AI, Machine Learning and their Applications

A *Brief* Comprehensive Overview

Ekaba Bisong

Data Science Lead,
T4G Limited

Table of Contents

1. Overview: Machine Learning (ML)
2. Principles of Supervised Learning
3. Supervised Learning Algorithms
4. Unsupervised Learning Algorithms
5. Reinforcement Learning Algorithms
6. Applications of Machine Learning
7. Conclusion

Overview: Machine Learning (ML)

What is Machine Learning

Machine learning is a field of study that gives computers the ability to learn without being explicitly programmed.
- Arthur Samuel, (1959)

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.
- Tom Mitchell, (1997)

Learning from Data

- ML: trains a computer to learn and make inferences from data
- Data: Expensive to collect and “clean”

Without data, there is no learning.

Schemes of Learning

ML: Categorized into (at least) three components based on its approach.

The three predominant schemes of learning are:

- Supervised
- Unsupervised
- Reinforcement Learning

Supervised Learning: Overview

- Supervised Learning: Each data point is associated with a label
- Goal: Teach the computer using this *labeled* data
- Learning: The computer learns the patterns from data
- Inference: Makes decisions about “unknown” samples

Supervised Learning: Overview

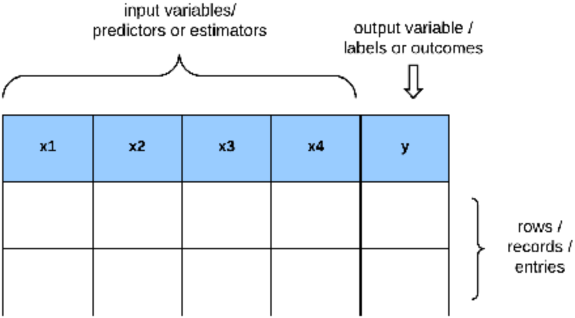
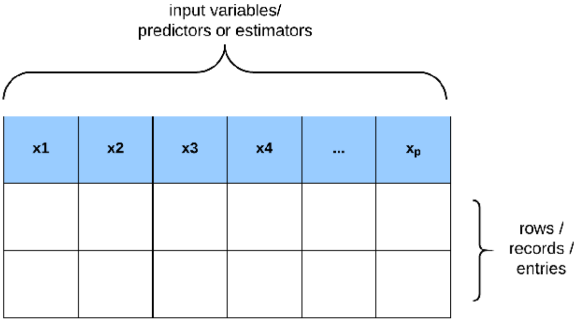


Figure 1: Supervised Learning.

Unsupervised Learning: Overview

- Unsupervised Learning: No corresponding labels - no guidance
- Goal: Computer attempts to determine data's *unknown* structure
- Scheme: By "grouping" similar samples together adaptively

Unsupervised Learning: Overview



x1	x2	x3	x4	...	x_p

Figure 2: Unsupervised Learning.

Reinforcement Learning: Overview

- Reinforcement Learning: Agent interacts with an Environment
- Scheme: A “feedback configuration”
- Method: Chooses an action from the set of actions
- Learning: Based on the responses from the Environment

Reinforcement Learning: Overview

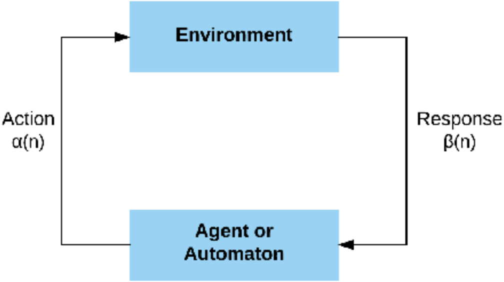


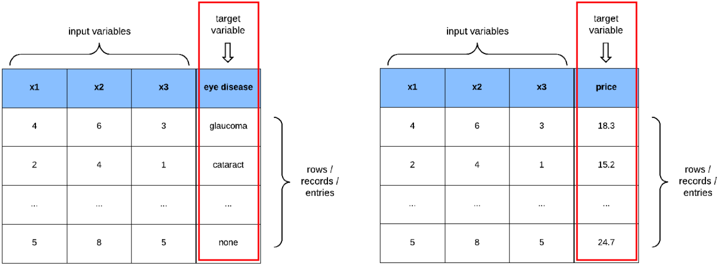
Figure 3: Reinforcement Learning.

Principles of Supervised Learning

Classification vs. Regression

- Supervised Learning: Two problems - Classification and Regression
- Classification: Output or target variable is a category or class
  - Don't Use All Patterns: Prototype Reduction Schemes
  - Use only "Border" Patterns: Border Identification Algorithms
- Regression: Target function is real-valued

Classification vs. Regression



input variables

target variable

x1	x2	x3	eye disease
4	6	3	glaucoma
2	4	1	cataract
...	...	...	...
5	8	5	none

rows / records / entries

input variables

target variable

x1	x2	x3	price
4	6	3	18.3
2	4	1	15.2
...	...	...	...
5	8	5	24.7

rows / records / entries

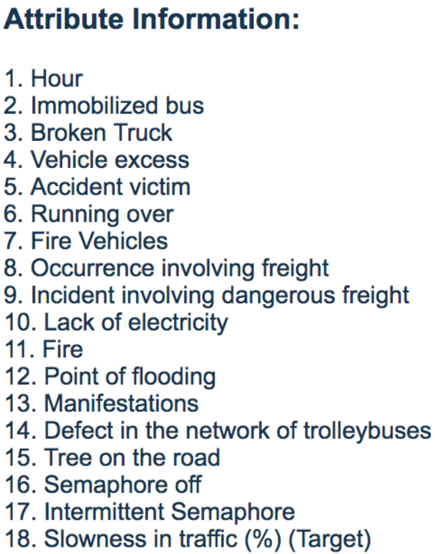
(a) Classification Example.

(b) Regression Example.

