STA 5820
Chapter 1
Introduction

Kazuhiko Shinki

Wayne State University

Probability and Statistics

What is the difference between probability and statistics?

- Probability is to design a casino. - based on mathematical assumptions
- Statistics is to play in a casino. - based on data

(By Alexander Korostelev, a professor Emeritus at WSU)

Statistics is developed based on probability theory.

Probability and Statistics

Example: What is the probability to get the head side up when you spin a penny?

Probabilists' answer: (We assume) 50%.

Statisticians' answer: Spin a penny a thousand times. If the head side appears 210 times, the estimated probability is 21%. To be on the safe side, the standard error for this estimate is $\sqrt{0.5 \times 0.5/1000} \approx 0.016$. Therefore, based on the statistical theory, we are 95% confident that the head probability is between $21\% \pm 1.96 \times 1.6\% = 21\% \pm 3.1\%$).

Notes: In fact, the probability of head is around 20%. Source: `https://www.smithsonianmag.com/science-nature/gamblers-take-note-the-odds-in-a-coin-flip-arent-quite-5050-145465423/`.

(Source: `https://www.smithsonianmag.com/science-nature/gamblers-take-note-the-odds-in-a-coin-flip-arent-quite-5050-145465423/`.)

Data Science

Data Science is to extract knowledge or insights from data (Wikipedia).

Data science is perceived as a combination of three components:

- Mathematical/statistical modeling of the data.
- Computer implementation of the models above.
- Data analysis and implication to the real-world problems.

Machine learning

Machine learning is a branch of mathematical/statistical modeling of the data. It emphasizes

- automatic detection of the pattern, and
- forecasting performance (rather than interpretation and correctness of the models).

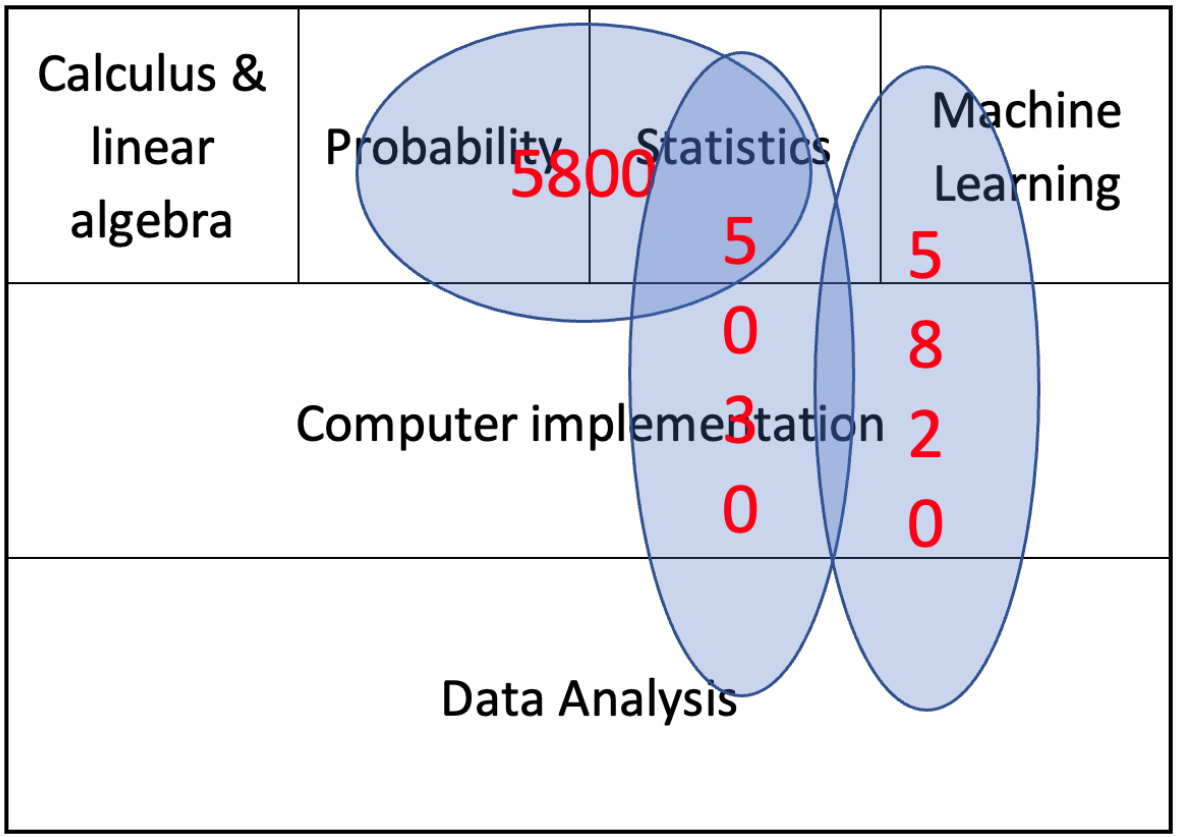
Example: Based on 1000 hand-written characters, build a model for optical character recognition algorithm.

