Mathematical Background

Yuh-Jye Lee

Data Science & Machine Intelligence Lab
Dept. of Applied Math @ NCTU

February 22, 2017

Outline

1. Probability and Statistics

2. Probability and Inference

Outline

Random Variable

Definition

A *random variable* is a real-valued function for which domain is a sample space

- Example

  For a coin toss, the possible outcome is head or tail. The number of heads appearing in one fair coin toss can be described using the following random variable:

$$X = \begin{cases} 1, & \text{if head} \\ 0, & \text{if tail} \end{cases}$$

with probability function given by:

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

Probability Distribution

Definition

If X is discrete random variable, the function given by $P(X = x)$ for each x within the range of X is called probability distribution of X.

- Example
  Let the random variable X be denoted as the total number of heads. The probability distribution of heads obtained in the **four** tosses of a fair coin can be written as follows:

$$P(X = x) = \frac{\binom{4}{x}}{2^4}, \text{ for } x = 0, 1, 2, 3, 4.$$

Probability Density Distribution

Definition

A function with values $f(x)$, defined over the set of all real numbers, is called a probability density function of the continuous random variable X if and only if

$$P(a \leq X \leq b) = \int_a^b f(x)dx,$$

for any real constants a and b with $a \leq b$

- Example

  The p.d.f of normal distribution is defined as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

where μ is the mean and σ is the standard deviation.

Conditional Probability

Definition

The conditional probability of an event A, given that an event B has occurred, is equal to

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Example
  Suppose that a fair die is tossed once. Find the probability of a 1 (event A), given an odd number was obtained (event B).

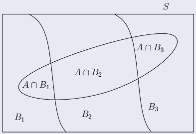
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = \frac{1}{3}$$

- Restrict the sample space on the event B

Theorem

Assume that $\{B_1, B_2, \ldots, B_k\}$ is a partition of S such that $P(B_i) > 0$, for $i = 1, 2, \ldots, k$. Then

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i).$$



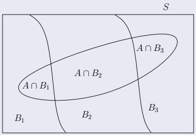
- Note that $\{B_1, B_2, \ldots, B_k\}$ is a partition of S if
  1. $S = B_1 \cup B_2 \cup \ldots \cup B_k$
  2. $B_i \cap B_j = \emptyset$ for $i \neq j$

Bayes' Rule

Bayes' Rule

Assume that $\{B_1, B_2, \ldots, B_k\}$ is a partition of S such that $P(B_i) > 0$, for $i = 1, 2, \ldots, k$. Then

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}.$$



Expected Value

Definition

If X is a discrete random variable and $P(X = x)$ is the value of its probability distribution at x, the expected value of X is

$$\mu = E(X) = \sum_x x \cdot P(X = x).$$

Correspondingly, if X is a continuous random variable and $f(x)$ is the value of its probability density at x, the expected value of X is

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx.$$

- $E(aX + bY) = aE(X) + bE(Y)$, *linear operator*

Variance
Measures of how far a set of numbers are spread out

Definition

If X is a discrete random variable and $P(X = x)$ is the value of its probability distribution at x, the expected value of X is

$$Var(X) = E([X - E(X)]^2) = \sum_x (x - \mu)^2 \cdot P(X = x).$$

Correspondingly, if X is a continuous random variable and $f(x)$ is the value of its probability density at x, the expected value of X is

$$Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx.$$

- $Var(X) = E(X^2) - (E(X))^2$

Bernoulli Distribution

A trial is performed whose outcome is either a "success" or a "failure". The random variable X is a $0/1$ indicator variable and takes the value 1 for a success outcome and is 0 otherwise. p is the probability that the result of trail is a success. Then

$$P(X = 1) = p \text{ and } P(X = 0) = 1 - p$$

which can equivalently be written as

$$P(X = i) = p^i(1 - p)^{1-i}, \; i = 0, 1$$

Tossing a *fair* coin, the parameter $p = 0.5$. If X is Bernoulli,

1. $E(X) = p$,
2. $Var(X) = p(1 - p)$
3. Who knows p?

Probability and Inference

- The outcome of tossing a coin is $\{Heads, Tails\}$
- We use a random variable $X \in \{0, 1\}$ to indicate the outcome
- Suppose that we have a random sample: $\mathbf{X} = \{x^t\}_{t=1}^N$
- How to *estimate* the parameter p?