Example Dataset: Urban traffic in Sao Paulo

Traffic in Sao Paulo (Brazil)

- Data Set Characteristics: Multivariate, Time-Series
- Number of Instances: 135
- Number of Attributes: 18
- Target: Slowness in traffic (%)



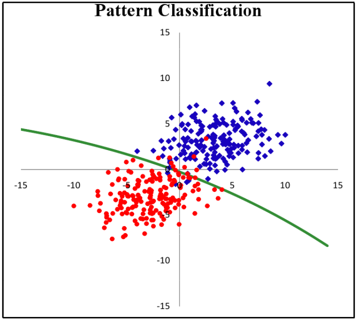
Attribute Information:

1. Hour
2. Immobilized bus
3. Broken Truck
4. Vehicle excess
5. Accident victim
6. Running over
7. Fire Vehicles
8. Occurrence involving freight
9. Incident involving dangerous freight
10. Lack of electricity
11. Fire
12. Point of flooding
13. Manifestations
14. Defect in the network of trolleybuses
15. Tree on the road
16. Semaphore off
17. Intermittent Semaphore
18. Slowness in traffic (%) (Target)

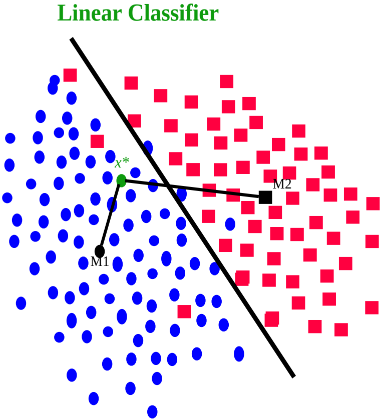
Pattern Classification: What it is...

Categorize feature into classes:*/*; Discriminant Function

- Bayesian classifier Optimal
- NN classifier Later...
- SVM Later...
- Decision Tree Later...
- Parzen window Not visited

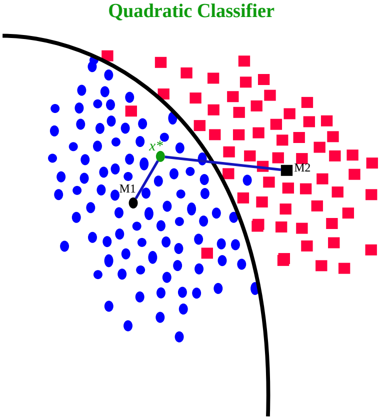


Optimal Classifier for Equal-Variance Distributions



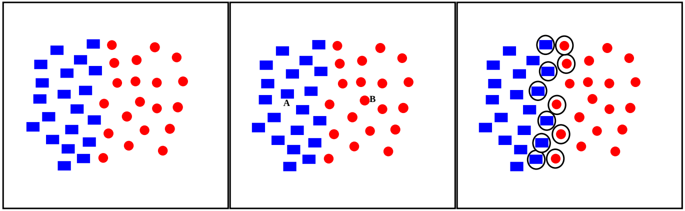
- $\sum_1 = \sum_2$
- Linear Classifier
- Compare with means

Optimal Classifier for Non-equal-Variance Distributions



- $\sum_1 \neq \sum_2$
- Quadratic Discriminant
- Compare with means
- *Mahalanobis* distance

Prototypes and Border Patterns



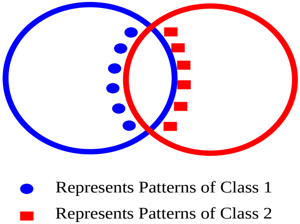
(c) Training Set. **(d)** Prototypes. **(e)** Border Patterns.

Figure 4: Border Patterns vs Prototypes

- Border points that are required in classification!

“Border” Methods

- Border patterns: Close to *neither* the means nor the class boundaries


● Represents Patterns of Class 1
■ Represents Patterns of Class 2

Separated classes

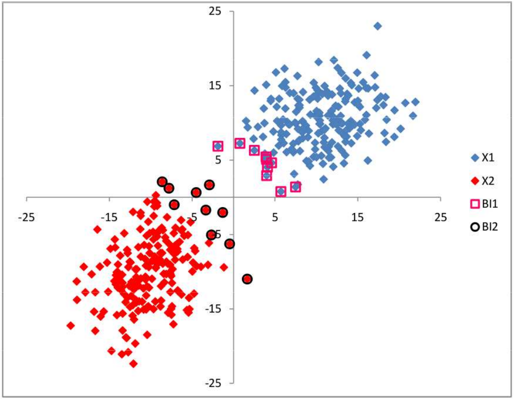


Figure 5: Border Patterns: Fairly separable classes obtained by ABBI.

Accessing when learning has occurred

- Memorization: “Remember” previous examples
- Generalization: Observe prior patterns to resolve unseen patterns

Learning: Use observed datapoints to generalize

Training and Validation data

Goal of ML: Predict/Classify unseen observations.

Partition Data:

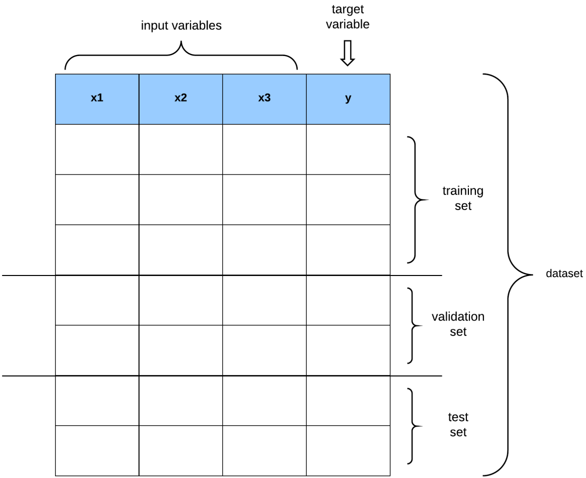
- Training set for training the model
- Validation set for fine-tuning the model parameters
- Test set for assessing the model's performance

Common Strategy:

- Split 60% of the dataset for Training
- 20% for Validation
- 20% for Testing.

