established governance framework, regulatory requirements, budget constraints and the preferences of the individuals involved. Three general

Artificial Intelligence Risk Certificate

guidelines can support management, modelers, and validators in establishing a risk modelling culture.

Awareness: Be aware of the limitations and assumptions of risk modelling. Understand your company's history with QRMs, the risks they entail, and the validation processes in place. Stay informed about market practices. Recognize that the world is constantly changing.

Transparency: Transparently communicate the assumptions, limitations, and documentation of QRMs. Provide detailed documentation with executive summaries. Document the decision-making process during model development and all validation activities, including unsuccessful attempts. Engage in open communication with end users.

Experience: Learn from past modeling endeavors and apply relevant lessons. Emphasize proper project management and develop prototypes early on. Collect data and continuously improve quantitative skills. Establish and maintain libraries of reusable code. Seek input from other modelers and consider external experts for validation activities.

By adhering to these guidelines, organizations can foster a modeling culture that views validation as an opportunity for insight and value creation, rather than a mere regulatory obligation.

3.3.1 Model Governance Policies

Model risk Framework rely on written policies. The model Risk Functions should provide clear definitions and guidelines addressing key issues such as:

- Defining what constitutes a model and where the distinction lies between an end-user tool or non-model and a model. A model is often described as a simplified representation or abstraction of reality designed to understand, analyze, or predict aspects of the real world. Whether a calculation constitutes a model can be a gray area, but the presence of uncertainty may indicate that the calculation risks to the level of being a model. This makes consistent adherence to a robust model-risk framework even more important.
- Defining model risk and agreeing on a definite for internal use. The definitions should align with regulatory requirements.
- Identification and specification of a model's risk tier as this drives the detail and frequency of model validation activities.
- Establishing a definition for model validation and setting expectations for the validation process.

Artificial Intelligence Risk Certificate

- Specifying the information that model owners should provide to enable model-risk assessment.
- Documenting the institution's model-risk appetite and outlining procedures to address unacceptable levels of model risk.

These policies are often further developed at lower levels, specifically for risk models used in computing economic capital or based on risk type.

3.3.2 Model Documentation

Strong governance also includes documentation of model development and validation that is sufficient detailed to allow parties unfamiliar with a model to understand how the model operates, as well as its limitations and key assumptions.

3.3.3 Model Inventory and Model Landscape

To facilitate model risk governance, institutions should maintain a model inventory, which enhances transparency regarding the number of models in use and tracks their usage, changes and approvals. A comprehensive model inventory should be included.

Artificial Intelligence Risk Certificate

Model Owner, user and developer information.

Purpose and brief description of the model including applicable products.

Description of the model methodology.

Model Classification including whether the model is AI/ML model.

Identification of whether the model was developed in house or by a vendor.

Details about the model's use and frequency of use.

Any restrictions and limitations of the model use.

Materiality of the model use/ model results.

Overview of key assumptions, known weaknesses, and management overlays.

Locations of model documentation and validation reports.

Access to program code and relevant databases if available.

Identification of model interdependencies and whether they are upstream or downstream.

Identification of important non-models such as qualified analytical tools or end-user tools.

History of model updates, approvals and validation activities

Artificial Intelligence Risk Certificate

History of validation findings and remediation.

Description of the model performance monitoring frequency.

Identification of whether the model is used to meet regulatory requirements.

The model inventory of large, complex institutions, may contain hundreds of models, necessitating the use of suitable technology tools for management. However, the inclusion of models in the inventory depends on the applicable definition of a model for which SR11-7 offers the following guidance:

“the term model refers to a quantitative method, system or approach that applies statistical, economic, financial or mathematical theories, techniques and assumptions to process input data into quantitative estimates.”

The decision to label a model an AI/ML algorithm can be complicated, as an AI/ML algorithm may not correspond to the traditional idea of a model, at least at first glance. An example would be an AI-driven chat bot designed to resolve queries and offer tailored financial advice. Clearly an underlying model is at work, but understanding it in the context of traditional models that yield quantitative output can be a challenge.

A well defined inventory prevents critical issues from being overshadowed by numerous spreadsheets,

Artificial Intelligence Risk Certificate

focusing attention on models that have significant implications.

Models should not be viewed in isolation, as outputs from one model often serve as input for others. Transparency regarding these interdependencies is vital and can be achieved using a model landscape. This graphical or list-based representation illustrates the interactions among models, emphasizing crucial models and their impact on key business decisions. While maintaining clarity, model landscapes should avoid overwhelming users with excessive information. Aggregate Model risk is influenced not only by interdependencies among models but also by shared assumptions, methodologies, or other factors that can effect multiple models simultaneously. In addition, similar models may be used in different parts of the organization or used with different parameter assumptions.

When it comes to AI/ML applications in the model inventory, more weight must be put on some new aspects like the following ones:

Complexity of methodology and design: With AI/ML models, complexity of the model Design becomes more relevant than ever. AI/ML models Learn autonomously in the range provided by the developer and this needs to be addressed when specifying and comparing model complexity. This can include the chosen methodologies that determine the level of intrepertability or indicators of the level of transparency, such as the number of

Artificial Intelligence Risk Certificate

hidden layers in a neural network or the number of parameters.

