Foundations of Statistical and Machine Learning for Actuaries -

Welcome and Foundations

Edward (Jed) Frees, University of Wisconsin - Madison
Andrés Villegas Ramirez, University of New South Wales

July 2025

Schedule

Day and Time	Presenter	Topics	Notebooks for Participant Activity
Monday Morning	Jed	Welcome and Foundations Hello to Google Colab	Auto Liability Claims
	Jed	Classical Regression Modeling	Medical Expenditures (MEPS)
Monday Afternoon	Andrés	Regularization, Resampling, Cross-Validation	Seattle House Sales
	Andrés	Classification	Victoria road crash data
Tuesday Morning	Andrés	Trees, Boosting, Bagging	
	Jed	Big Data, Dimension Reduction and Non-Supervised Learning	Big Data, Dimension Reduction, and Non-Supervised Learning
Tuesday Afternoon	Jed	Neural Networks	Seattle House Prices Claim Counts
	Jed	Graphic Data Neural Networks	MNIST Digits Data
Tuesday 4 pm	Fei	Fei Huang Thoughts on Ethics	
Wednesday Morning	Jed	Recurrent Neural Networks, Text Data	Insurer Stock Returns
	Jed	Artificial Intelligence, Natural Language Processing, and ChatGPT	
Wednesday After Lunch	Dani	Dani Bauer Insights	
Wednesday Afternoon	Andrés	Applications and Wrap-Up	

Monday Morning IA - Welcome and Foundations

The Concern: *Robots took the jobs of factory workers. Will artificial intelligence take the jobs of Actuaries?*

This module covers:

- the bases of statistical and machine learning,
- topics covered in this course and how we will approach them, and
- the role of Google Colaboratory (colab).

Welcome to the course

Presenters

- Edward (Jed) Frees
- Andrés Villegas Ramirez
- Fei Huang
- Dani Bauer

What is Machine Learning, Deep Learning, and Artificial Intelligence?

All three can be described as the effort to automate intellectual tasks normally performed by humans.

- **Machine learning** is the science (and art) of programming computers so they can learn from data.
- [Machine learning is the] field of study that gives computers the ability to learn without being explicitly programmed. - Arthur Samuel, 1959

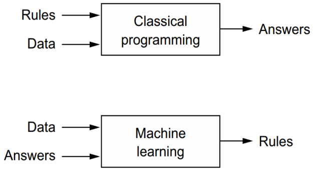


Figure 1.2 Machine learning: a new programming paradigm

{*Credit*: Chollet, F. (2021). Deep Learning with Python}

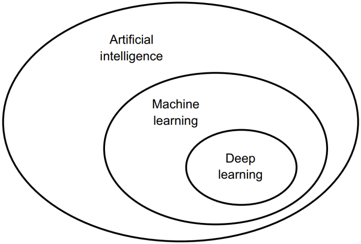


Figure 1.1 Artificial intelligence, machine learning, and deep learning

- **Deep learning** is a specific subfield of machine learning: a new take on learning representations from data that puts an emphasis on learning successive layers of increasingly meaningful representations.
  - The "deep" in "deep learning" stands for this idea of successive layers of representations.
  - The number of layers is called the *depth* of the model.

{*Credit*: Chollet, F. (2021). Deep Learning with Python}

- Experts believed that human-level **artificial intelligence** could be achieved by having programmers handcraft a sufficiently large set of explicit rules for manipulating knowledge stored in explicit databases.
  - This approach is known as symbolic AI.
  - It was the dominant paradigm in AI from the 1950s to the late 1980s, and it reached its peak popularity during the “expert systems” boom of the 1980s.
- Some of you may recall the “model office” approach used in actuarial science, particularly with life insurance cash flows, in the 60’s and 70’s. They included:
  - Projections of Capital and surplus, Production Distribution of business (new and in force), Investment return, Gross premium rates, Premium taxes, Commissions and other field expenses, General administrative expenses, Mortality rate, Morbidity rate, Persistency rate, Cash-value basis, Reserve basis, Policyholder dividends, Premium-mode distribution, Reinsurance cost, and Federal income tax.
  - Very complicated!

Quantitative Foundations of Insurance

A bit of history to emphasize that the importance of quantiative methods - these are the root of our discipline.

- We have long used data for insurance pricing.
  - An early basis is provided by John Graunt's Observations on Bills of Mortality (1662), reprinted in Graunt (1939).
  - A more complete analysis of births and deaths by age was subsequently conducted by Edmund Halley in his Breslau Life Tables (1693), reprinted in Halley (1977).
- Theoretical milestones are the basis of a disciplined approach
  - Lundberg's 1909 treatise on the collective theory of risk provided an important impetus to the development of stochastic processes.
  - In the realm of statistics, a lead figure is Harald Cramér, an insurance analyst, actuary, and statistician, who developed some of the theoretical cornerstones of actuarial science, e.g., Cramér (1930)
  - Also notable is the book by Bühlmann (1970) that introduced the stochastic model approach toward non-life insurance.