Statistical learning

Statistical learning is mostly the same as machine learning, and is not a common word besides our textbook's title. Most machine learning models have a statistical foundation, and the word can be interpreted as machine learning with emphasis on statistical justification.

Prerequisites

Data science requires a variety of knowledge and ideas - mathematics, probability, statistics, machine learning, algorithms, computer implementation, data analysis and domain knowledge. Think flexibly to learn new concepts you have not studied before.



How are data perceived in statistics?

In statistics, data are perceived as a realization of random variables. This is different from some other fields of study.

Example: How is one's academic ability determined?

A statistician's answer:

$$\textit{Ability} = \textit{Genes} + \textit{Environment} + \textit{Error} \tag{1}$$

The error may be ultimately explained by genes and environment, but it is reasonable to assume that there is an uncertain component since we do not have all information on genes and environment.

Supervised/Unsupervised learning

Supervised learning is a modeling problem when a response variable (called label) is available for the data set. That is, the "answer" is known for the sample data.

Example 1: Estimating a formula for height of students given that weight of the students. Height and weight data are available for 100 students. (regression).

Supervised/Unsupervised learning

Example 2: Establish a criterion to distinguish 3 species of flowers, when characteristics (color, height, diameter etc.) and species of 90 flowers are available. (classification)

Supervised/Unsupervised learning

Unsupervised learning is a modeling problem when a label is not available for the dataset. That is, the "answer" is unknown for the sample data.

Example 1: Developing a business cycle indicator, when 30 economic time series data (such as monthly unemployment rate and monthly industrial production) are available. (dimension reduction)

Supervised/Unsupervised learning

Example 2: Establish a criterion to distinguish 3 species of flowers, when characteristics (color, height, diameter etc.) of 90 flowers are available but species of them are unknown (classification).

Supervised/Unsupervised learning

Semisupervised learning is a modeling problem when a label is available for a part of sample data.

An Example: Developing a translation algorithm from English to Chinese, when side-by-side translation is available for 1,000 documents, but 10,000 documents are available only in English.

An overview of the textbook

- Ch.2 Statistical Learning
- Ch.3 Linear Regression
- Ch.4 Classification
  - Logistic regression, LDA, QDA, KNN.
- Ch.5 Resampling Methods
- Ch.6 Linear Model Selection and Regularization
  - Subset selection, shrinkage, dimension reduction.
- Ch.7 Moving Beyond Linearity
  - Polynomial regression, splines, local regression, GAMs.
- Ch.8 Tree-based Methods
- Ch.9 Support Vector Machines
- Ch.10 Deep learning (neural network)
- Ch.12 Unsupervised Learning
  - Principal component analysis, K-mean clustering.

How to use R

Install R and then RStudio into your computer by the next class. Both are needed to do homwork, the take-home exam and the final project. R and RStudio are open source and available for Windows, Mac OS X and Linux.

- R is available on the CRAN (Comprehensive R Archive Network) Website at `https://cran.r-project.org`. R is a statistical language/software.
- R studio is available on the RStudio Website `https://www.rstudio.com`. This is an integrated development environment (i.e., an add-on) to R.

Memo