Data usage: Data drives the complexity of the AI/ML methodology and thus the difficulty in assessing the model components. Influencing factors to be evaluated are the volume of required data or number of data features, the complexity of data structures, the quality of data and whether there are variable interactions and transformations.

Output parameters: A further decisive factor is whether the model in question is based on supervised machine learning with delimited output parameters or unsupervised learning, in which there is no direct way to evaluate output accuracy.

Model recalibration: An institution might determine if the model in question is static or requires continuous recalibration. Thereby, the complexity varies depending on whether a potential recalibration of the model would require an entire redevelopment, or if the initial model structure might be maintained and only retraining the model with recent data would be required.

Model Risk Ranking: Factors may differ across firms, but can generally be expected to include materiality, complexity, and exposure. Consideration of materiality is essential for efficient allocation of resources within the model risk management process. Prioritization can be based on various factors like economic, operational, or reputational consequences of misused models, usage by different parties, or impact on decisions, financial

Artificial Intelligence Risk Certificate

statements, or regulatory reporting. Model complexity comes into play when a model might otherwise be risk-ranked as low based on materiality, but the model complexity is high, which would elevate the risk ranking. Potential exposure should also be considered. For example, if an otherwise simple model affects published financial reports, the model risk ranking may be elevated.

While materiality, complexity, and exposure assessments are crucial, minimum standards of model risk assessment should be applied to all models regardless of their materiality, as the absence of such assessments can introduce its own model risk.

Model Performance Monitoring Metrics: Performance monitoring needs to be tailored to the model type. For example, accuracy, precision, recall, F1, Area under the ROC curve, confusion matrices, and so forth are commonly used for classification models. Regression models may be monitored with MAE, MSE, R2, RMSE...

3.4 Roles and Responsibilities

Model Risk management relies on the involvement of people in various roles and the implementation of effective devices and tools.

Board of Directors and senior management: The board of directors and senior management have the overall responsibility for establishing a model risk management framework that aligns with the broader risk management strategy of the institution. They set the model-risk appetite for the firm and ensure that qualified resources are available for developing, implementing, operating and validating risk models on an ongoing basis. Formal approval or rejection of a model falls under the purview of senior management, with validation results serving as a reliable basis for decision making. Neglecting validation can undermine the effectiveness and acceptance of models despite the investments made in their development.

Model Risk Function: To implement an effective model risk management approach, the model risk function should:

Develop and maintain a model risk management framework.

Maintain a model inventory, combined with the material aspects of the model landscape.

Set and maintain thresholds for model materiality related to business decisions.

Artificial Intelligence Risk Certificate

Set and maintain thresholds for model materiality related to business decisions.

Establish general validation standards and ensure that risk model validation activities occur.

Report validation results and other model risk-related conclusions to the board of directors.

Generally, the model risk function needs to reside at a senior enough level within the organization to have an impact on decision making and cover various risk types such as market risk, credit risk, operational risk, liquidity risk, enterprise risk, and so on. In large institutions, enterprise-level model risk management may require dedicated personnel with sufficient stature within the organization.

Model Owner: The model owner assumes a critical role in the effective framework of model risk management. In an ideal configuration, model ownership is assigned to parties using the model outputs within each business unit. This approach introduces a sense of accountability, incentivizing model users to understand the model, its assumptions, limitations, strengths and weaknesses. Typically, a model owner oversees a number of models, ensuring their proper development, validation, approval, findings, remediation, implementation, use, ongoing performance monitoring and reporting, and change management. Model owners are responsible for capturing model characterization within the model

Artificial Intelligence Risk Certificate

inventory and model landscape, helping to assess the model's materiality based on thresholds set by the model risk function, and providing all necessary information for validation and OPM and backtesting activities. In complex organizations, model ownership may be shared among several individuals or committees. Regardless, it is crucial to ensure that resources using model results fully comprehend the model's assumptions, limitations and weakness.

Model Developers: Are central to the creation of models and inherently contribute to model risk. Their decisions throughout the modelling cycle, based on considerations of alternatives, have an impact. Proper documentation of modeling decisions saves time and facilitates subsequent model validation and risk management processes. Modelers act as both producers and recipients of validation results. Feedback on errors, flaws, or inconsistencies detected during validation activities is crucial to improving models.

Model Users: Are critical to successful model validation as they can provide valuable feedback throughout validation and model performance-monitoring activities. Establishing effective feedback channels between model users and developers is indispensable. In the development phase, the translation of business language used by users into technical language understood by developers can enhance mutual understanding. As noted above, one of the model users may be assigned the role of model owner.

Artificial Intelligence Risk Certificate

Model Validators: Play a critical role in evaluating models and their performance. Validators provide independence and objectivity within the validation process, and can add value in terms of reviewing assumptions, methodologies, and so on. Validators often have prior experience in the first and third line of defense, model development, or even with vendor firms, and the validation and governance teams can bring multiple perspectives to bear on the validation. Validators communicate any errors, flaws, or inconsistencies detected during validation to model owners and developers, driving model improvement. The model risk team also decides on the model risk ranking and model risk tiering based on the materiality results recommended by model owners.

Internal and external auditors: Typically have specific roles in reviewing risk models, focusing on controls, consistency, regulatory compliance, and adherence to internal standards. Although their activities can be seen as meta-validation or validation of validation, they should not be considered the primary model risk management or validation functions. They are a vital and necessary independent check of the validation process and outputs.