This is commonly known as the 60/20/20 rule.

Training and Validation data



input variables

target variable

x1

x2

x3

y

training set

validation set

test set

dataset

Figure 6: Partitioning the Dataset.

The Bias/ Variance Tradeoff

- Bias/Variance tradeoff: Critical for assessing model performance
- Bias: Oversimplifies the learning problem; fails to generalize
- Variance: *Closely* learns the irreducible error of the dataset.
    - Leads to high variability for unseen observations

The Bias/ Variance Tradeoff

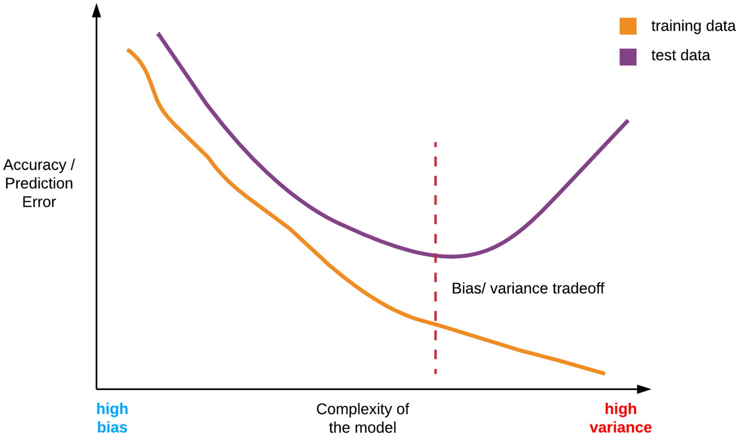


Figure 7: Bias and variance.

The Bias/Variance Tradeoff

- Goal: Have a model that generalizes to new examples
- Finding this middle ground is called the Bias/Variance tradeoff.

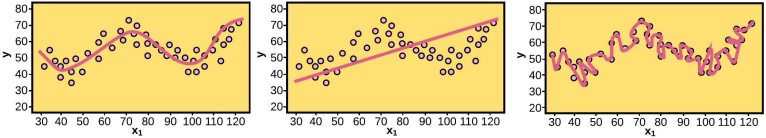


Figure 8: Left: Good fit. Center: Underfit (Bias). Right: Overfit (Variance).

Evaluation Metrics

- Evaluation Metrics: Measure the performance of our ML model

Classification

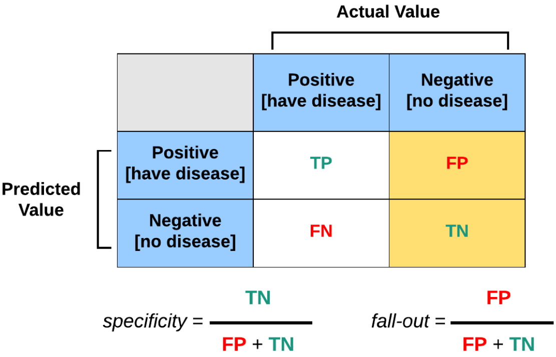
- Confusion matrix.
- AUC-ROC (Area under ROC curve).

Regression

- Root Mean Squared Error (RMSE).
- R-squared (R^2).

Evaluation metrics: Confusion matrix

Confusion matrix: A popular metric for assessing ML performance.



		Actual Value	
		Positive [have disease]	Negative [no disease]
Predicted Value	Positive [have disease]	TP	FP
	Negative [no disease]	FN	TN

$$\textit{specificity} = \frac{\mathbf{TN}}{\mathbf{FP} + \mathbf{TN}} \qquad \textit{fall-out} = \frac{\mathbf{FP}}{\mathbf{FP} + \mathbf{TN}}$$

Figure 9: A Typical Confusion Matrix.

Evaluation metrics: Root Mean Squared Error (RMSE)

- RMSE: Evaluation metric for regression ML problems
- Goal of RMSE: Difference between original and predicted targets

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}.$$

Resampling

- Resampling: Important concept for evaluating the ML performance
- Method: Select a subset of the available dataset
- Train/Test: Train on the data subset and test on the remainder

Methods for resampling include:

- The validation set technique
- The Leave-one-out cross-validation technique (LOOCV)
- The *k*-fold cross-validation technique.

Resampling: Validation Set

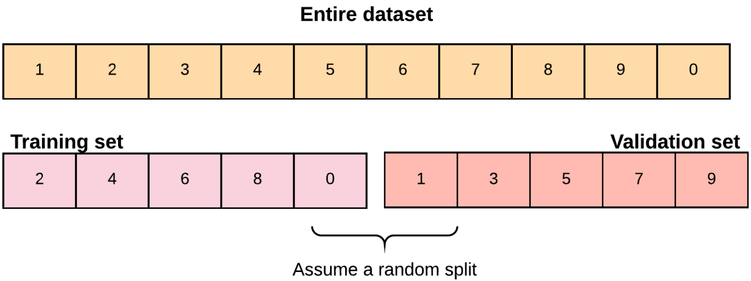
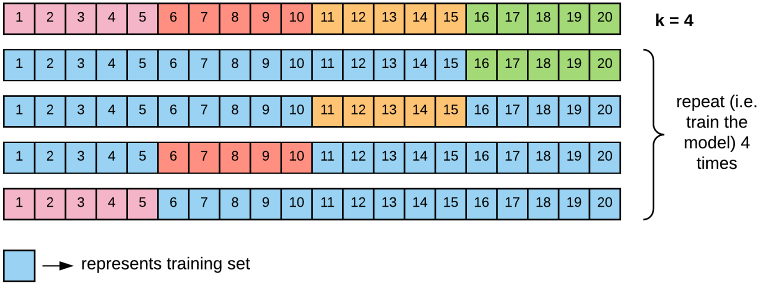


Figure 10: Validation Set.

