MAT555E: Statistical Data Analysis for Computational Sciences

Fall22-Lecture 01: Introduction

Gül İnan

İstanbul Technical University

Self Introduction

- Education
  - B.Sc., M.Sc., and Ph.D. in Statistics from Middle East Technical University, Ankara, Turkey.
  - Post-doctoral research experience on high-dimensional data analysis at School of Statistics, University of Minnesota, MN, USA.
- Current research interests
  - High-dimensional data analysis problems,
  - Data splitting problems,
  - Computational statistics, and
  - With applications in health care and industrial data.

Course Information

- Course Website
  - https://mat555e-fall22.github.io/ (for syllabus).
- Course GitHub organization
  - https://github.com/MAT555E-Fall22 (for lecture materials and Jupyter Notebooks).
- Course on Ninova (for announcements and grade list)
  - https://ninova.itu.edu.tr/Ders/27718/Sinif/81564.
- GitHub Classroom (for collecting exam papers and paper presentations/reports)

What is statistical learning?

What is statistical learning?

With advances in computer and information technology, **vast** amounts of **complex** data being **generated**, **collected**, and **stored** in many fields such as finance, genetics, and industry etc.

What is statistical learning?

- More specifically,
  - **Advances in data collection technologies** (high-throughput chips in experimental sciences, internet for social sciences, digitized documents in humanities, high-volume transactions in finance, medical imaging in medicine, sensors, drones, satellite images),
  - Rapidly increasing **computing power**,
  - **Faster communication** between devices with **faster connection speeds**,
  - Better **database management systems**, and
  - **Cheaper data storage devices** with unlimited capacities, such as cloud storage,
- supports the explosive **growth of data** (in different types) in a great variety of fields.

What is statistical learning?

- Statistical learning methods are tools used to **discover the hidden patterns and relationships** in **complex** and most probably **large data** sets.
- Unlike the conventional statistical methods, the main objective of statistical learning methods is **making predictions**.
- Some statistical learning methods may be very **computationally intensive**.

Who uses predictive modelling and statistical learning?

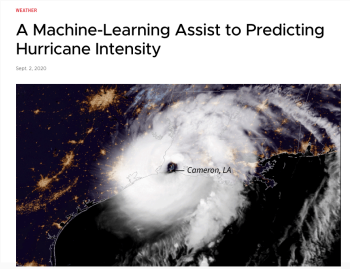
- **Academia** (**Scientists**)

Who uses predictive modelling and statistical learning?

- **A team of medical doctors** trained a machine learning algorithm on brains scans of more than 400 patients with early and later stage Alzheimer's patients with other neurological conditions, and healthy controls.
- The developed **algorithm** could accurately **predict** whether someone had Alzheimer's disease or not in **98 percent of cases**.

Who uses predictive modelling and statistical learning?

- **A team of atmospheric researchers** trained a machine learning algorithm on storms from 1998 to 2008 and tested it using a different set of storms, from 2009 to 2014.
- For hurricanes whose winds increased by at least 35 mph (mile per hour) within 24 hours, the researchers' model had a **60% higher probability of detecting** the rapid-intensification event compared to the current operational forecast model.



Who uses predictive modelling and statistical learning?

- Industry

Who uses predictive modelling and statistical learning?

Who uses predictive modelling and statistical learning?

- Our **media-focused ML research**, development, and opportunities related to the following areas:
  - Computer vision: video understanding search and match cut tools,
  - VFX and Computer graphics: matting/rotoscopy, volumetric capture to digitize, actors/props/sets, animation, and relighting,
  - Audio and Speech, and
  - Content: understanding, extraction, and knowledge graphs.

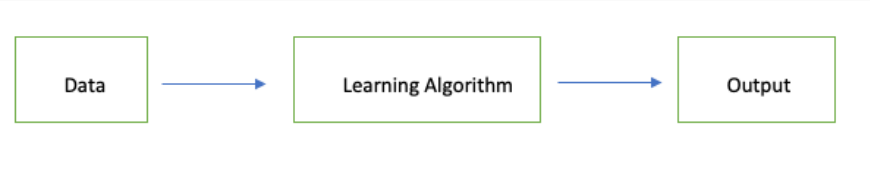
Who uses predictive modelling and statistical learning?

