Machine Learning Mathematics -Statistics

Machine Learning II

Lecture 2-a

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

- Machine learning combines statistics and computer science fields.
- Statistics, probability, estimation and confidence intervals are some of main topics in machine learning in the statistical part.
- Linear algebra is another mathematical skill which is highly necessary in machine learning.
- Optimization theory and calculus are widely used in machine learning algorithms.

Why we need math?

- There are so many machine learning codes out there, and they are fairly simple to run.
- You need to download the packages and library to run the machine learning algorithms.
- However, to get some useful results and meaningful performance, you need to have a good mathematical background.

- After this lecture, you will get a glimpse of what types of mathematical skills you need to practice.
- In this lecture, we go over the main topics and further materials will be provided to you in order to strengthen you mathematical skills.

Definition of Probability?

- Relative Frequency.
- Subjective Probability.
- Axiomatic Probability.

- $a \in A$ a is member of A.
- $\cup$ is union, $\cap$ is intersection.
- $\sum$ is summation.
- $\int$ is integral.
- R is set of real numbers.
- **a,b,c** vector.
- **A,B,C** Matrix.
- $\frac{\sigma}{\sigma x} f(x)$ is a function.
- $y = f(x)$ is a function.
- $\frac{d}{dx} f(x)$ is a function.
- $||A||$ is norm A.
- a,b,c is set.
- $\emptyset$ is empty set.
- $\subset$ is subset.
- $y = f(\mathbf{x})$

My note - Venn Diagram

Axiomatic Probability

- A probability needs to satisfy three properties (Kolmogorov, 1956):
- A probability must be nonnegative.
- The sum of the probabilities across all events in the entire sample space must be 1.
- For any two mutually exclusive events, the probability that one or the other occurs is the sum of their individual probabilities.

- $P(A) >= 0$
- $P(s) = 1$
- $P(A \cup B) = P(A) + P(B)$ A and B are disjoint.
- $A \cap B = \emptyset$ is disjoint.
- $A \cap B = P(A) \times P(B)$ if A and B are independent.

My note - Example rolling a die

- $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- $P(B|A) = \frac{P(A \cap B)}{P(A)}$
- $P(A \cap B) = P(B|A) \times P(A)$
- $P(A \cap B) = P(A|B) \times P(B)$
- $P(B|A) = \frac{P(A|B) \times P(B)}{P(A)}$ BAYES RULE.

- A random variable is a mapping from sample space to the real line.
- $F_x(\lambda) = P(s : x(s) <= \lambda)$ Distribution function.
- $f_x(\lambda) = \frac{d}{d\lambda} F_x(\lambda)$ Probability Density function.
- $F_x(\lambda) = \int_{-\infty}^{\lambda} f_x(\mu) d\mu$
- $E[x] = \int_{-\infty}^{\infty} \lambda f_x(\lambda) d\lambda$

Random variables and density functions

Probability distribution

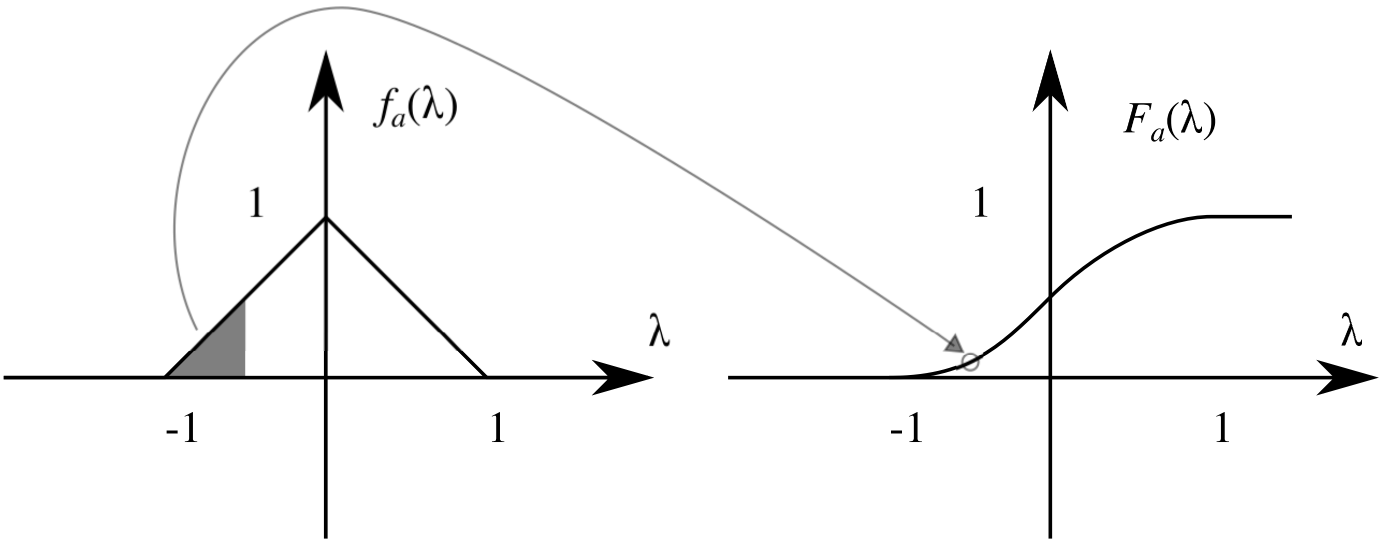
$$F_a(\lambda) = P(a \leq \lambda)$$

Probability density

$$f_a(\lambda) = \frac{\delta F_a(\lambda)}{\delta \lambda}$$

$$F_a(\lambda) = \int_{-\infty}^{\lambda} f_a(\gamma) d\gamma$$

Example density and distribution functions



Gaussian

$$f_a(\lambda) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\lambda-\mu)^2}{2\sigma^2}}$$