Non-Life Insurance

- Includes insurance products that differs from life insurance
  - “Non-life” is mainly used in Europe, “general” is used in the United Kingdom, and “property and casualty” is used in the US and Canada
- Non-life insurance pricing
  - Probably the first use of GLM model in non-life insurance was the paper “Two Studies in Automobile Insurance Ratemaking” by Bailey and LeRoy (1960)
  - This paper addressed the problem of classification ratemaking.
  - They introduced a minimum bias model that was later discovered (Mildenhall, 1999) to be equivalent to GLM modeling.
- Currently, actuarial pricing in car insurance is often based on 40 to 50 covariates that discriminate among policyholders.

- **Challenges**:
  - The majority of explanatory variables are of categorical type, a statistical analysis faces complications such as, e.g., the sparsity of the underlying design matrix.
  - Covariates interact in a nontrivial way, making proper estimation a challenging task.
  - Claims frequency modeling is a rare event prediction problem where actuaries try to find systematic effects in data that are heavily dominated by the noisy (random) part.
- Claims Reserving in Non-Life Insurance
  - Claims reserving in non-life insurance is concerned with predicting claims cash flows that can last over multiple years
  - Stochastic models were introduced by Mack (1993) and Renshaw & Verrall (1998)
  - The stochastic basis for this individual claims reserving view was introduced in the 1990s by Arjas (1989) and Norberg (1993, 1999)

Life Insurance

- Life and pension insurance insures life and protects against disability and death of individual policyholders over possibly their entire lifetimes.
- As a consequence, life insurance is very much concerned with predicting mortality and longevity trends over several decades into the future.
- Clearly, the more statistical part of life and pension insurance modeling is mortality forecasting.
    - The most popular stochastic mortality projection models are the Lee & Carter (1992) model and the Cairns et al. (2006) model.
    - Because life and pension insurance are of a long-term nature, product design is especially important to guaranteeing long-term financial stability.

Reinsurance

- The reinsurance industry provides specific services and products so that the primary (direct) insurers, referred to as ceding companies, can cap or reengineer their insurance product portfolios.
- The reinsurance industry, especially, faces serious challenges emerging from climate change due to its perceived impact on climatological catastrophes and related insurance covers.

Sources

{

- Analytics of Insurance Markets, Frees, 2015
- Recent Challenges in Actuarial Science, 2021, by Paul Embrechts and Mario V. Wüthrich

}

Statistical and Machine Learning

Statistical Learning and Data Modeling

- One way to motivate an algorithmic development is through the use of a **data model**.
- With a “probability” or “likelihood” based model, our main goal is to understand the target (Y) distribution, typically in terms of the explanatory variables.
- Classical data models are particularly useful for:
  - the goal of explanation
  - understand the uncertainty of our predictions
  - interpretability.

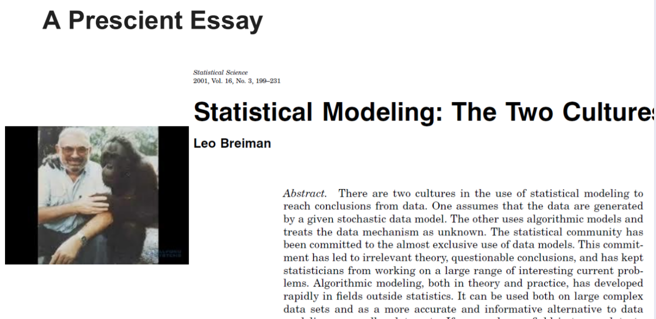
Machine Learning and Algorithmic Modeling

- Idea underpinning **algorithmic modeling**
  - One variable, Y, is determined to be a target variable.
  - Other variables, $X_1, X_2, \ldots, X_p$, are used to understand or explain the target Y.
  - Goal - determine an appropriate function $f(\cdot)$ so that $f(X_1, X_2, \ldots, X_k)$ is a useful predictor of Y.
- Classic Special Case: Linear Regression with functions

$$f(x_{i1}, \ldots, x_{ik}) = \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = \mathbf{x}_i' \boldsymbol{\beta}.$$

Algorithmic Modeling Culture

- Emphasizes algorithmic fitting particularly in complex problems such as voice, image, and handwriting recognition.
- Algorithmic methods are especially useful when the goal is prediction.
- Does not emphasize the distribution of outcomes
- In addition to linear regression, algorithmic fitting methods include ridge and lasso regression, as well as regularization methods.



A Prescient Essay

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Culture

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data

Algorithmic Modeling Scope of Applications

With the algorithmic approach, we can do so much more than with the traditional data modeling approach. In particular:

- **Big Data**. Examples of big data include text documents, videos, and audio files that are also known as *unstructured* data.