Resampling: k-Fold Validation



k = 4

repeat (i.e. train the model) 4 times

🟦 → represents training set

Figure 11: *k*-fold Validation.

Supervised Learning Algorithms

Classes of Supervised Algorithms

Supervised ML Algorithms

- Linear
- Non-linear and
- “Ensemble” methods

Linear Algorithms

- Linear Methods: Methods with linear Discriminant in feature space

Examples of linear algorithms are:

- Linear regression
- Fisher's Discriminant
- Logistic regression
- Support Vector Machines (SVMs)

Linear Regression

- Assumption: Output is modeled as a linear combination of inputs
- Output type: Real-valued outputs
- Goal of the linear model: Find “line” to “best” approximate data

Linear Regression

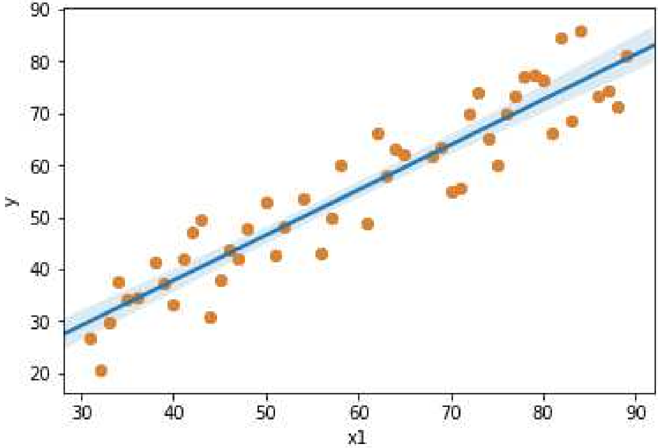


Figure 12: Scatter plot of x1 (x-axis) and y (y-axis) with a Regression Line.

Linear regression: adding non-linearity

- Adapt linear regression: For nonlinear datasets.
- Adding nonlinearity: Leads to *Polynomial Regression*.

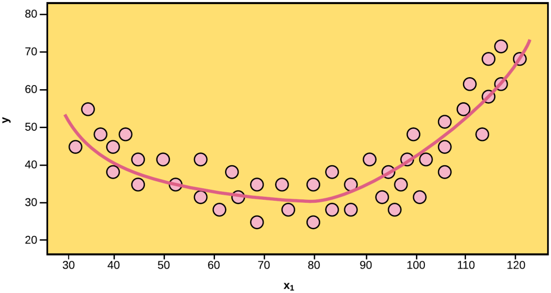


Figure 13: Adding Polynomial Terms to Approximate Dataset.

Logistic Regression

- Logistic regression: Supervised ML algorithm for classification.
- The logistic function: Known as the *logit* or the *sigmoid* function
- Impact: Constrains the output of the cost function between 0 and 1

The Logit or Sigmoid Model

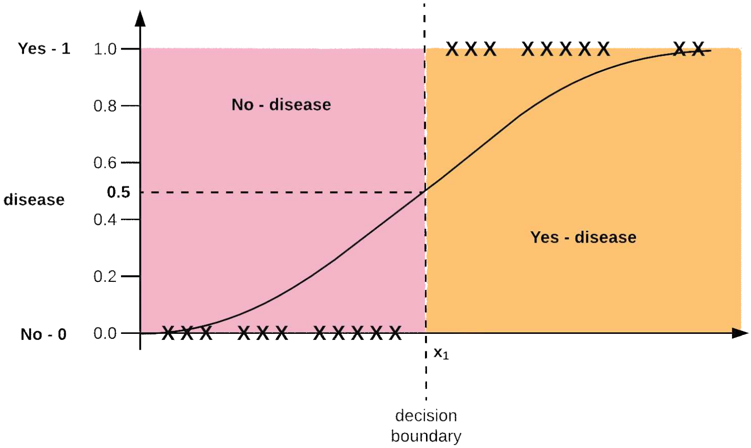


Figure 14: Class Decisions Using Logistic Function

Multinomial Logistic Regression

- Multinomial/Multiclass: Contains more than 2 classes
- The softmax function: The probability of belonging to a class

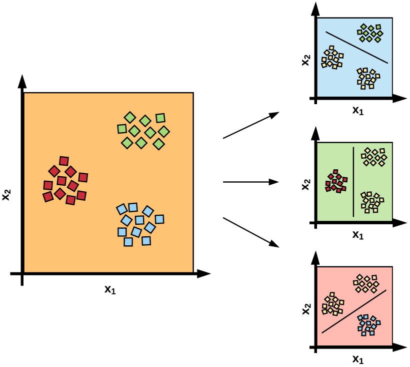


Figure 15: An Illustration of Multinomial Regression

Support Vector Machines (SVM)

- SVM: A Classification ML algorithm
- Linearly Separable data: Infinite set of discriminants
- Goal: Find the "*best*" discriminant

Support Vector Machines

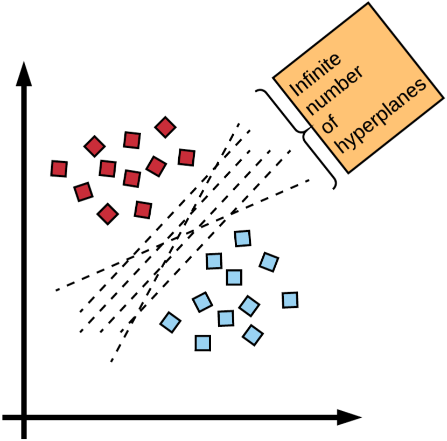


Figure 16: Infinite Set of Linear Discriminants

Hyperplanes

- A hyperplane: Linear separator for 2 classes in *d*-dimensional space
- In 2D: Line; In 3D: Plane
- In $>$ 3D: Hyperplane.