Regulators: Regulatory authorities may also review risk models' validation processes and outcomes as part of their supervision. Their assessments may include examinations of internal audits, ensuring compliance with regulatory requirements. Firms should consider the

Artificial Intelligence Risk Certificate

regulatory framework when establishing internal guidelines to align with supervisory expectations.

3.4.1 Communication and Interaction

Effective communication among all relevant groups is vital. Some banks may follow written rules and deliver results according to prescribed reporting lines, and others may rely more on direct communication. Establishing a model risk committee or validation committee by risk type can ensure comparable information levels among participants and shared responsibility for difficult decisions.

3.5 Model Review and Model Changes

Model changes are an important aspect of model governance, and they should be treated with proper care and oversight. Here's how model governance typically addresses model challenges.

Change Management Process: Model Governance establishes a structured change-management process that outlines the steps and controls required for making changes to the models. This process includes activities such as change requests, impact assessments, documentation and validation with approval.

Artificial Intelligence Risk Certificate

Documentation: Comprehensive documentation is crucial in model governance to ensure transparency and accountability. All model challenges and associated information, including the reason for change, modifications made, validation results, and approvals, should be documented thoroughly.

Independent Review: To enhance objectivity and mitigate potential conflicts of interest, model governance often includes an independent review of model changes. This review is typically performed by individuals or teams not involved in the model development or modification process. They assess the changes of accuracy , reliability, compliance and potential risks.

Approval documentation: Final approval for model changes is obtained through the defined model-governance process. All approved changes should be documented in detail and stored in the model inventory, including the rationale, decision and any associated conditions or limitations.

Review of proposed changes to the model: When a change to a model is under consideration, whether it be for a new usage, the model performance indicates that a recalibration or rebuilt is necessary, or for another reason, the model owner communicate this change to the model risk management team. The team then assesses the change of materiality. If the change is material, a change-based validation is scheduled.

Artificial Intelligence Risk Certificate

Implementation and Monitoring: Approved model changes are implemented in a controlled manner, with proper documentation, including any changes to performance monitoring. Once implemented, Ongoing performance continues to be conducted to ensure that the changes achieve their intended objectives and remain compliant with regulatory requirements.

Given the growing volume of AI/ML models, the potential need for more frequent retraining models, and the greater impact of model errors, automated ongoing controls, where practical and responsible become more relevant. The review framework of AI/ML models differs from the traditional models in three notable ways.

Frequency: For AI/ML models, frequent monitoring is key. Although the FRB SR11-7 proposes on page 10 that “banks should conduct a periodic review—at least annually but more frequently if warranted—of each model to determine whether it is working as intended,” more-frequent review may be necessary for AI/ML models. There are several reasons for this, essentially stemming from the fact that AI/ML models, especially deep learning models, tend to have a higher level of complexity with many more parameters than traditional models. This complexity can lead to undesirable outcomes including unexpected behavior under certain circumstances that weren’t encountered during training. In addition, these types of models may be more susceptible to data drift, i.e. performance deterioration due to changes over time in the underlying data distribution due to evolving market conditions,

Artificial Intelligence Risk Certificate

consumer behavior or environmental factors. Although traditional models can also be affected by data drift, this is a greater concern for AI/ML models because they are often less transparent than traditional ones. For this reason, monitoring dashboards may be useful.

The frequency of assessing whether the model works appropriately might be based on the following main indicators:

Observed changes in Key input values.

Established KPI which might focus on early warning signs regarding the data or functioning of the model, such as potential biases or unachieved targets, as well as external events like regulatory, legal and technological changes.

The business volume of the model, given that the frequent assessment will lead to results that are more volatile.

Validation Stakeholders: AI/ML models are increasingly being used by, and are affecting, an even-wider range of stakeholders. These interactions involve not just input data, but the models outputs as well, and have brought heightened risk concerns around areas such as data protection regulation and ethical Policies. As such, areas that should be involved as controlling instances from the beginning of development should expand to encompass

Artificial Intelligence Risk Certificate

stakeholders from additional domains such as HR, compliance and operational risk.

Furthermore, the existing approval structure might require modifications. A higher familiarity with AI/ML models is necessary to avoid the rejection of models due to a lack of understanding of the underlying functioning.

Validation Content: The following topics gain relevance in situations where only the inputs and outputs of a model are observable, that is, black box models:

Focus on explainability: The SR11-7 request that the computer code implementing the model be subject to rigorous quality and change control procedures to ensure that the code is correct, that it cannot be altered except by approved parties, and that all changes are logged and can be audited. Because ML/AI models are often less than transparent, they may be more like black-box vendor models than internally developed models where code can be reviewed. In this case, the model owners are still responsible for understanding the inputs, outputs, benchmarking and assumptions. Explainability is key, and the outcomes of these models must be rigorously tested and compared with traditional models where possible.

Benchmarking: The comparison of the model results to classical models is recommended to understand the reason for any deviation.

Artificial Intelligence Risk Certificate

Assessment based on different data sets: The SR11-7 advice on analyzing the “in-sample fit and model performance in holdout samples” is just one of the main applicable tests for AI/ML models, see page 14 of the SR11-7.

