MACHINE LEARNING DATA - INTRODUCTION

Institute of Technology of Cambodia

October 16, 2023

Who Am I

- UN Lykong (un.lykong@gmail.com, 0969512202)
- Class page `https://github.com/Lykong123/AI_course`
- A lecturer, Institute of Technology of Cambodia

Definition

- Machine learning(ML) involves computers discovering how they can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks. It is a subset of Artificial Intelligence.
- `https://en.wikipedia.org/wiki/Machine_learning`

This Class

- Less math
  - Math are everywhere in ML. But most of it is applying simple math. If you need to understand the hard one, it is simpler to take a book.
- More intuition
  - I want you to understand the large picture. You can focus on what you like.
- No learning by heart
  - Remember that the concept exist, so you can google it.
- And some practice
  - Huge amount of resources available for free.

This class

- This class is based on
  - Countless Wikipedia and blogs (use them too!).
- Some books
  - `https://dataminingbook.info/`
  - `http://ema.cri-info.cm/wp-content/uploads/2019/07/2019BurkovTheHundred-pageMachineLearning.pdf`

Class Overview

- Data description, preparation.
- Unsupervised machine learning (beyond k-means)
- Supervised machine learning (beyond linear regression)
- Deep learning (applicable)
- Project/other topics

Class Overview

- Your past experience on machine learning?
- What is machine learning for you?

TYPE OF DATA

Data Types

Data types : What kind of data (feature, variables) can we encounter?

- People
  - Name, Age, Gender, Revenue, Birth Date, Address, etc.
- House/Apartment
  - Surface area, Floor, Address, number of rooms, number of Windows, Elevator, etc.

Data Types

- Continuous features
  - time, weight, income, temperature, etc.
- Categorical features
  - male, female, red, blue. etc
  - Categorical data divided into nominal and ordinal

Data Types

- Nominal
  - From “names”. No order between possible values
  - Color, Gender, Animal, Brand, etc. (Numbers:Participant ID, class. . .)
- Ordinal
  - Order between values, but not numeric
  - Size[small, medium, large], [Satisfied, . . . , Unsatisfied]

Missing Values

- Real life datasets are full of missing values
  - Impossible data: hair color for a bald person
  - More generally, failed to obtain them
- Few ML method can deal with missing values
  - Imputation
  - Naive: fill with average value
  - Use ML to fill missing values (other problems, introduce biases...)

Data Quality

- Data coming from the real world is often incorrect
  - Malfunctioning sensors (speed...)
  - Human error or falsification (e.g., entered 100 instead of 1.00)
  - Undocumented change (e.g., Bicycle sharing station was removed..)
- If the data is plausible, no simple solutions
- Two common problems can be detected
  - Out-of-range values (e.g., a person's weight is negative or above 1000kg...)
  - Zeros (weight or the person is 0. but in many cases, zero is possible too..)

DESCRIBING A VARIABLE

Describing Values

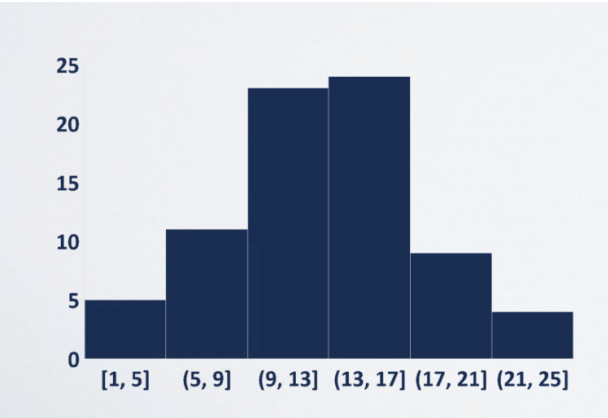
- Mean/Average
  - Be careful, not necessarily representative!
- Median
  - Be careful, not necessarily representative!
- Mode
  - Not necessarily representative!
- Min/Max

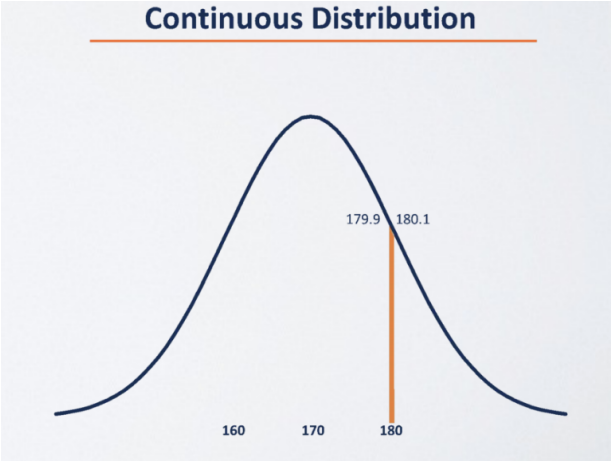

Median
Mean

Distribution

What is a distribution?

- A description of the frequency of occurrence of items
- A generative function describing the probability to observe any of the possible events
- Discrete or continuous





Variance

Variance:
- expectation of the squared deviation of random variable from its mean

$$Var(X) = \sigma^2 = E[(X - \mu)]^2$$

Also expressed as average squared distance between all elements

$$\sigma^2 = \frac{1}{N^2} \sum_{i<j} (x_i - x_j)^2$$

STANDARD DEVIATION

- Squared root of the Variance

$$\sigma = \sqrt{\sigma^2} = \sqrt{E[(X - \mu)^2]}$$

VARIABLE INTERACTIONS

Covariance Matrix

- Covariance matrix K

$$cov(X, Y) = K_{XY} = E[(X - E[X])(Y - E[Y])^T$$

or

$$\Rightarrow cov(X, X) = Var(X)$$

- Extension of Variance to multivariate data
- How much observation X differs from the mean? and Y?
- Multiply the respective divergences of X and of Y for each item
- Take the average

- Covariance is hardly interpretable by itself
- If , divergences tend to be in the same direction
- Normalize it to obtain the "correlation coefficient"

Correlation Coefficient

Pearson correlation coefficient:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

- Normalize the Covariance by the Standard deviation
- Independent from magnitude, i.e., no need to have normalized data
- Value in -1, +1
  - +1 means a perfect positive linear correlation, i.e., X = aY
  - -1 a negative one i.e., X=-bY
- O can mean many different things

Feature Scaling

We want to use euclidean distance to compute the "distance" between 2 people based on attributes age(y), height(m), weight(m)

- a = (y:20,m:1.82,g:80 000), b=(y:20,m:1.82,g:81000), c=(y:90,m:1.50,g=80 020)
  - d(a,b) = 1000.0005
  - d(a,c) = 72.8
- That is not what we expected from our expert knowledge!
  - We should normalize/standardize data

Feature Scaling

- Rescaling (Normalization):

$$\chi' = \frac{\chi - min(\chi)}{max(\chi) - min(\chi)} : [0, 1]$$

- Mean normalization:

$$\chi' = \frac{\chi - average(\chi)}{max(\chi) - min(\chi)} : 0 = mean$$

- Standardization (z-score normalization):

$$\chi' = \frac{\chi - \overline{\chi}}{\sigma}$$

- 0:mean, -1/+1:1 standard deviation from the mean

Experiments

- Go to the webpage of the class and do today's experiments
- The "advance" section is not mandatory, you can do it if you have time