Introduction to Machine Learning

Statistical Machine Learning

Varun Chandola

March 29, 2021

Outline

Contents

1 Statistical Machine Learning - Introduction

Statistical Machine Learning

Functional Methods

- $y = f(\mathbf{x})$
- Learn $f()$ using training data
- $y^* = f(\mathbf{x}^*)$ for a test data instance

Statistical/Probabilistic Methods

- Calculate the *conditional probability* of the target to be y, given that the input is $\mathbf{x}$
- Assume that $y|\mathbf{x}$ is *random variable* generated from a *probability distribution*
- Learn parameters of the distribution using training data

2 Introduction to Probability

3 Random Variables

- A variable whose value depends on a random phenomenon
  - Mapping random processes to numbers (or values)
- Usually denoted using an upper case letter, $X, Y, \ldots$
- A random variable has:
  - A **domain**: Set of possible values that X can take (denoted as $\mathcal{X}$)
  - A **probability measure** ($f()$) that assigns the probability of X to belong to a subset of $\mathcal{X}$, i.e., $P(X \in S | S \in \mathcal{X})$, with two requirements:
    * $0 \leq f(S) \leq 1$
    * $\sum_i f(S_i) = 1$, where $S_1, S_2, \ldots$ are mutually disjoint subsets of $\mathcal{X}$ and $\cup_i S_i = \mathcal{X}$
- An instance of the probability measure is a **probability distribution** which assigns probability to every element in $\mathcal{X}$

The notion of a random event and a random variable are closely related. Essentially, any random or probabilistic event can be represented as a random variable X taking a value x.

The quantity $P(A = a)$ denotes the probability that the event $A = a$ is true (or has happened). Another notation that will be used is $p(X)$ which denotes the distribution. For discrete variables, p is also known as the **probability mass function**. For continuous variables, p is known as the **probability density function**.

A probability distribution is the enumeration of $P(X = x), \forall x \in \mathcal{X}$.

Two basic types of random variables

Discrete Random Variable

- $\mathcal{X}$ is finite/countably finite
- $P(X = x)$ or $P(x)$ is the probability of X taking value x
  - Categorical?? Categorical variables are those for which the possible values cannot be ordered, e.g., - a coin toss can produce a heads or a tail. The outcome of the toss is a categorical random variable.

Continuous Random Variable

- $\mathcal{X}$ is infinite
- Probability of any one value is 0
- Can only talk about range of values:

$$P(a < X \leq b)$$

- We define the **probability density function** at any location, $p(x)$ or $f(x)$

$$P(a < X \leq b) = \int_a^b p(x)dx$$

$p(x)$ or the pdf for a continuous variable need not be less than 1 as it is not the probability of any event. But $p(x)dx$ for any interval dx is a probability and should be less than 0.

Notation

- X - random variable ($\mathbf{X}$ if multivariate)
- x - a specific value taken by the random variable (($\mathbf{x}$ if multivariate))
- $P(X = x)$ or $P(x)$ is the probability of the event $X = x$
- $p(x)$ is either the **probability mass function** (discrete) or **probability density function** (continuous) for the random variable X at x
  - Probability mass (or density) at x

Basic Rules

- For two events A and B:
  - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
  - **Joint Probability**
    * $P(A, B) = P(A \wedge B) = P(A|B)P(B)$
    * Also known as the *product rule*
  - **Conditional Probability**
    * $P(A|B) = \frac{P(A,B)}{P(B)}$

Note that we interpret $P(A)$ as the probability of the random variable to take the value A. The event, in this case is, the random variable taking the value A.

- Given D random variables, $\{X_1, X_2, \ldots, X_D\}$

  $P(X_1, X_2, \ldots, X_D) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \ldots P(X_D|X_1, X_2, \ldots, X_D)$

- Given $P(A, B)$ what is $P(A)$?
  - Sum $P(A, B)$ over all values for B

$$P(A) = \sum_b P(A, B) = \sum_b P(A|B = b)P(B = b)$$

  - **Sum rule**

4 Bayes Rule

- Computing $P(X = x|Y = y)$:

Bayes Theorem

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P(X = x)P(Y = y|X = x)}{\sum_{x'} P(X = x')P(Y = y|X = x')}$$

Bayes Theorem: Example

- Medical Diagnosis
- Random event 1: A *test* is positive or negative (X)
- Random event 2: A person has cancer (Y) – yes or no
- What we know:
  1. Test has accuracy of 80%
  2. Number of times the test is positive when the person has cancer

  $$P(X = 1|Y = 1) = 0.8$$

  3. Prior probability of having cancer is 0.4%

  $$P(Y = 1) = 0.004$$

  Question?