Assessment of specific cases: As in the case of traditional models, extreme situations should be assessed. This not only encompasses stress testing through extreme input data values, but also the analysis of specific cases in which the decision has been made in favor of an obligor that was exactly on the hedge of being neglected.

Backtesting: Common sensitivity and backtesting methods like the MSE might become ineffective, given that there may no longer be a straightforward relation between inputs and outputs with general AI/ML in contrast to classical regression. K-Fold cross validation techniques might become more suitable in this context.

Reporting Component: The assessment of model outputs as part of validation to verify that they are accurate, complete, and informative and that they contain appropriate indicators of model performance and limitations is relevant regarding the detection of potential biases in the model outcomes. For example, in the case of VaR model, the Portfolio P&L distribution should be monitored in this context.

3.6 Recommendations for Establishing Model Risk Governance Framework for ML/AI

The main recommendations are:

Begin with existing model Framework: Even though ML/AI introduces new challenges for model risk management, enhanced model risk framework should not start from scratch. Financial institutions and regulators have gained much experience in risk model validation in recent years that could serve as solid basis for model governance topics related to AI/ML.

Consider the new role of Data: The new paradigm suggest that AI/ML is model free, and everything depends only on the data. Though that may not literally be true, the more important role of data needs to be addresses within model risk governance frameworks. This addresses issues related to various forms of bias, overfitting, population drift, and regime changes. The new uses of AI/ML in the financial industry will reveal many further challenges related to data.

Add new perspectives to your model inventory: when it comes to model risk classification, AI/ML will increase the relevance of ethical aspects due to data bias, explainability of model results and the role of recalibration process. To facilitate a comprehensive model risk management framework, these attributes need to be considered when filling the model inventory.

4. Model Validation

When conducting validation of a specific QRM it is essential to keep several general issues in mind.

Terminology: In the literature on model validation in quantitative finance, alternative terms such as model review, Model Quality assurance, model evaluation, and model vetting have been proposed to describe the validation process.

Focus on suitability: The ultimate goal of validation is to assess whether the model is appropriate and effectively used for its intended purpose. Initial validation is critical and helps to establish the model's credibility. After that, OPM will be performed as proposed and approved as part of the validation process. However, it is crucial to ensure that the validation activities remain within the domain of the model's intended application.

Recursive validation: Validation activities should be recursive, meaning that they should not be immune to critical examination. It is essential to ensure that the validation process does not rely on defective or inappropriate data and remains within the model's intended application domain. This emphasizes the need for a challenge of validation activities. Validation will not occur in a vacuum. There needs to be good communication among all participating parties.

No valid Mode: Because models are simplified representations of reality under certain assumed conditions, there is no perfect model. The goal of

Artificial Intelligence Risk Certificate

validation exercise is thus not to test for a valid or fully validated model, but rather to subject the model to a series of attempts to invalidate it. Successful validation implies that the model has withstood rigorous testing, although ongoing validation efforts should be continuous as new challenges may arise. Model validation is the process of rigorously testing the model, inputs, outputs, governance and so on to try to identify the extent of residual model risk and assess mitigating controls.

Usefulness versus validity: Models do not have to be perfect to be useful and used. If weakness and limitations are identified during the development or validation process, the insights gained can be utilized to improve the model or restrict its application, leading to an enhance understanding of both the model's capabilities and limitations. In this context it is important to remember that model weaknesses can also arise from regime shifts, hence the need for Ongoing monitoring and validation over time.

Effective Challenge: SR11-7 sates that the effective challenge should encompass a critical analysis by objective and informed parties that are able to identify crucial model assumptions and limitations, and to spot and communicate relevant model weaknesses. This requires independence from the model development process, a high level of competence with respect to validation activities, and a sufficient degree of influence to spark model improvements.

Artificial Intelligence Risk Certificate

By considering these general issues during the validation process, it is possible to enhance the effectiveness and credibility of the model, driving continuous improvement and adaptation.

4.1 Design: Bad choice and Misspecification, Parameter uncertainty

Designing a QRM involves making crucial decisions that determine its capabilities and limitations. Errors made during this stage may only become apparent during future crises. Let us briefly examine the motivations for developing a QRM and explore the role of model developers in the process.

4.1.1 Motivation

The motivation behind developing a new QRM or replacing an existing one is not arbitrary. The need for a QRM typically arises due to various factors.

Changing conditions: New models may be required if a business enters new markets, develops new products, or wants to respond to customer request. These models must be validated.

Regulatory Pressure: Regulatory finding that a model is not performing as intended may drive the need for redevelopment or replacement of a model.

Artificial Intelligence Risk Certificate

Recent Crisis: Losses experienced in recent crises may highlight the urgency for enhanced or more sensitivity modeling.

Internal Concerns: Senior management or other stakeholders may express discomfort with the existing quantitative modeling practices, driven by change in the business model, increased exposure to innovative products, or changing market conditions.

Innovation and Business Growth: Businesses may need to respond to new or evolving customer demands, spurring the need for new models or the extension of existing models to accommodate new products. Business may also acquire new lines of business, requiring additional models.

4.1.2 Model Developers

Developing QRMs requires personnel with expertise in quantitative fields such as mathematics, econometrics, and statistics, among others. They also need strong product knowledge. These professionals, commonly known as quants, often possess advanced academic degrees in mathematics, physics, engineering or computer science. Their scientific background enables them to tackle challenging modeling problems creatively. However, it is crucial that modelling teams involve individuals with significant experience, including exposure to past crisis, to ensure skepticism and out-of-the-box thinking.