- The **Video Quality Analysis (VQA) group** in Amazon Prime Video trains **computer vision** models to watch video and spot issues that may compromise the customer viewing experience, such as blocky frames, unexpected black frames, and audio noise.

How is statistical learning carried out?

What is the statistical learning pipeline?

- Most people think that...



What is the statistical learning pipeline?

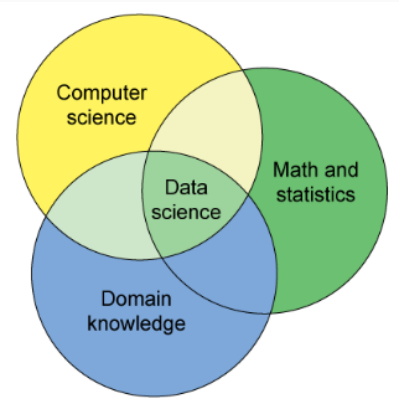
- Oftentimes, we hope statistical learning to go beyond computer simulation and to tackle real world problems.

The complete statistical learning life cycle

1. Quantitative formulation of the problem and data collection,
2. Data cleaning,
3. Learning algorithm,
4. Model assessment and validation, and
5. Domain knowledge used to decide the relevance of the output.

Data Science at the intersection of three pillars

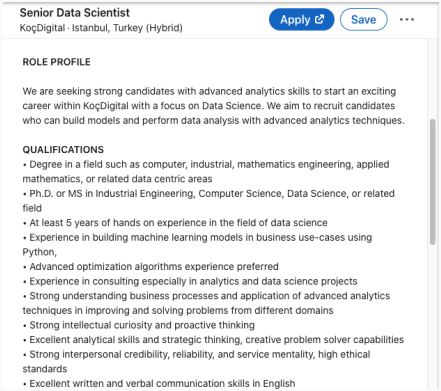
- Many people believe data science to be at the intersection three pillars: **statistics**/**mathematics**, **computer science**, and **domain knowledge** (business, operation research, medicine, music, architecture, history etc).



Field expectations

- Understand domain problem and put into a statistical prediction problem with an appropriate method.
- Explain the relevance of the method chosen.
- Understand the pros and cons of the method.
- Know how to implement chosen model through a machine learning software/library.
- Know how to perform data exploration (including EDA and visualization).
- Communicate results.

Recent job posts



Senior Data Scientist
KoçDigital · Istanbul, Turkey (Hybrid)

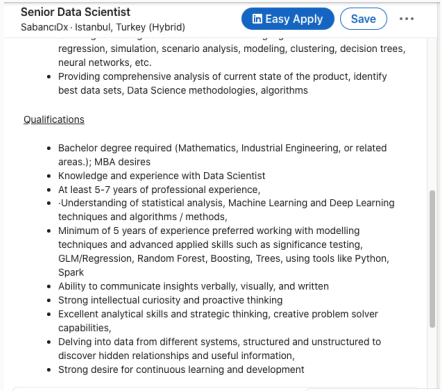
Apply | Save | ...

ROLE PROFILE

We are seeking strong candidates with advanced analytics skills to start an exciting career within KoçDigital with a focus on Data Science. We aim to recruit candidates who can build models and perform data analysis with advanced analytics techniques.

QUALIFICATIONS

• Degree in a field such as computer, industrial, mathematics engineering, applied mathematics, or related data centric areas
• Ph.D. or MS in Industrial Engineering, Computer Science, Data Science, or related field
• At least 5 years of hands on experience in the field of data science
• Experience in building machine learning models in business use-cases using Python,
• Advanced optimization algorithms experience preferred
• Experience in consulting especially in analytics and data science projects
• Strong understanding business processes and application of advanced analytics techniques in improving and solving problems from different domains
• Strong intellectual curiosity and proactive thinking
• Excellent analytical skills and strategic thinking, creative problem solver capabilities
• Strong interpersonal credibility, reliability, and service mentality, high ethical standards
• Excellent written and verbal communication skills in English



Senior Data Scientist
SabancıDx · Istanbul, Turkey (Hybrid)