  If I test positive, does it mean that I have 80% rate of cancer?

- Ignored the prior information
- What we need is:

  $$P(Y = 1|X = 1) = ?$$

- More information:

  - False positive (alarm) rate for the test
  - $P(X = 1|Y = 0) = 0.1$

$$P(Y = 1|X = 1) = \frac{P(X = 1|Y = 1)P(Y = 1)}{P(X = 1|Y = 1)P(Y = 1) + P(X = 1|Y = 0)P(Y = 0)}$$

$$P(Y = 1|X = 1) = \frac{P(X = 1|Y = 1)P(Y = 1)}{P(X = 1|Y = 1)P(Y = 1) + P(X = 1|Y = 0)P(Y = 0)} = \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} = 0.031$$

Classification Using Bayes Rule

- Given input example $\mathbf{x}$, find the true class

$$P(Y = c|\mathbf{X} = \mathbf{x})$$

- Y is the random variable denoting the true class
- Assuming the **class-conditional** probability is known

$$P(\mathbf{X} = \mathbf{x}|Y = c)$$

- Applying Bayes Rule

$$P(Y = c|\mathbf{X} = \mathbf{x}) = \frac{P(Y = c)P(\mathbf{X} = \mathbf{x}|Y = c)}{\sum_c P(Y = c'))P(\mathbf{X} = \mathbf{x}|Y = c')}$$

Independence and Conditional Independence

- One random variable does not depend on another
- $A \perp B \iff P(A, B) = P(A)P(B)$
- Joint written as a product of marginals

Two random variables are independent, if the probability of one variable taking a certain value is not dependent on what value the other variable takes. Unconditional independence is typically rare, since most variables can influence other variables.

- **Conditional Independence**

$$A \perp B | C \iff P(A, B|C) = P(A|C)P(B|C)$$

Conditional independence is more widely observed. The idea is that all the information from B to A "flows" through C. So B does not add any more information to A and hence is independent conditionally.

Expectation

- Let $g(X)$ be a function of X
- If X is discrete:

$$\mathbb{E}[g(X)] \triangleq \sum_{x \in \mathcal{X}} g(x)P(X = x)$$

- If X is continuous:

$$\mathbb{E}[g(X)] \triangleq \int_{\mathcal{X}} g(x)p(x)dx$$

Properties

- $\mathbb{E}[c] = c$, c - constant
- If $X \leq Y$, then $\mathbb{E}[X] \leq \mathbb{E}[Y]$
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- $\mathbb{E}[aX] = a\mathbb{E}[X]$
- $var[X] = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - \mu^2$
- $Cov[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- Jensen's inequality: If $\varphi(X)$ is convex,

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$$

- Expected value of a random variable

$$\mathbb{E}[X]$$

- What is most likely to happen in terms of X?
- For discrete variables

$$\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} xP(X = x)$$

- For continuous variables

$$\mathbb{E}[X] \triangleq \int_{\mathcal{X}} xp(x)dx$$

- **Mean** of X (μ)

While the probability distribution provides you the probability of observing any particular value for a given random variable, if you need to obtain one representative value from a probability distribution, it is the expected value. Another way to understand it is that a probability distribution can give any sample, but the expected value is the **most likely sample**.

Another way to explain the expectation of a random variable is a *weighted average* of values taken by the random variable over multiple trials.

- Spread of the distribution

$$\begin{aligned} var[X] &\triangleq \mathbb{E}((X - \mu)^2) \\ &= \mathbb{E}(X^2) - \mu^2 \end{aligned}$$

$$\begin{aligned} var[X] &\triangleq \mathbb{E}((X - \mu)^2) \\ &= \int (x - \mu)^2 p(x)dx \\ &= \int x^2 p(x)dx + \mu^2 \int p(x)dx - 2\mu xp(x)dx \\ &= \mathbb{E}(X^2) - \mu^2 \end{aligned}$$

5 Different Types of Distributions

Discrete

- Binomial, *Bernoulli*
- Multinomial, *Multinoulli*
- Poisson
- Empirical

Continuous

- Gaussian (Normal)
- Degenerate pdf
- Laplace
- Gamma
- Beta
- Pareto

Discrete Distributions

Binomial Distribution

- X = Number of heads observed in n coin tosses
- Parameters: n, θ
- $X \sim Bin(n, \theta)$
- Probability mass function (*pmf*)

$$Bin(k|n,\theta) \triangleq \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

Bernoulli Distribution

- Binomial distribution with $n = 1$
- Only one parameter ($\boldsymbol{\theta}$)

The pmf is nothing by the number of ways to choose k from a set of n multiplied by the probability of choosing k heads and rest $n-k$ tails.

