Introduction to Machine Learning

Probability

I Foundations of Murphy book

Mingchen Gao

August 29, 2024

Outline

Contents

1 Introduction to Probability

- Probability that a coin will land heads is 50%[1]
- **What does this mean?**

Probability and statistics have a strong role in machine learning. For more details on probability theory refer to excellent textbooks on this topic [3, 1].

But how does one interpret probability? What does it mean that the probability of rain tomorrow is 70%? More importantly, what kind of action does one take based on this probabilistic knowledge? In statistics, there are two *schools of thought* which interpret probability in two different ways. We will discuss them next.

Frequentist Interpretations

- Number of times an event will be observed in n *trials*
- What if the event can only occur once?
  - *My winning the next month's powerball.*
  - *Polar ice caps melting by year 2050.*

Frequentists interpret the probability in terms of the outcome over multiple experiments in which the occurence of the event is monitored. So for the coin example, the frequentist interpretation of the 0.5 probability is that if we toss a coin many times, almost half of the times we will observe a head. A drawback of this interpretation is the reliance on multiple experiments. But consider a different scenario. If someone claims that the probability of the polar ice cap melting by year 2050 is 10%, then the frequentist interpretation breaks down, because this event can only happen zero or one times. Unless, there are multiple parallel universes!

Bayesian Interpretations

- **Uncertainty** of the event

[1] Dr. Persi Diaconis showed that a coin is 51% likely to land facing the same way up as it is started.

- Use for making decisions
    - What is the probability of an email is spam?

Bayesian interprets the same probability as a measure of *uncertainty* regarding the outcome of the event at any given time. Note that this view does not rely on multiple experiments. The Bayesian view allows one to use the probability to take future decisions through multiplicative priors.

While Bayesian statistics is favored more by Machine Learning community, many of the basic probability concepts are the same.

2 Random Variables

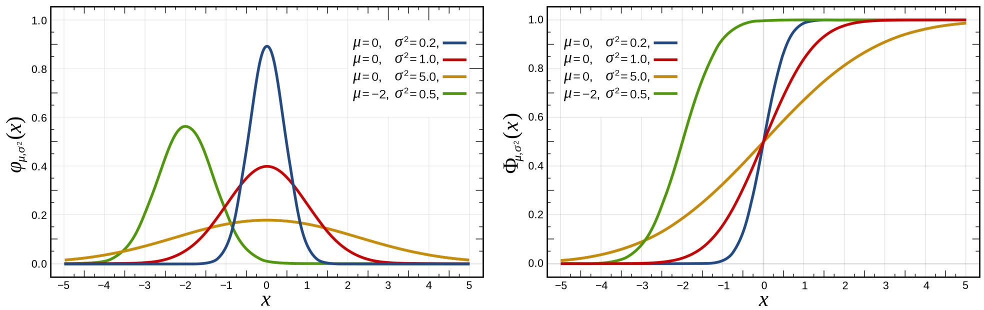
- X - random variable ($\mathbf{X}$ if multivariate)
- x - a specific value taken by the random variable (($\mathbf{x}$ if multivariate))
- Can take any value from $\mathcal{X}$
- **Discrete Random Variable** - $\mathcal{X}$ is finite/countably finite
- **Continuous Random Variable** - $\mathcal{X}$ is infinite
- $P(X = x)$ or $P(x)$ is the probability of X taking value x
- $p(x)$ is either the **probability mass function** (discrete) or **probability density function** (continuous) for the random variable X at 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**.