Easy Apply | Save | ...

regression, simulation, scenario analysis, modeling, clustering, decision trees, neural networks, etc.
- Providing comprehensive analysis of current state of the product, identify best data sets, Data Science methodologies, algorithms

Qualifications

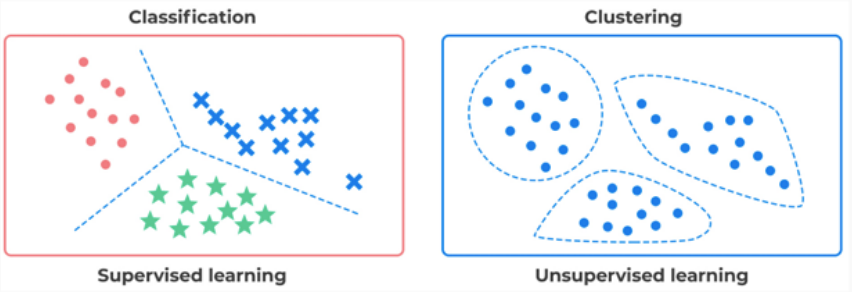
- Bachelor degree required (Mathematics, Industrial Engineering, or related areas.); MBA desires
- Knowledge and experience with Data Scientist
- At least 5-7 years of professional experience,
- ·Understanding of statistical analysis, Machine Learning and Deep Learning techniques and algorithms / methods,
- Minimum of 5 years of experience preferred working with modelling techniques and advanced applied skills such as significance testing, GLM/Regression, Random Forest, Boosting, Trees, using tools like Python, Spark
- Ability to communicate insights verbally, visually, and written
- Strong intellectual curiosity and proactive thinking
- Excellent analytical skills and strategic thinking, creative problem solver capabilities,
- Delving into data from different systems, structured and unstructured to discover hidden relationships and useful information,
- Strong desire for continuous learning and development

Categories of statistical learning problems

Categories of statistical learning problems

1. Supervised learning
2. Unsupervised learning

Categories of statistical learning problems



Supervised learning set-up

- Each data point consists of two parts:
  - Outcome measurement: Y (also dependent variable, response, target).
  - p-dimensional input measurement: X (also inputs, regressors, covariates, features, independent variables).
- We have N training data points $(y_1, x_1), \ldots, (y_n, x_N)$.
  - **Regression:** Y is quantitative (e.g., price, blood pressure).
  - **Classification:** Y takes values in a finite, un-ordered set (survived/died, digit 0-9, cancer stage of tissue sample).

Objectives of supervised learning

- With the help of training data, we would like to:
  - Accurately predict unseen test cases,
  - Understand which inputs affect the outcome, and how,
  - Assess the quality of our predictions and inferences.

Unsupervised learning set-up

- No outcome variable!
- p-dimensional input measurement X.
- We have N training data points $x_1, \ldots, x_N$.

Objectives of unsupervised learning

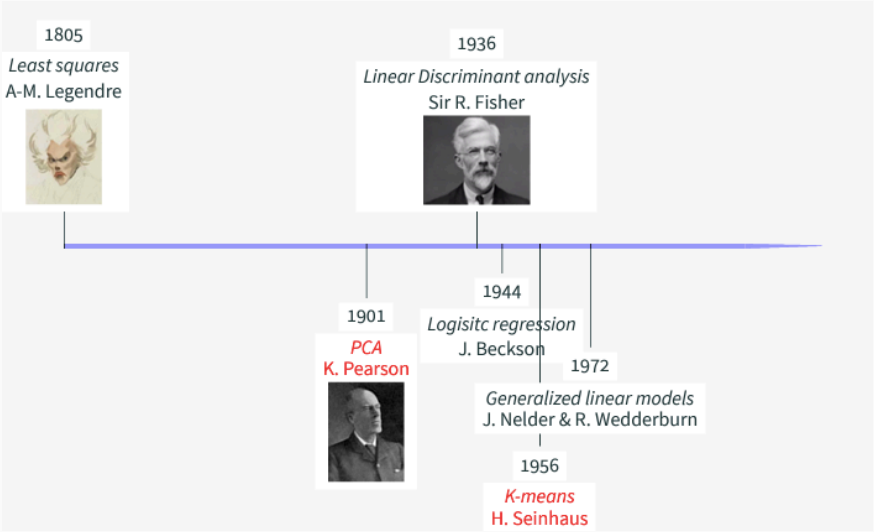
- Objects are fuzzy
  - Find underlying simple structure.
  - Find groups of samples that behave similarly.
  - Find features that behave similarly.
  - Find linear combinations of features with the most variation.
- Useful for exploratory data analysis and get ideas of building supervised learning models.