Artificial Intelligence Risk Certificate

Finding the balance: One of the most challenging aspects of QRM design is designing and constructing a set of potential future scenarios and weights to them. During this step, modeler must be mindful of the difficulties associated with risk definition and the issues surrounding statistical usage discussed previously. For instance, the uncritical use of normal distributions and the sole reliance on correlations to model dependence in financial models has been criticized for underestimating risk and contributing to crises. Well known examples include the Black Schoel Model for Equity derivatives valuation, the Gaussian Copula model for collateralized debt obligations, and regulatory models such as the Basel framework for credit risk. Although these models exhibit robustness and transparency in their assumptions, applying them without reflection may result in significant consequences.

Strike a Balance: To mitigate shortcomings, experienced users may qualitatively adjust for the limitations of such models. It is essential to strike balance between the robustness of models and the need to account their shortcomings. Transparent communication and awareness of the potential risks involved are critical in leveraging the benefits of QRMs while avoiding potential pitfalls.

4.2 Modeling: Numerical and Statistical Issues

During the modeling process there are several choices to be made when translating a mathematical model into approaches that can be implemented on a computer.

4.2.1 Discretization

In some cases, it is more practical to describe the world using continuous variables that can take any values within a certain range. However, there may be cases where variables are discretized, assuming a finite number of values. This might occur for purposes of reducing complexity in modeling path-dependent events such as mortgage prepayments, or when modeling events that occur only at discrete time intervals. In addition, models may involve complex equations or formulas that lack closed form of analytic solutions. Numerical Approximation via discretization is a key solution methodology and should yield very close results when properly implemented.

In discretization, time intervals are divided into days, months or years, and monetary gains or losses are rounded to cents or other discrete measures. The discretization should accurately reflect the underlying physical or mathematical principles of the continuous model, ensuring that the discrete model is a faithful representation.

Artificial Intelligence Risk Certificate

While the benefits of discretization can make the model implementation feasible due to reduction in complexity, running time and memory usage, it can introduce challenges as well. Care should be taken in discretization to account for errors, stability, convergence, computational complexity, boundary conditions, numerical precision, and the like. The choice of time step is crucial, as with some systems too large a time step can lead to numerical instability, whereas too small a step can lead to excessive computational time. Consistency is also an important consideration.

Model Results obtained with discretization can differ significantly from expected results without discretization. To address this, it is important to include parameters to avoid instability in the algorithm. These parameters allow for evaluating the effect of discretization on model results in separate test environments without time and memory usage restrictions. In some cases, adaptive discretization methods can be used to concentrate computational effort in areas of the model where more precision is required, improving efficiency and accuracy.

4.2.2 Approximations

Involves replacing complex or hard to evaluate model components with alternative methods that produce similar results more conveniently. However, choosing an approximation fixes the precision, and there is no room

Artificial Intelligence Risk Certificate

for improvement except by selecting a different approximation. Common examples include the use of price sensitivities or Taylor series approximation in market risk modelling. Although these approximations simplify calculations, they are valid for simple financial instruments like plain bonds and small changes in risk factors. Including instruments based on sensitivities is preferable to excluding them entirely, as long as their limitations are understood.

When measuring the same risk over different time horizons, approximations can be useful. An example would be scaling up the VaR at one time point by multiplying it by the square root of time. This approximation assumes normal distribution and temporal independence, which may not hold in practice. It is often convenient but rarely empirically justified to assume a distribution as normal. The validity of these approximations must be examined via appropriate statistical tests and their impact measured, and control over their usage may require effort to ensure accuracy.

4.2.3 Numerical Evaluation

In contrast to approximation, numerical evaluation means evaluation of a model component with the potential for arbitrary precision. Therefore, methods for numerical evaluation are frequently equipped with parameters allowing control of precision – so discretization may serve as a tool here. These methods

Artificial Intelligence Risk Certificate

are theoretically available for all model components, backed by mathematical justifications under certain conditions. However, meeting these conditions in practice can be challenging, requiring a trade off between running time and precision. Theoretical confirmation alone is rarely sufficient, and extensive testing is necessary to supplement it.

Numerical analysis, the branch of mathematics responsible for numerical evaluation, cover computation of integrals, evaluating non-elementary functions, solving systems of equations, linear algebra products, interpolation, numerical differential equations solutions, optimization, and more. Implementing Numerical evaluation involves algorithmic description and computer system implementation. Stability and convergence are crucial factors, whereby small errors should not lead to unbounded imprecision. Floating-point arithmetic used by computers can also introduce rounding errors that may affect stability, requiring careful consideration.

4.2.4 Random Numbers

In some cases, QRMs rely heavily on probability theory, necessitating the assignment of values, realizations to random variables modeling risk factors. In cases where the distributions of random variables are not empirically given, samples must be drawn from these distributions. Generating random numbers is challenging for both

Artificial Intelligence Risk Certificate

humans and computers. Computers utilize pseudo random number generators which depend on an initial seed, as they cannot produce actual random numbers. It is typically better to rely on well designed deterministic pseudo RNGs rather than introducing randomness in the choice of RNG. The quality of an RNG is assessed using specialized test suites, considering criteria such as speed and reproducibility.