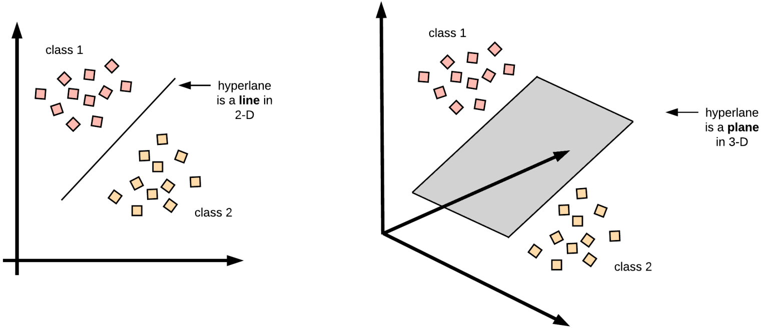


Figure 17: Separator: 2D - *Line*, 3D - *Plane*. If $d > 3$, can't visualize

Determining the Optimal or “Best” Hyperplane

- “Best” Hyperplane: Largest *margin* from the boundary vectors

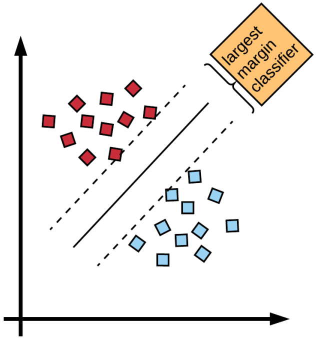


Figure 18: The Largest Margin Classifier

Support Vectors

- Boundary Points: Support Vectors
- Function: Formal algorithm to find largest margin

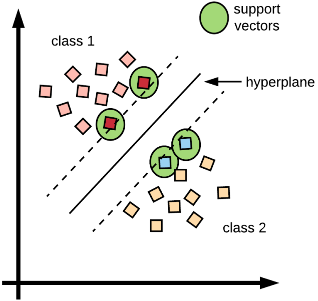


Figure 19: Support Vectors

Non-linear Algorithms

- Non-Linear methods:, non-parametric methods.
- Assumption: does not assume any structural form of the dataset.

Examples of non-linear algorithms are:

- K-Nearest Neighbors,
- Classification & regression trees.
- Support vector machines, and
- Neural networks.

Artificial Neural Networks (ANN)

- ANN: connectionist agents that transfer information.
- Higher representations: formed as data transfers between neurons.
- Deep Learning: forms the foundation for deep learning techniques.

An artificial neural network is composed of:

- An input layer,
- Hidden layer(s), and
- An output layer.

Artificial Neural Networks (ANN)

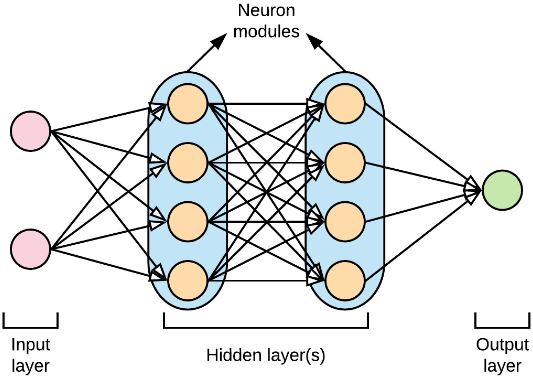


Figure 20: Neural Network Architecture.

Training the Network

- Backpropagation: process of training a neural network.
- How does it train: by adjusting the weights of the network.
- Feedforward algorithm: computes inputs + weights and bias acted upon by an activation.

The Backpropagation Algorithm

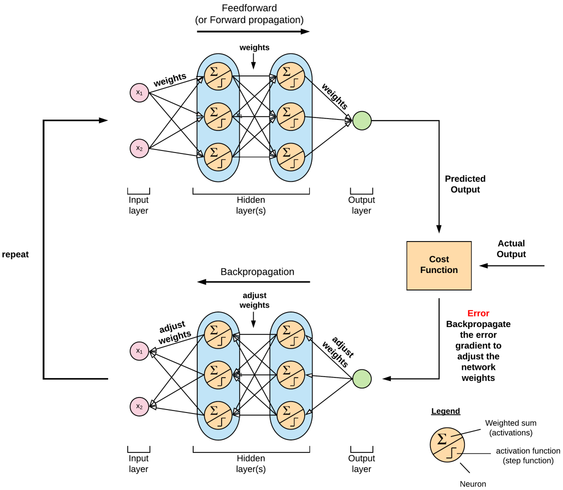


Figure 21: Backpropagation.

Ensemble algorithms

Ensemble methods: combine “weak learners” for better generalization

- Boosting (stochastic gradient boosting), and
- Bagging (random forests).

Decision Trees

- Decision trees: known as classification and regression trees (CART).

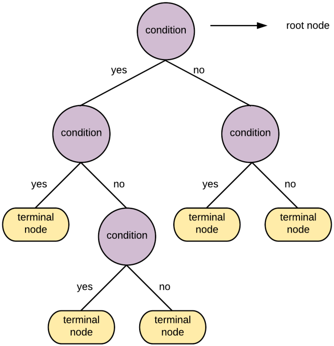


Figure 22: Illustration of a decision tree.

Regression and Classification with CART

- Regression trees: the output is continuous.
- Classification trees: output is categorical.
- Terminal node of regression tree: average of data points in region.
- Terminal node of classification tree: highest occurring class in area.
- Random splits: CART involves randomly splitting the set of attributes into distinct regions.

Regression and Classification with CART

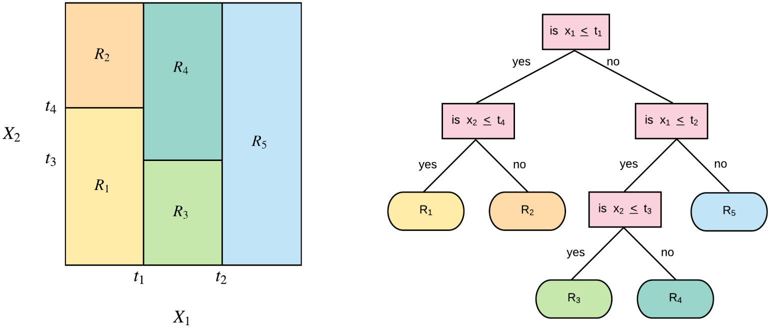


Figure 23: Left: An example of splitting a 2-Dimensional dataset into sub-trees/ regions using the recursive binary splitting technique. Right: The resulting tree from the partitioning on the left.