A brief history of statistical learning

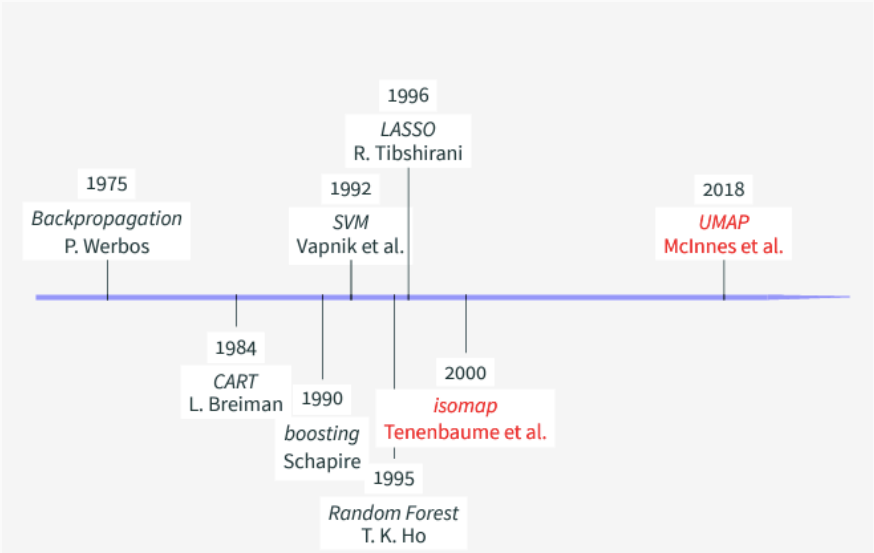
A brief history of statistical learning

- Statistical learning only emerged as a new important sub-field in statistics in the 90's.
- Though **the term statistical learning is new**, many of the concepts were developed long ago.

Timeline


1805
Least squares
A-M. Legendre
1936
Linear Discriminant analysis
Sir R. Fisher
1901
PCA
K. Pearson
1944
Logisitc regression
J. Beckson
1972
Generalized linear models
J. Nelder & R. Wedderburn
1956
K-means
H. Seinhaus

Timeline


1996
LASSO
R. Tibshirani
1975
Backpropagation
P. Werbos
1992
SVM
Vapnik et al.
2018
UMAP
McInnes et al.
1984
CART
L. Breiman
1990
boosting
Schapire
2000
isomap
Tenenbaume et al.
1995
Random Forest
T. K. Ho

Course Topics

- Formulation of a statistical prediction problem.
- Exploratory data analysis.
- Simple linear regression, multiple linear regression.
- Shrinkage methods and regularization.
- Model assessment and bias-variance trade-off.
- Logistic regression and classification.
- Linear discriminant analysis. Quadratic discriminant analysis.
- Tree based methods.
- Support vector machines
- Principal component analysis. Factor analysis.
- Clustering methods.

References

- https://www2.stat.duke.edu/courses/Fall21/sta521.001/post/week01-1/ (with permission of Prof. Yuansi Chen)
- https://online.stat.psu.edu/stat508/lesson/1a

Announcement: Prospective Graduate Students

- I am planning to supervise 1-2 graduate students for **master thesis projects** starting from Spring23/Summer23.
- The problem/project in my mind suits best for students with the following background:
  - Industrial engineering degree (major or minor degree),
  - Experience in mathematical statistics, mathematical optimization, and data analysis,
  - Good knowledge of programming in Python (able to write code from scratch),
  - Research oriented and hard-working, and
  - Strong communication skills (both verbal and written).
- If you have the relevant background, we can talk about what we can do together in **early January 2023**!