Gamma

$$f_a(\lambda) = \frac{1}{\Gamma(k)\theta^k} \lambda^{k-1} e^{-\frac{\lambda}{\theta}}$$

Maxwell-Boltzmann

$$f_a(\lambda) = \sqrt{\left(\frac{m}{2\pi kT}\right)^3 4\pi\lambda^2 e^{-\frac{m\lambda^2}{2kT}}}$$

Joint density

$$f_{a_1,a_2}(\lambda_1,\lambda_2) = f_{a_1|a_2}(\lambda_1|\lambda_2)f_{a_2}(\lambda_2)$$

Marginal density

$$f_{a_1}(\lambda_1) = \int_{-\infty}^{\infty} f_{a_1,a_2}(\lambda_1,\lambda_2)d\lambda_2$$

Conditional density

$$f_{a_2|a_1}(\lambda_2|\lambda_1) = \frac{f_{a_1,a_2}(\lambda_1,\lambda_2)}{f_{a_1}(\lambda_1)}$$

Bayes rule

$$f_{a_1|a_2}(\lambda_1|\lambda_2) = \frac{f_{a_2|a_1}(\lambda_2|\lambda_1)f_{a_1}(\lambda_1)}{f_{a_2}(\lambda_2)}$$

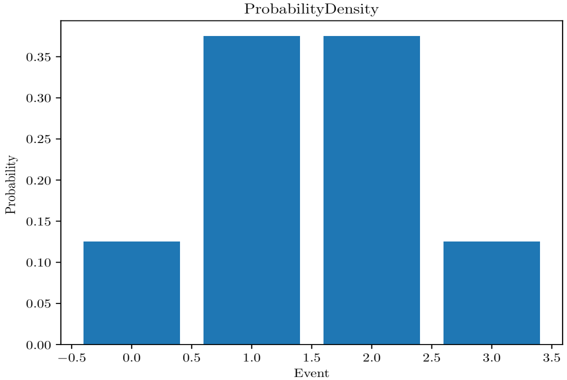
- If some parameters of the density are unknown, we can collect samples of the random variable and estimate the parameters.
- For example, given a set of independent samples from a Gaussian density, we can estimate the mean using the average value.

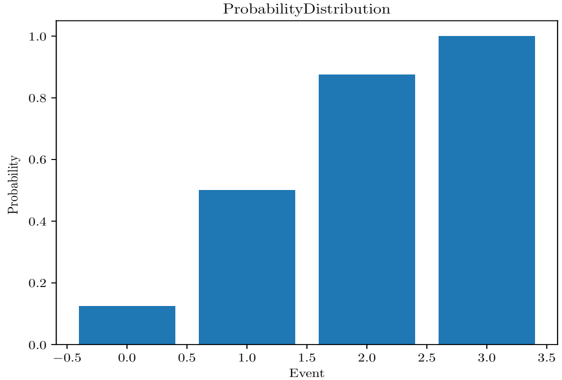
$$\hat{\mu} = \frac{1}{Q} \sum_{i=1}^{Q} a_i$$

- Therefore, the probability distribution for the number of heads occurring in three coin tosses is

Count of Heads (X)	P(X)	P(X<=X)
0	1/8	1/8
1	3/8	4/8
2	3/8	7/8
3	1/8	1

$$P(X) = \begin{cases} \frac{1}{8} & if\ x = 0 \\ \frac{3}{8} & if\ x = 1{,}2 \\ \frac{1}{8} & if\ x = 3 \\ 0 & if\ otherwise \end{cases}$$

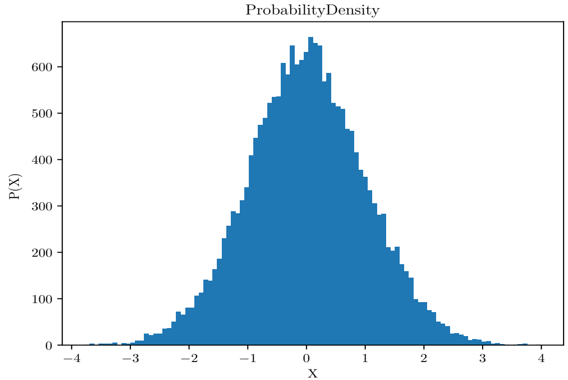




- When the sample space consists of continuous outcomes (ex: people's heights) we cannot use probability mass for a specific outcome.
- Because the probability mass for a specific outcome will be zero.
- In other words, the probability of someone's height being exactly 67.214139084.
- Discretize the space into a finite set of mutually exclusive and exhaustive intervals.
- Calculate the probability mass in each interval.
- Use the ratio of probability mass to interval width.
- This ratio is called the Probability Density

The Normal Probability Density Functions

- Perhaps the most famous probability density function is the normal distribution, also known as the Gaussian distribution
- The probability density function of normal distribution is
- $$p(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$



Python GitHub Repo

- Check my GitHub and answer the exercises.
  - https://github.com/amir-jafari/Machine-Learning/tree/master/Python-Math