Data Sources	**Algorithms**
Mobile devices	Statistical learning
Auto telematics	Artificial intelligence
Home sensors (Internet of Things)	Structural models
Drones, micro satellites	
Data	**Software**
Big data (text, speech, image, video)	Text analysis, semantics
Behavioral data (including social media)	Voice recognition
Credit, trading, financial data	Image recognition
	Video recognition
Source : Stephen Mildenhall, Personal Communication	

Scope of Learning - Supervised and Unsupervised Learning

With many variables, we have the opportunity to think about some of them as “inputs” and others “outputs” of a system.

- Models based on input and output variables are known as **supervised learning methods**.
    - When the target variable is a continuous variable, supervised learning methods are called **regression methods**.
    - When the target variable is a categorical variable, supervised learning methods are called **classification methods**.
- **Unsupervised learning methods** - data are treated the same and there is no artificial divide between “inputs” and “outputs.”

Some Variations

- Naturally, there are many variants of these two basic themes:
    - **Semi-supervised learning**. Since labeling data is usually time-consuming and costly, you will often have plenty of unlabeled instances, and few labeled instances. Some algorithms can deal with data that's partially labeled. This is called semi-supervised learning. Most semi-supervised learning algorithms are combinations of unsupervised and supervised algorithms.
    - **Reinforcement learning.** Reinforcement Learning is a branch of machine learning focused on making decisions to maximize cumulative rewards in a given situation. It interacts with the data!

Some Terminology

- Regression has traditionally been applied in many different fields
- Computer science machine learning has their own set of terms

Target Variable	**Features**
$\mathbf{Y}$	$\mathbf{X}'s$, or Explanatory Variables
Dependent variable	Independent variable
Response	Treatment
Output	Input
Endogenous variable	Exogenous variable
Predicted variable	Predictor variable
Regressand	Regressor

Algorithmic Modeling Fitting Methods

Many of these algorithms take an approach similar to linear regression.

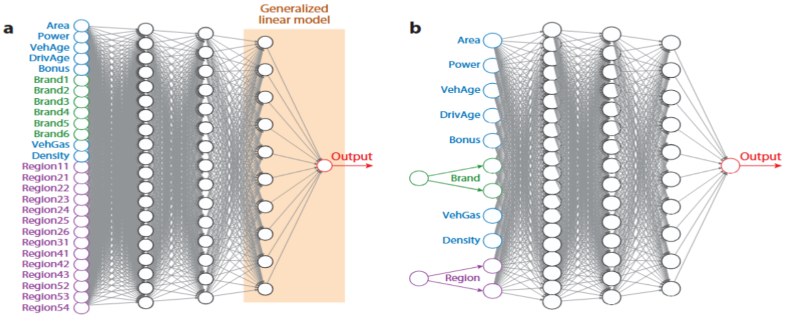


Figure 1

(*a*) Feed-forward neural network of Expression 4 of depth $d = 3$. (*b*) Feed-forward neural network using embedding layers of dimension 2 for categorical covariates (*green and purple*).

{ *Credits:* Recent Challenges in Actuarial Science, 2021, by Paul Embrechts and Mario V. Wüthrich }

Topics Addressed in this Course

- The Foundations
- Statistical learning tools
  - Logistic Regression and Generalized Linear Models
  - Regularization, Resampling, Cross-Validation
  - Classification
  - Trees, Boosting, Bagging
- Machine learning tools
  - Big Data, Dimension Reduction and Non-Supervised Learning
  - Neural Networks
  - Graphic Data Neural Networks
  - Recurrent Neural Networks, Text Data
  - Artificial Intelligence, Natural Language Processing, and ChatGPT

Topics Addressed in this Course 2

- Guest Lecturers:
  - Fei Huang
  - Dani Bauer
- We will not cover **Data**. See the online Chapter Two of Loss Data Analytics, Edition Two for a discussion of data considerations in terms of:
  - data types,
  - data structure and storage,
  - data cleaning,
  - big data issues, and
  - ethical issues.

Course Learning Approach, with Google Colab

We combine lecture with hands-on learning in the form of "labs."

- Jupyter notebooks provide a handy way to combine executable code, code outputs, and text into one connected file.
    - You can take a look at the course notebooks by going to nbviewer site. Then, enter the course Github repo URL https://github.com/OpenActTextDev/ActuarialRegression, select the folder and then a notebook that you want to view.
    - To interact with notebook, go to Colab!
- Google colaboratory (colab for short) is a cloud-based system of servers designed to process machine learning code.
    - We will use the free base system - you only need a (free) Google account.
    - Colab handles both R and python code - perfect for our needs.
    - Machine learning applications often depend upon large datasets and utilize computationally intensive algorithms - Colab is designed to accommodate these demands.

Session IA - Welcome and Foundations - Summary

This module covered:

- The bases of statistical and machine learning,
  - What is statistical and machine learning
  - Quantative foundations of insurance
- Topics covered in this course and how we will approach them, and
- The role of Google Colaboratory (colab).
- During lab, participants may follow the notebook Auto Liability claims