Random Forests

- Random Forest: an ensemble of decision tree classifiers.
- How: selects random samples from fully grown trees.
- Result: trees are averaged to smoothen out the variance.

Random Forests

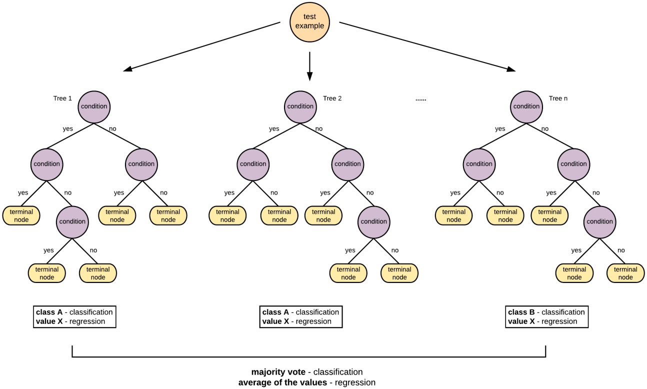
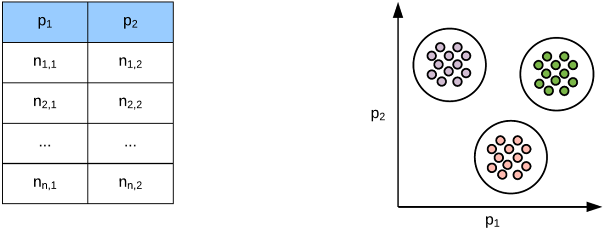


Figure 24: Take a majority vote to determine the final class in the classification case and the average of the values in each trees to determine the predicted value in the regression case.

Unsupervised Learning Algorithms

Clustering

- Clustering: groups homogenous datapoints into partitions.
- Application: to find number of distinct groups among data samples.



p_1	p_2
$n_{1,1}$	$n_{1,2}$
$n_{2,1}$	$n_{2,2}$
...	...
$n_{n,1}$	$n_{n,2}$

Figure 25: An illustration of clustering in a 2-D dimensional space.

Clustering

- Challenge: perform clustering in higher dimensional spaces.
- Algorithms: k-means and hierarchical clustering.
- K-means: used when number of classes are already known.
- Hierarchical: the number of clusters is unknown.

Clustering

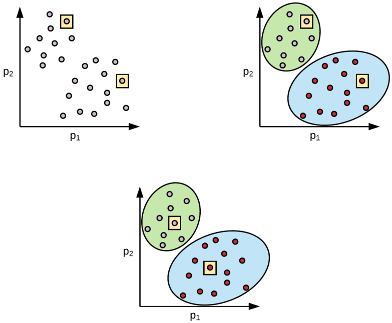
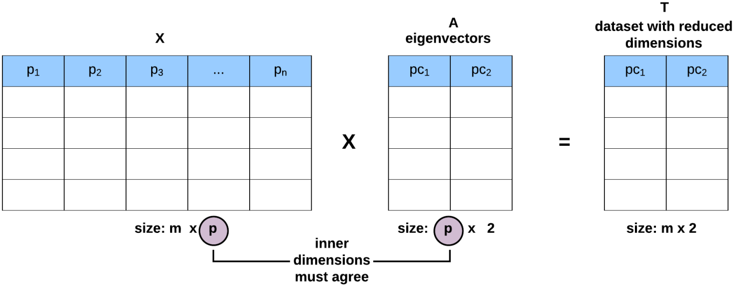


Figure 26: k-Means Clustering with k = 2. Top-Left: Randomly pick a point for each k. Top-Right: Iteratively assign each point to its closest cluster centroid. Bottom: Update the cluster centroids for each of the k clusters. Repeat until the algorithm resolves in a stable clustering.

Principal Component Analysis (PCA)

- Principal components: vectors to capture variability in feature space.
- Goal: to find a low-dimensional feature sub-spaces.
- Use: data visualization.

Principal Component Analysis (PCA)



X

p_1	p_2	p_3	...	p_n

size: m x p

X

A
eigenvectors

pc_1	pc_2

size: p x 2

=

T
dataset with reduced dimensions

pc_1	pc_2

size: m x 2

inner dimensions must agree

Figure 27: Reducing the dimension of the original dataset.

Reinforcement Learning Algorithms

Classes of Reinforcement Learning Algorithms

Reinforcement Learning Algorithms:

- Dynamic programming methods,
- Monte Carlo methods,
- Temporal-difference methods, and
- Learning Automata methods.

Key Concepts of Reinforcement Learning

- agent or automaton: interacts with an Environment.
- environment: the nature of the world.
- state: changes in the Environment over time.
- policy: mapping from states to probability of choosing an action.
- reward: the desirability of a state or action.
- value function: long-term utility of states considering future states.

Dynamic Programming (DP)

- DP: requires model of Environment to compute optimal policies.
- Dynamic programming algorithms:
    - iterative policy evaluation
    - generalized policy iteration
    - value iteration

Monte Carlo Methods

- Monte Carlo: does not assume complete knowledge of Environment.
- Model: They are model-free.
- How do they learn: From experience as *sample episodes*.
- on-policy methods: the agent finds the best policy that still explores.
- off-policy methods: the agent explores, but still learns a deterministic optimal policy.

Temporal-Difference (TD) Methods

- TD: combines ideas from Monte Carlo and dynamic programming.
- Methods for learning: long-term predictions of dynamical systems.