Multinomial Distribution

- Simulates a K sided die
- Random variable $\mathbf{x} = (x_1, x_2, \ldots, x_K)$
- Parameters: n, θ
- $\theta \leftarrow \Re^K$
- θ_j - probability that j^{th} side shows up

$$Mu(\mathbf{x}|n,\boldsymbol{\theta}) \triangleq \binom{n}{x_1, x_2, \ldots, x_K}\prod_{j=1}^{K}\theta_j^{x_j}$$

Multinoulli Distribution

- Multinomial distribution with $n = 1$
- $\mathbf{x}$ is a vector of 0s and 1s with only one bit set to 1
- Only one parameter (θ)

$$\binom{n}{x_1, x_2, \ldots, x_K} = \frac{n!}{x_1!x_2!\ldots x_K!}$$

Continuous Distributions

Gaussian Distribution

$$\mathcal{N}(x|\mu,\sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- Parameters:

1. $\mu = \mathbb{E}[X]$
2. $\sigma^2 = var[X] = \mathbb{E}[(X-\mu)^2]$

- $X \sim \mathcal{N}(\mu, \sigma^2) \Leftrightarrow p(X = x) = \mathcal{N}(\mu, \sigma^2)$
- $X \sim \mathcal{N}(0, 1) \Leftarrow X$ is a **standard normal random variable**
- **Cumulative distribution function**:

$$\Phi(x; \mu, \sigma^2) \triangleq \int_{-\infty}^{x} \mathcal{N}(z|\mu, \sigma^2)dz$$

Gaussian distribution is the most widely used (and naturally occuring) distribution. The parameters μ is the mean and the mode for the distribution. If the variance σ^2 is reduced, the cdf for the Gaussian becomes more "spiky" around the mean and for limit $\sigma^2 \leftarrow 0$, the Gaussian becomes infinitely tall.

$$\lim_{\sigma^2 \leftarrow 0} \mathcal{N}(\mu, \sigma^2) = \delta(x - \mu)$$

where δ is the **Dirac delta function**:

$$\delta(x) = \begin{cases} \infty & \text{if } x = 0 \\ 0 & \text{if } x \neq 0 \end{cases}$$

6 Handling Multivariate Distributions

Joint Probability Distributions

- Multiple *related* random variables
- $p(x_1, x_2, \ldots, x_D)$ for $D > 1$ variables $(X_1, X_2, \ldots, X_D)$
- Discrete random variables?
- Continuous random variables?
- What do we measure?

Covariance

- How does X vary with respect to Y

- For linear relationship:

$$cov[X, Y] \triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

For discrete random variables, the joint probability distribution can be represented as a multi-dimensional array of size $O(K^D)$ where K is the number of possible values taken by each variable. This can be reduced by exploiting conditional independence, as we shall see when we cover *Bayesian networks*.

Joint distribution is trickier with continuous variables since each variable can take infinite values. In this case, we represent the joint distribution by assuming certain functional form.

Covariance and Correlation

- $\mathbf{x}$ is a d-dimensional random vector

$$cov[\mathbf{X}] \triangleq \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top]$$
$$= \begin{pmatrix} var[X_1] & cov[X_1, X_2] & \cdots & cov[X_1, X_d] \\ cov[X_2, X_1] & var[X_2] & \cdots & cov[X_2, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ cov[X_d, X_1] & cov[X_d, X_2] & \cdots & var[X_d] \end{pmatrix}$$

- Covariances can be between 0 and ∞
- Normalized covariance $\Rightarrow$ **Correlation**
- *Pearson* Correlation Coefficient

$$corr[X, Y] \triangleq \frac{cov[X, Y]}{\sqrt{var[X]var[Y]}}$$

- What is $corr[X, X]$?
- $-1 \leq corr[X, Y] \leq 1$
- When is $corr[X, Y] = 1$?
  - $Y = aX + b$

Multivariate Gaussian Distribution

- Most widely used joint probability distribution

$$\mathcal{N}(\mathbf{X}|\mu, \mathbf{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^{\top}\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu)\right]$$

References