Maximum Likelihood Estimation

Likelihood Function

The probability to *observe* the random sample $\mathbf{X} = \{x^t\}_{t=1}^N$ is

$$\prod_{t=1}^{N} p^{x^t}(1-p)^{1-x^t}$$

Why don't we choose the parameter p which will maximize the probability for observing the random sample $\mathbf{X} = \{x^t\}_{t=1}^N$?

Based on MLE, we will choose the parameter p

$$p = \frac{\sum_{t=1}^N x^t}{N}$$

Sample Mean, Variance, and Standard deviation

Sample Mean

The mean of a sample of n measured responses $y_1, y_2, \ldots, y_n$ is given by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

The corresponding population mean is denoted by μ.

Sample Variance

The variance of a sample of measurements $y_1, y_2, \ldots, y_n$ is given by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

The corresponding population variance is denoted by σ^2.

Applying Baye's Rule to Classification

Credit Cards Scoring: Low-risk vs. High-risk

- According to the past transactions, some customers are low-risk in that they paid back their loan and the bank profited from them and other customers are high-risk in that they defaulted.
- We would like to *learn* the class "*high-risk customer*"
- We observe customer's *yearly income* and *savings*, which we represent by two *random variables* X_1 and X_2
- The *credibility of a customer* is denoted by a *Bernoulli* random variable C where $C = 1$ indicates a high-risk customer and $C = 0$ indicated a low-risk customer

Applying Baye's Rule to Classification

How to make the decision when a new application arrives?

- When a new application arrives with $X_1 = x_1$ and $X_2 = x_2$
- If we know the probability of C *conditioned on* the observation $X = [x_1,\ x_2]$ our decision will be
  - $C = 1$ if $P(C = 1|[x_1,\ x_2]) > 0.5$
  - $C = 0$ otherwise
- The probability of error we made based on this rule is

$$1 - \max\{P(C = 1|[x_1,\ x_2]), P(C = 0|[x_1,\ x_2])\} < 0.5$$

- Please note $P(C = 1|[x_1,\ x_2]) + P(C = 0|[x_1,\ x_2]) = 1$

The *Posterior Probability*: $P(C|\mathbf{x}) = \frac{P(C)P(\mathbf{x}|C)}{P(\mathbf{x})}$

- $P(C = 1)$ is called the *prior probability* that $C = 1$
- In our example, it corresponds to a probability that a customer is high-risk, *regardless* of the $\mathbf{x}$ value.
- It is called the *prior probability* because it is the knowledge we have *before* looking at the observation $\mathbf{x}$
- $P(\mathbf{x}|C)$ is called the *class likelihood* and is the *conditional probability* that an *event belonging to the class C* has the associated observation value $\mathbf{x}$
- $P(\mathbf{x})$, the *evidence* is the probability that an observation $\mathbf{x}$ to be seen, regardless of whether it is a positive or negative example

All above information can be extracted from the past transactions *(historical data)*

The *Posterior Probability*: $P(C|\mathbf{x}) = \frac{P(C)P(\mathbf{x}|C)}{P(\mathbf{x})}$

- Because of normalization by the evidence, the posteriors sum up to 1
- In our example, $P(X_1, X_2)$ is called the *joined probability* of two random variables X_1 and X_2
- Under the assumption, these two random variables X_1 and X_2 are *probability independent*, we have $P(X_1, X_2) = P(X_1)P(X_2)$
- It is one of key assumptions of *Naive Bayes' Classifier*
- Although it is *over simplified* the problem it is very easy to use for real applications

Extend to Multi-class classification

- We have K mutually and exhaustive classes; $C_i, \;\; i = 1, 2, \ldots, K$
- For example, in *optical digit recognition*, the input is a *bitmap image* and there are 10 classes
- We can think of that these K classes define a *partition* of the *input space*
- Please refer to the slides of the *Partition Theorem* and *Baye's Rule*
- The Bayes' classifier choose the class with the highest posterior probability; that is we choose C_i if

$$P(C_i|\mathbf{x}) = \max_k P(C_k|\mathbf{x})$$

- Question: Is it very important to have $P(\mathbf{x})$, the evidence?