By understanding and managing discretization, making appropriate approximations, utilizing numerical evaluation methods and generating random numbers effectively, QRMs can be more robust, accurate and reliable.

4.3 Implementation: Software Engineering Data

Having made the modeling choices from the last section, the model is then ready to be implemented in software.

4.3.1 Model Implementation Tasks

These tasks constitute the setup of the model:

Model design (task M0) : The model is designed and documented in a way that translation into computer code is possible.

Artificial Intelligence Risk Certificate

Core implementation (task C0): The inner workings of the model are implemented as a black box with specified interfaces.

System implementation (task S0): The model core is integrated into a new or existing system that provides user interface, collects input data and parameters, schedules computations, feeds the core, processes output data and keeps a history of the computations performed.

4.3.2 Model Adaptation Tasks

Because no model lasts forever, reviews may detect weaknesses that demand adaptation. The subscript t will be used for the corresponding tasks to show their evolving character.

Model Adaptation (task Mt): The model and its documentation are adapted or even replaced in a way that adaptation of existing computer code is possible.

Core Adaptation (task Ct): The model Core, and perhaps its interfaces, are adapted or even replaced.

System Adaptation (task St): The system around the model core is adapted or even replaced.

Both model design and model adaptation belong to the realm of pen and paper, the other tasks to the realm of computers.

Artificial Intelligence Risk Certificate

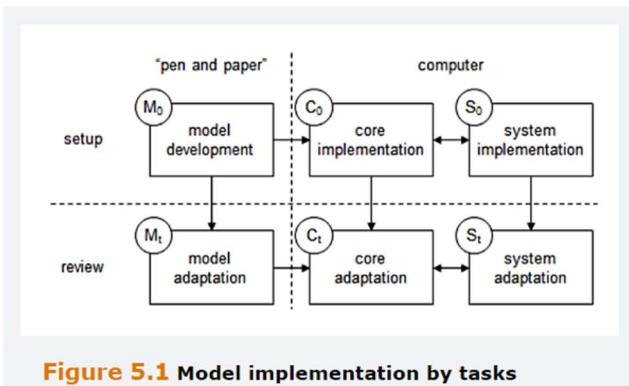


Figure 5.1 Model implementation by tasks

Examining the tasks and their interactions, particularly regarding the individuals involved, provides valuable insights. Let's start with the setup tasks, M₀, C₀ and S₀. In some cases, M₀ individuals, who are quants with a background in mathematics or other quantitative disciplines, may also perform C₀. This can be advantageous as it allows programmers to have a deeper understanding of the model they are translating into code. However, it can also result in sloppy documentation and potentially bad programming style. Although the number of errors may remain small, detecting individual errors becomes more challenging.

In general, M₀ and S₀ tasks are typically performed by different individuals, whereas C₀ and S₀ tasks may be done by the same people. Different programming languages may be used for C₀ and S₀, and S₀ often requires specialized skills in areas such as graphical front end design, large databases and hardware, such as large computer farms.

Artificial Intelligence Risk Certificate

C0 individuals often face pressure from both M0 and S0. Consequently, code produced during C0 tends to be highly optimized but difficult to read. C0 programmers may also introduce ad hoc shortcuts and hacks which can conflict with M0 intentions.

The complexity deepens during the review process. The individuals responsible for M0, C0 and S0 may have left the company, or some tasks may have been performed by external consultants. If the model is no longer understood, the adaptation task, Mt, can exacerbate the situation. Conversely, if M0 and Mt are undertaken by the same people, there is a risk of ignorance. If the core has become a black-box rather than just having well defined interfaces, the adaptation task Ct might introduce errors into previously functional C0 code. On the other hand, if C0 programmers are also involved in Ct, poor programming style could thrive. Adapting large or highly complex systems is even more problematic due to their size and/or complexity, and sometimes replacement may be the only feasible option.

Further refinements to consider include over optimism, inadequate budgeting, midstream redefinition of goals by end users or the systems group, misunderstandings, miscommunication, lack of understanding from upper management, egoism, programming errors, lack of software quality assurance, inappropriate design architecture, disparate hardware, incompatible databases and programmers being reassigned. The subsequent analysis will delve more deeply into programing errors, a.k.a bugs.

Artificial Intelligence Risk Certificate

Division by zero: A variable occurring as a divisor in an equation is assumed not equal to zero.

Arithmetic overflow: A calculation produces a result that cannot be represented by the data type of the variable which the result is assigned.

Buffer Overflow: A program tries to write to memory that has not been allocated.

Problems with Loops: Does counting start at zero or at one, and is the condition for termination of the loop < or <= and does the loop always terminate?

Problems with recursion: Does the recursion always terminate?

Bugs may cause a program to:

To Crash Regularly – Which makes it easier to locate the bugs.

To crash only occasionally and irreproducible – Which does not help to locate the bugs, but at least indicates that something is wrong.

Not to crash but to produce weird-looking results – Which is problematic because these might also be a consequence of bad model design or bad data.

To produce inconspicuous but nevertheless wrong results – Which may be catastrophic.

4.4 Processes and Misinterpretation of Results

Having implemented the model on a computer, the next crucial step involves running the model software and using the corresponding model results.

