Statistics 101C - Introduction to Statistical Models and Data Mining

Shirong Xu

University of California, Los Angeles

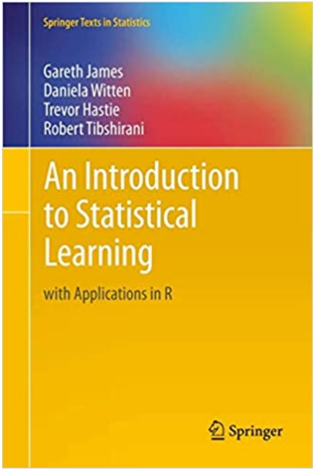
shirongxu56@ucla.edu

September 25, 2024

The goal of this course

- You learn the basic methods and concepts of statistical learning
- You can apply some statistical models to analyze a real dataset

Textbook



- free download from UCLA library
- Cover Ch 2, 4, 5, 6, 8, 9 and 10
- R or Python

Offcie Hour

- **Me**:
  - Email: shirongxu56@ucla.edu
  - Office: Bolter Hall 9401
  - Office Hours: Friday 3:00 - 5:00 pm
  - Questions: Post it on Bruinslearn and I will reply to them on each Friday

- **TA and Grader: Zhi Zhang (2A and 2B) and Alex Chen (1A and 1B)**:
  - Email: zzh237@g.ucla.edu & aclheexn1346@g.ucla.edu
  - Office Hours: available on the first discussion.

Grading Scheme

- Homework: 30%
  - 5 homework assignments: each takes up 6% or 6 points
  - Submit Homework on CCLE website
  - Late homework is acceptable but at most get 80%.
  - If you submit your homework late, just **email the grader**. (Lec 1: Alex Chen) and (Lec 2: Zhi Zhang)

Grading Scheme

- Homework: 30%
  - 5 homework assignments: each takes up 6% or 6 points
  - Submit Homework on CCLE website
  - Late homework is acceptable but at most get 80%.
  - If you submit your homework late, just **email the grader**. (Lec 1: Alex Chen) and (Lec 2: Zhi Zhang)
- Mid-term exam: 40% - On Nov. 14 (Thursday)

Grading Scheme

- Homework: 30%
  - 5 homework assignments: each takes up 6% or 6 points
  - Submit Homework on CCLE website
  - Late homework is acceptable but at most get 80%.
  - If you submit your homework late, just **email the grader**. (Lec 1: Alex Chen) and (Lec 2: Zhi Zhang)
- Mid-term exam: 40% - On Nov. 14 (Thursday)
- Final Project: 30%.

Grading Scheme

- Homework: 30%
  - 5 homework assignments: each takes up 6% or 6 points
  - Submit Homework on CCLE website
  - Late homework is acceptable but at most get 80%.
  - If you submit your homework late, just **email the grader**. (Lec 1: Alex Chen) and (Lec 2: Zhi Zhang)
- Mid-term exam: 40% - On Nov. 14 (Thursday)
- Final Project: 30%.
- Grade Scale (Ranking):
  - 10%: A+,
  - 10%-40%: A,
  - 40%-70%: A-,
  - 70%-85%: B+,
  - 85%-90%: B,
  - 90%-100%: B- and Below

Final Project

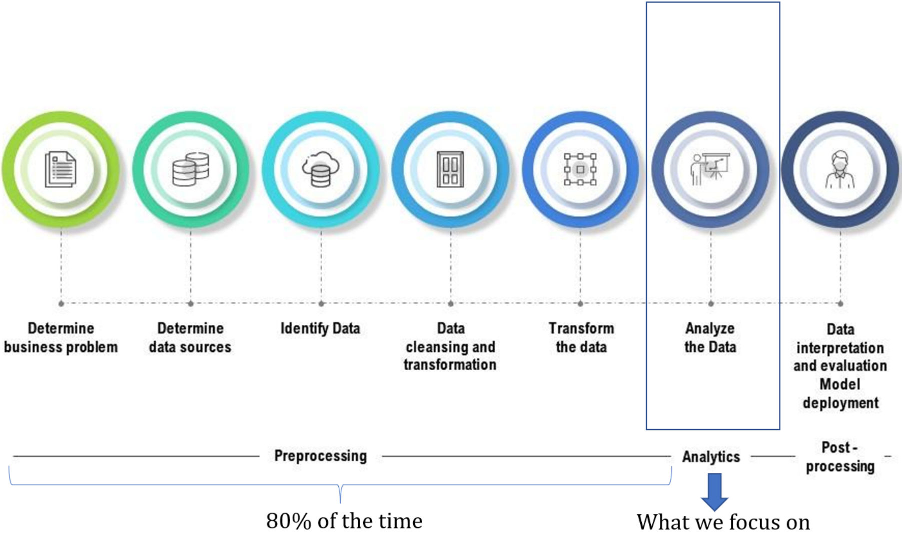
- Group Project: 4-6 people
- Dataset: the dataset will be available around the midterm
- **Output**: a cute paper (at least 2 page but less than 10 pages) describing how you analyze the dataset. It should contains
  - How do you pre-process the data?
  - What kind of models you apply to the pre-processed dataset?
  - Any interesting results or conclusion?