Examples of these methods are:

- Sarsa: *on-policy* control algorithm.
- Actor-critic methods: TD methods with separate memory structures.
- Q-learning: TD *off-policy* control algorithm.
- R-learning: *off-policy* TD method.

Learning Automata Methods

- Learning Automata: methods for solving learning problems.
- Fixed Structure Stochastic Automata (FSSA):
  automata learns optimal policies in random environments.
- Variable Structure Stochastic Automata (VSSA):
  action probabilities updated using a reinforcement scheme.

FSSA Algorithms

Examples of FSSA Algorithms are:

- The Two Action Automaton with Memory, $L_{2N,N}$:
  uses $2N$ states and 2 actions to learn optimal policy.
- The Krinsky Automaton:
  similar to $L_{2N,N}$ when response of the Environment is *unfavourable*.
- The Krylov Automaton:
  similar to $L_{2N,N}$ when the response of the Environment is *favourable*.

VSSA Algorithms

Examples of VSSA Algorithms are:

- The Linear Reward-Penalty (L_{R-P}) Scheme:
  Transition probabilities are updated based on the linear update rule.
- The Linear Reward-Inaction (L_{R-I}) Scheme:
  Updates the action probability on reward; does nothing on penalty.
- The Linear Inaction-Penalty (L_{I-P}) Scheme:
  Updates the action probability on penalty; does nothing on reward.

Applications of Machine Learning

Applications of Machine Learning

Survey of ML application areas.

- Machine Learning for Fraud Detection.
- Machine Learning in Cyber Security.
- Machine Learning in Healthcare.
- Machine Learning in Web Analytics.
- Machine Learning in Product Recommendation.

Machine Learning for Fraud Detection

- Challenge: scarce record of fraudulent transactions.
- Rule-based systems: no longer adequate to tackle credit card fraud.
- Anomaly detection: Learn non-fraudulent patterns.

Machine Learning in Cyber Security

- Cyber attacks: has impacts on the digital, physical and social world.
- Traditional cyber-security tools: are mostly rule-based.
- Threats at scale: cannot combat large-scale coordinated cyber attacks.
- Learning from data: to combat digital and cyber threats.
- Note: attackers also wield the tools of AI and ML.

Machine Learning in Healthcare

- Augment care: by leveraging health records and patient information.
- Relevance: classification of medical imagery to identify cancer.
- Pushing frontiers: genomics and precision medicine.

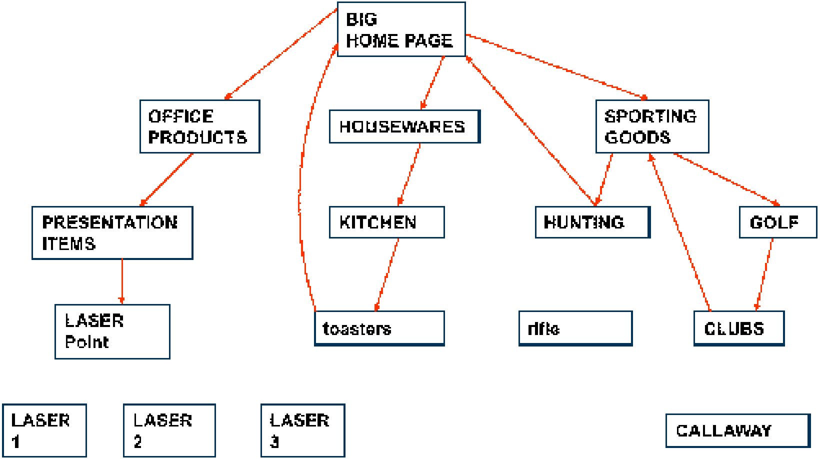
Machine Learning: Web Analytics

Visitor type in an online store:

- casual: website visitors: no intention to buy.
- potential: website visitors: want to buy, maybe not from this site.

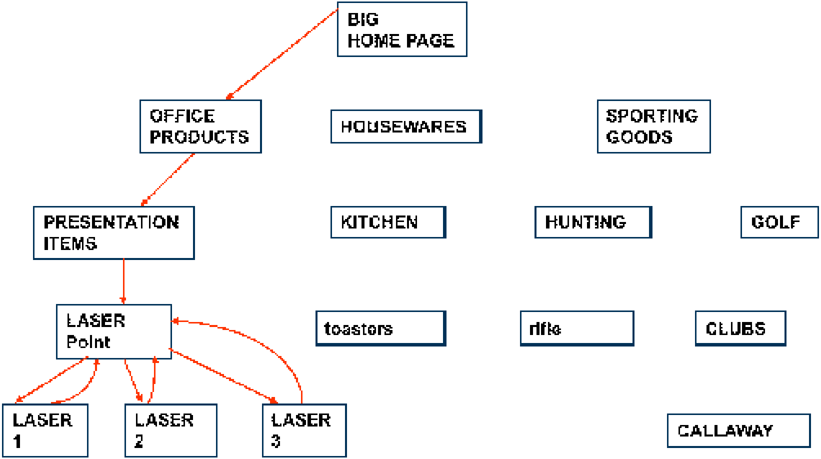
Machine Learning: Web Analytics

Casual Website visitors



Machine Learning: Web Analytics

Potential Website visitors



Machine Learning: Web Analytics

Assessment of “exit” websites:

- Objective: assess the pages where visitors leave the website (exit).
- Goal: improve these “exit” webpages.

Machine Learning: Web Analytics

- Bounce Rate: number of visitors that visit a website and leave right away (they do not go further).