4.4.1 Running a QRM

Can be challenging, involving tasks such as data gathering, processing, feeding into the model, producing results and further analysis and reporting.

Robustness: QRMs can become complex, intertwining with multiple systems. This complexity increases the risk of failure and should be minimized. Efforts should be made to eliminate exotic systems or to standardize interfaces. It is important to avoid reinventing the wheel, when setting up QRMs. Achieving compatibility over time can be challenging particularly when historical data are required for calibration, but certain risk factors may have been introduced after the relevant historical period, rendering the data unavailable.

Speed: The total response time required for QRMs varies greatly. Credit risk Models with quarterly reporting and extensive data gathering processes may take weeks, whereas market risk models with daily reporting and real time pre deal checking demand faster processing. The overall running time is determined by the slowest task, so parallel processes should be monitored to avoid delays caused by individual slow systems. Over

Artificial Intelligence Risk Certificate

optimization for specific hardware can be risky, as model performance may be affected if the hardware becomes obsolete or support for it is discontinued.

Flexibility: A desirable aspect of a QRM is its ability to be run in two ways, scheduled runs and ad hoc runs. Scheduled runs provide regular numbers with minimal manual intervention, utilizing stored environments for quick recovery of computation. Emphasis is placed on process and IT related reliability and safety. Ad hoc runs, on the other hand, provide additional information on short notice, usually requiring manual intervention and manipulation of input data. Staff responsible for ad hoc runs should be knowledgeable about the model's interfaces and allocate appropriate time resources. However, it is important to monitor and prevent a gradual shift from scheduled mode to ad hoc mode, ensuring that model results are correctly interpreted and accepted.

4.4.2 Misinterpreting the results of QRM

Can have significant implications for decision making and risk management.

Lack of comprehension: Misinterpretation often occurs when users lack a comprehensive understanding of the risk model, its underlying assumptions and the limitations of the outputs. It is important for users to have the necessary knowledge and expertise to interpret the results correctly.

Artificial Intelligence Risk Certificate

Neglecting Model Limitations: Risk models have inherent limitations, and users should be aware of these limitations when interpreting the results. Models may assume certain conditions or may not fully capture all relevant risks, and failing to acknowledge these limitations can lead to a flawed understanding of the associated risks.

Overemphasis on point estimates: QRMs often provide estimates with a certain level of uncertainty. Misinterpretation can occur when users focus solely on the point estimates without considering the associated range of potential outcomes or probabilities. Understanding the probability distribution and ranges of potential outcomes is crucial for proper interpretation. Providing confidence intervals and probabilistic interpretation of model outputs in both the development and validation stage can be helpful in this regard.

Failure to consider qualitative factors: QRMs tend to focus on quantitative measures, such as probabilities, statistics and numerical outcomes. However, qualitative factors should not be overlooked. Factors such as market conditions, regulatory changes, and other external influences can significantly affect risk, and their exclusion from the interpretation can lead to a distorted understanding of potential outcomes.

Ignoring the context of the decision: Misinterpretation can occur when QRM results are considered independently, without considering the specific context

Artificial Intelligence Risk Certificate

of the decision being made. Risk models should be used as a tool to inform decision making within the appropriate context, incorporating broader business strategies, risk appetite another relevant factors.

Confirmation Bias: Users may unintentionally interpret QRM results in a way that aligns with their preconceived beliefs or desired outcomes. Confirmation bias can cloud judgement and compromise the objective interpretation of the model results. It is essential to approach a QRM results with an open mind and a commitment to unbiased analysis.

Fostering a culture that encourages open dialogue, critical thinking, and cross functional collaboration can enhance the interpretation process and guard against misinterpretation biases.

4.5 Final Thoughts

There are several main aspects in establishing an effective validation framework taking AI/ML applications into account.

Team qualification: The validation team, in addition to having knowledge of and hands-on expertise in AI/ML techniques, ought to have strong foundations in product, econometric, statistical and computer science knowledge, as well as a proactive approach to staying abreast to the latest AI/ML advancements. A deep understanding is needed beyond the traditional

Artificial Intelligence Risk Certificate

statistical and computer science knowledge, which may necessitate specific training, the incorporation of specialist in this domain and if need be the use of external experts.

End to end review: The entire model validation framework needs to be reviewed and adapted to the likelihood that AI/ML algorithms will eventually be subject to regulatory review and approval. This will not only raise challenges in model design, but will have implications related to data, implementation, monitoring, documentation and use.

More complex is not always better: An appropriate balance needs to be found between the model performance and all the other factors. Furthermore, validation teams need to find a balance between regulatory-only-driven position and a strictly performance-oriented approach. Striking this balance requires a nuanced and insightful cost-benefit analysis.

Questions and Answers Module 5 from GARP – Data and AI Governance

1. When developing and testing a model, the developer should be sure to calibrate the model against the full data set to obtain the best fit and predictive capabilities?

False, the dataset should be split into training and testing, or also validation segments. Training the model on the full dataset will lead to overfitting, with poor performance expected on new unseen data.

2. Steps in the model development process include defining the model objective and scope, collection of data and preprocessing, feature engineering, choice of the appropriate model and validating the model?

False, Model validation is a separate process that occurs independently of, and generally following the MDP.