Final Project - Grading

Reviewing your final projects: two-round reviews

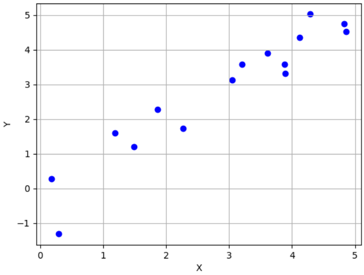
1. First round review
2. First round review releases: initial score and comments
3. Second round review: determine whether you submit a revised version (Up to you). If not, the initial score will be the final score.
4. Second round review release: final score.

Statistical learning in the real world



Data Mining and Statistical Models

- Suppose we observe a dataset:

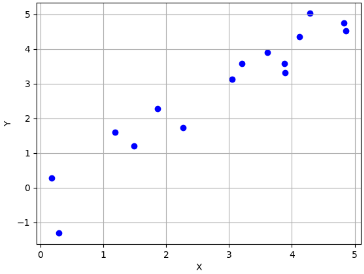


- What is data mining?

- What is a statistical model?

Data Mining and Statistical Models

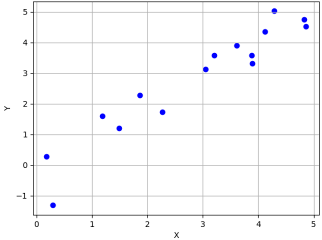
- Suppose we observe a dataset:



- **What is data mining**?
  - Data mining is a process of discovering patterns in large data sets involving methods at the intersection of machine learning and statistics.

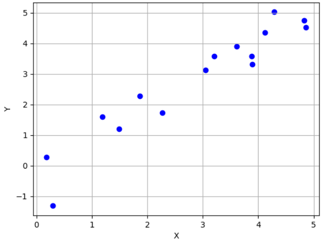
Statistical Model

Statistical model: is a mathematical model that embodies a set of statistical **assumptions** concerning the generation of sample data.



Statistical Model

Statistical model: is a mathematical model that embodies a set of statistical **assumptions** concerning the generation of sample data.

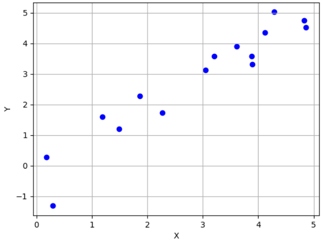


Assumptions:

- $Y = f(X) + \epsilon$, where f is a true model

Statistical Model

Statistical model: is a mathematical model that embodies a set of statistical **assumptions** concerning the generation of sample data.

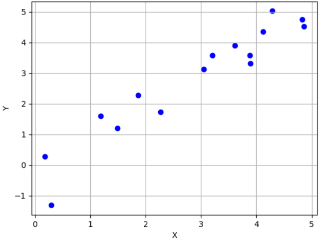


Assumptions:

- $Y = f(X) + \epsilon$, where f is a true model
- X and ϵ are independent

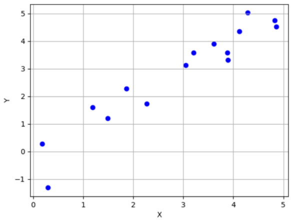
Statistical Model

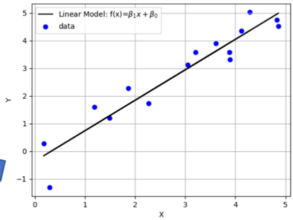
Statistical model: is a mathematical model that embodies a set of statistical **assumptions** concerning the generation of sample data.