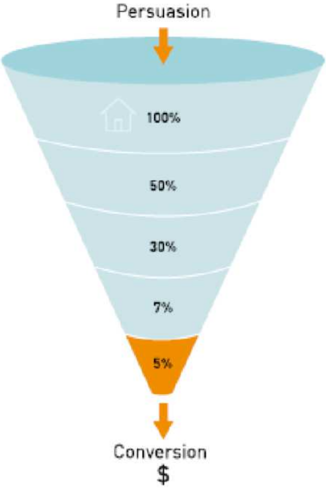
Bounce vs Exit:

- Bounce Page:
    - Page A -¿ Leaving home page
- Exit:
    - Page A -¿ Leaving website
    - Page A -¿ C -¿ Page A -¿ Leaving website
    - Page A -¿ B -¿ again Page A -¿ Leaving website

Machine Learning: Web Analytics

Conversion Rate: the percentage of visitors who complete the funnel

- Purchases.
- Social shares.
- Email Signup.
- Contacts via form & phone.
- File downloads.
- Video views.



Machine Learning: Web Analytics

Type of Visitors:

- Unique visitors: number visiting the site multiple times per period.
- New visitors: visits the web site for the first time per period.
- repeat visitor: unique visitors with two or more visits per period.
- return visitor: unique visitor who is not a new visitor.

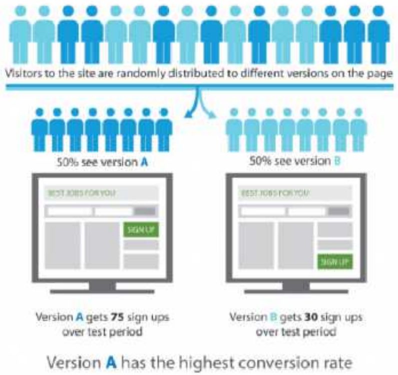
Machine Learning: Web Analytics

A/B testing:

- A change at a time.
- Run both versions together.
- Pick a winner and start again.

Machine Learning: Web Analytics

Randomized experiments:



Visitors to the site are randomly distributed to different versions on the page
50% see version A
50% see version B
BEST JOBS FOR YOU
SIGN UP
BEST JOBS FOR YOU
SIGN UP
Version A gets 75 sign ups over test period
Version B gets 30 sign ups over test period
Version A has the highest conversion rate

Machine Learning: Product Recommendation

Recommendation systems: allow users to get information especially tailored to:

- user requirements.
- goal.
- knowledge.
- interests.

Main methods:

- content-based: recommend similar to objects previously liked.
- collaborative filtering: based on similarities between users and objects.

Machine Learning: Product Recommendation

Explicit feedback vs. implicit:

- explicit: rating, “visit a website”, “add to favourite” etc.
- implicit: visit duration of a webpage.

Machine Learning: Product Recommendation

Some existing systems:

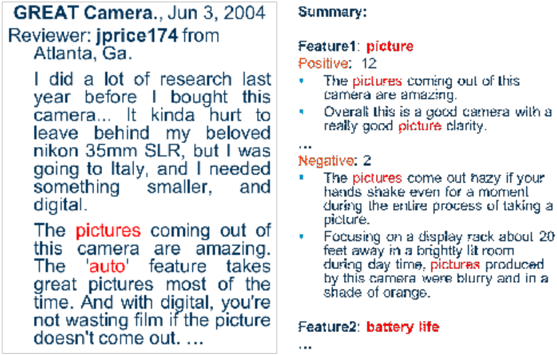
Machine Learning: Sentiment Analysis

Feature Based Analysis & Summarization:

- extracting product features.
- identifying opinion sentences and deciding sentiment (positive or negative).
- summarizing and comparing results.

Machine Learning: Sentiment Analysis

An example:



GREAT Camera., Jun 3, 2004

Reviewer: **jprice174** from Atlanta, Ga.

I did a lot of research last year before I bought this camera... It kinda hurt to leave behind my beloved nikon 35mm SLR, but I was going to Italy, and I needed something smaller, and digital.

The pictures coming out of this camera are amazing. The 'auto' feature takes great pictures most of the time. And with digital, you're not wasting film if the picture doesn't come out. ...

Summary:

Feature1: **picture**

Positive: 12

- The pictures coming out of this camera are amazing.
- Overall this is a good camera with a really good picture clarity.

...

Negative: 2

- The pictures come out hazy if your hands shake even for a moment during the entire process of taking a picture.
- Focusing on a display rack about 20 feet away in a brightly lit room during day time, pictures produced by this camera were blurry and in a shade of orange.

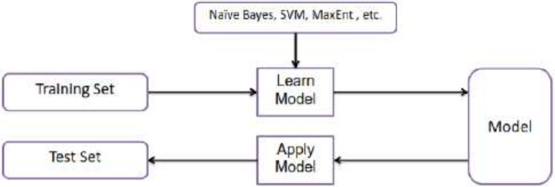
Feature2: **battery life**

...

....

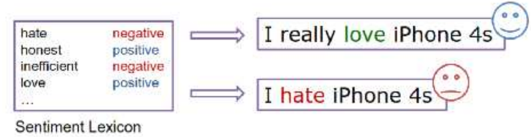
Machine Learning: Sentiment Analysis

Machine Learning Based Approach:



Machine Learning: Sentiment Analysis

Sentiment Lexicon Based Approach: Building a better dictionary.



Machine Learning: Sentiment Analysis

Sentiment lexicon: assigning score to words in text.

Lexicon Generation:

- Manual generation: tedious and time consuming.
- Corpus-based family.
- Dictionary-based family.

Conclusion

Summary

- Bibliography: [1, 2, 3, 4]

Questions?

References i

- Artificial intelligence for a smarter kind of cybersecurity, ibm.
  https://www.ibm.com/security/artificial-intelligence.
  Accessed: 2019-01-24.

- E. O. Bisong.
  Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners.
  Apress (Springer Nature), New York, NY, (In Review) 2019.

- K. S. Narendra and M. A. L. Thathachar.
  Learning Automata: An Introduction.
  Dover Publications, Mineola, NY, 2012.

References ii

R. S. Sutton and A. G. Barto.
Reinforcement Learning: An Introduction.
MIT Press, Cambridge, MA, 2011.