3. A good model testing plan includes system integration tests, performance tests and acceptance tests.

True, includes unit tests, components test, integration test, system test, system integration test, performance test, regression test and user acceptance test.

Artificial Intelligence Risk Certificate

4. If a model is considered a black box, such as a proprietary model provided by a third-party vendor, unrealistic and extreme data test cases such missing trades or incorrect data entry should not be tested?

False, whether the model is a white, gray or black box, the less realistic the parameters and input data, the better.

5. The model validation team is responsible for data selection and specification of any transformations, review of the model conceptual soundness, review of the quality of the model documentation and testing parameter stability and robustness under various scenarios?

False. Testing that stability of parameters and robustness to various scenarios is performed by the developers, who are also responsible for data selection and specification of any new transformations. Model validation just ensures that these have been done appropriately and documented.

6. Outcomes from models should be considered independently, without consideration of the context of the decision being made to ensure that all decision makers will arrive at the same results with the same input information.

Artificial Intelligence Risk Certificate

False, the context of the decision is critical, as misinterpretation can occur when model results are considered independently, without considering the specific context of the decision being made. Risk models should be used as a tool to inform decision making within the appropriate context, incorporating broader business strategies, risk appetite and other relevant factors.

7. TIME sensitive trading model which provide immediate prices for complex derivatives will be used as inputs to the Firm's VaR mode, which is intended to be run overnight on each trading day. In assessment of the options available for implementation, the head of trading desk argues that accuracy is more important than speed so no approximations on the VaR calculation should be used. Is this True or false?

False, VaR calculations are not as time sensitive as pricing models used for intraday trading and numerical approximations are often employed. For example, scaling up a time horizon is frequently used.

8. After a model has been designed, built and placed into production, a periodic model validation detects that the assumptions that the model was trained on are outdated and that the model owner has been compensating for this with Overlays. However, the model must be used for stress testing so model

Artificial Intelligence Risk Certificate

validation recommends that redevelopment of the model to include the new assumptions or a completely new development. However, the model developer stated that a rebuilt won't work on the existing infrastructure which cannot be changed, and that therefore a new development will have to occur as part of the system adaptation task. Is the statement true or false?

False. System adaptation describes a situation where the system around the model core is adapted or even replaced. This is in opposition to the system implementation task where the model core is integrated into a new or existing system. In this case, a core adaptation is first required.

9. After an equity portfolio allocation model was designed, built and placed into production for a firm's customer, a subsequent model validation finds that the model is also being used for crypto, which the model wasn't trained on, and which behave differently from equities. The model owner has been compensating for this by scaling the risk up by an additive factor based on the proportion of crypto vs. non cryptos in the portfolio. As the underlying code to perform portfolio allocation can be adapted to include cryptos without much additional work, the model validation recommends enhancements of the existing model be adapted to include risk drivers for crypto prices and volatilities. The existing user interface is already used

Artificial Intelligence Risk Certificate

by customers so will not be changed. The Implementation task that best fits this requirement is?

Core Adaptation - Once the implementation has been chosen and the model implemented, reviews may detect weakness that demand adaptation. Core adaptation is the process where the model core, and perhaps its interface, are adapted or even replaced. Since it sees that the existing code can mainly be reused with adaptations and a new data feed added to the interface, Core Adaptation is the best answer. There will be no system adaptation

10. Some fraud and AML models operated by a vendor require access to customer data to detect unusual transaction patterns that might trigger compliance alerts. A bank using the vendor models suspects that one of their customers is engaging in fraudulent behavior and decides to check their account history for transactions matching certain patterns. One concern is that the customer is trying to circumvent triggering suspicious activities reports, by opening multiple accounts at different banks under slightly different names and phone numbers which are used to move money around between accounts. The bank runs a query of all bank customers requesting name, social security, phone number, address and company names and sends this information history across all banks using this service. By doing this, is the bank in violation of data governance best practices?

Artificial Intelligence Risk Certificate

True, the bank needs to be extremely careful when sending non-publicly available costumer information to anyone, even their own vendors. Approval from the bank's data governance committee must be sought, and specification of whether the data is for single or recurring use must also be included.

11. A bank's model developer is building a credit scoring model to inform decisions on the initial credit line to extend once an applicant has been approved for the bank's new global travel credit card. The developer decides to use an online generative AI platform to help with cross validation and hyperparameter tunning to calibrate the model, pinpoint flaws in their code and suggest unit tests that can be included. The AI will also help with code documentation and change management. The developer informs the AI that as the potential for overfitting is a concern, it is to select the train appropriate train/test split and ensure that there is no data leakage. The developer prepares and uploads a CSV file of consumer features including, age, employment status, zip code, salary, credit bureau score, phone number and social security number to the AI platform and makes the request. The developer also pastes in the code with database connection script which the AI analyzes and returns with the requested hyperparameter tuning, code correction, documentation, unit tests and so on.

Artificial Intelligence Risk Certificate

The developer tests the model and the results look very good. When informing his manager of the great work AI has done, the manager expresses immediate concerns around data privacy. The developer states that the AI platform is completely private and there should be no concerns. The Developers statement is :

False, users of online generative AI systems need to be very careful on uploading any proprietary data at all. Anything uploaded to an AI platform has the potential to be exposed at some point or used in training successive models. The manager is absolutely correct.