Assumptions:

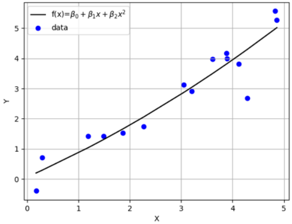
- $Y = f(X) + \epsilon$, where f is a true model
- X and ϵ are independent
- $\mathbb{E}(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$





Assumption: the relationship between X and Y is Linear!

Assumption: the relationship between X and Y is Quadratic!



Another Example - Ranking of Basketball players

Problem: The prefernce **ranking** of these basketball players

Data Collection - Pairwise comparisons

Kobe or Lebron?

We can let $y \in \{0, 1\}$ denote the binary choice. $y = 1$ means that your answer is kobe and $y = 0$ means lebron.

Question: How to model the observation y using a statistical model?

$$\mathbb{P}(\text{Kobe is chosen over Lebron}) = \frac{e^{\alpha_{kobe}}}{e^{\alpha_{kobe}} + e^{\alpha_{leborn}}}.$$

Here α_{kobe} can be understood as a **popularity parameter** of kobe. After collecting all data, we can estimate the popularity parameters of all basketball players and give a ranking based on $\alpha_{kobe}, \alpha_{leborn}, \alpha_{Curry}, \cdots$.

Parametric models v.s. Non-parametric

- **Parametric models**: Situations like linear regression, in which we can describe the functional form of $f(x)$ using a finite number of parameters are called parametric models. Like

$$f(x) = \beta_0 + \beta^T x$$

- Once we know the form of f, the estimation of f reduces to estimating the parameters β_0 and β.

Parametric models v.s. Non-parametric

- **Non-Parametric models**: Simply, a model that is not parametric. There are many different interpretations to this statement.
- In this course, a non-parametric models is one that does not make explicit assumptions about the form of f, like KNN and decision tree.
- **Question**: Is the number of K in KNN a parameter?

Parametric models v.s. Non-parametric

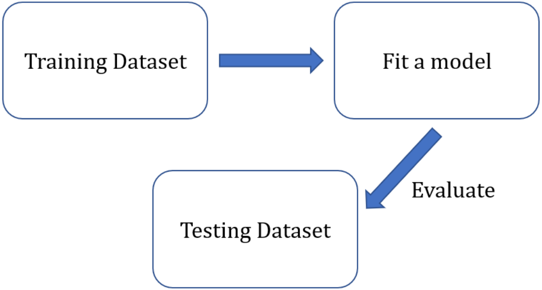
- **Non-Parametric models**: Simply, a model that is not parametric. There are many different interpretations to this statement.
- In this course, a non-parametric models is one that does not make explicit assumptions about the form of f, like KNN and decision tree.
- **Question**: Is the number of K in KNN a parameter? No. K in KNN is a hyperparameter.
- **Hyperparameters** are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning. Examples:
  - The number of layers and the width in deep neural networks.
  - The depth of decision tree
  - Learning rate in optimization algorithms (e.g. gradient descent)

Two cultures of models?

- **Inference**: develop a model that fits the data well. Then make inferences about the data-generating process based on the structure of such model.
- **Prediction**: Silent about the underlying mechanism generating the data and only care about accuracy of predictions. Machine learning researchers more care about whether a model is **state-of-the-art** (SOTA)

Training dataset v.s. Testing Data

- **Training data**: data used to fit a model
- **Testing**: data that were NOT used in the fitting process, but are used to test how well your model performs on unseen data.
- **Validation**: Usually, a validation dataset will be available for helping choose the best parameter of models.



Notations you should know

- For a random variable, the density function is $\mathbb{P}(x)$
- Expectation of a random variable X (denoted as $\mathbb{E}(X)$):

$$\mathbb{E}(X) = \int X\mathbb{P}(x)\,dx$$

- **Argmin and Argmax:**

$$x^* = \arg\min_x f(x) \quad \text{and} \quad x_0 = \arg\max_x f(x)$$

where $f(x^*) = \min f(x)$ and $f(x_0) = \max f(x)$

- **Function class $\mathcal{F}$:** (a set of functions)

$$\mathcal{F} = \{f(x) = \beta x : \beta \in \mathbb{R}\}$$

where $\mathbb{R}$ is the set of all real values.

- Minimize or maximize an objective with respect to a function class

$$f^* = \arg\min_{f \in \mathcal{F}} L(f)$$

where $L(f)$ is the objective function.

- **A linear regression example:**

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

It can be equivalently represented as

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

where $\mathcal{F} = \{f(x) = \beta x : \beta \in \mathbb{R}\}$