Discrete Examples

1. Coin toss ($\mathcal{X} = \{heads, tails\}$)
2. Six sided dice ($\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$)

Continuous Example: Gaussian distribution



Basic Rules

- For two events A and B:
  - **Union of two events**
    - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
  - **Joint Probability**
    - **product rule** $P(A, B) = P(A \wedge B) = P(A|B)P(B)$
    - **sum rule** Given $P(A, B)$ what is $P(A)$? Sum $P(A, B)$ over all values for B

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

- **Chain Rule of Probability**
  - Given D random variables, $\{X_1, X_2, \ldots, X_D\}$

$$P(X_{1:D}) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \ldots P(X_D|X_{1:D-1})$$

- **Conditional Probability**
  - $P(A|B) = \frac{P(A,B)}{P(B)}$

3 Bayes Rule

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

Bayes Theorem

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

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
  - $P(X = 1|Y = 0) = 0.1$

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

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

Classification Using Bayes Rule

- Given input example $\mathbf{X}$, Y is the random variable denoting the true class, we want to find the true class

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

- Assuming the **class-conditional** probability $P(\mathbf{X}|Y=c)$ and class prior $P(Y=c)$ are known
- Applying Bayes Rule

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

Independence and Conditional Independence

- One random variable does not depend on another
- $A \perp B \Longleftrightarrow P(A,B) = P(A)P(B)$
- 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**

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

A is conditionally independent of B given 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.

3.1 More About Conditional Independence

- Alice and Bob live in the same town but far away from each other
- Alice drives to work and Bob takes the bus
- Event A - Alice comes late to work
- Event B - Bob comes late to work
- Event C - A snow storm has hit the town
- $P(A|C)$ - Probability that Alice comes late to work given there is a snowstorm
- Now if I also know that Bob has come late to work, will it change the probability that Alice comes late to work? No, given the condition that there is a snow storm, the information of Bob has come late to work does not affect the probability of Alice comes late to work.
- What if I do not observe C? Will B have any impact on probability of A happening? The answer is yes, A and B are not completely independence.

4 Continuous Random Variables

- X is continuous
- Can take any value
- How does one define probability?
  - $\sum_x P(X = x) = 1$
- Probability that X lies in an interval $[a, b]$?
  - $P(a < X \leq b) = P(x \leq b) - P(x \leq a)$

- $F(q) = P(x \leq q)$ is the **cumulative distribution function**
- $P(a < X \leq b) = F(b) - F(a)$

Probability Density Function

$$p(x) = \frac{\partial}{\partial x} F(x)$$

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

- Can $p(x)$ be greater than 1?

$p(x)$ 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 1.

Expectation

- Expected value of a random variable

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

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

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

- For continuous variables

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

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

- 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
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- $\mathbb{E}[aX] = a\mathbb{E}[X]$
- Jensen's inequality: If $\varphi(X)$ is convex,
$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$$

Variance

- Spread of the distribution
$$\begin{aligned} var[X] &\triangleq \mathbb{E}((X - \mu)^2) \\ &= \mathbb{E}(X^2) - \mu^2 \end{aligned}$$
- Covariance $Cov[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

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

5 Different Types of Distributions

Discrete

- Binomial,*Bernoulli*
- Multinomial, *Multinolli*
- 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 of binomial is nothing but 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 tossing a K sided dice n times
- Random vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_K)$
- Parameters: $n, \boldsymbol{\theta} \leftarrow \Re^K$, θ_j - probability that j^{th} side shows up
- $\boldsymbol{x} \sim Mu(n, \boldsymbol{\theta})$

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

Multinoulli Distribution

- Multinomial distribution with $n = 1$
- $\mathbf{x}$ is a vector of 0s and 1s with only one bit set to 1, called **one-hot vector**
- 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. $\sigma^2 = var[X] = \mathbb{E}[(X - \mu)^2]$

- $X \sim \mathcal{N}(0, 1) \Leftarrow X$ 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: multi-dimensional array of size $O(K^D)$

- Continuous random variables: certain functional form

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

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

- 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

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

7 Transformations of Random Variables

Linear Transformations

- What is the distribution of $f(\mathbf{X})$ ($\mathbf{X} \sim p()$)?
  - Linear transformation:

$Y = \mathbf{a}^\top \mathbf{X} + b$

- $\mathbb{E}[Y] = a^\top \mu + b$
- $var[Y] = \mathbf{a}^\top \mathbf{\Sigma} \mathbf{a}$

$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$

- $\mathbb{E}[Y] = \mathbf{A}\mu + \mathbf{b}$
- $cov(Y) = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top$

- The Matrix Cookbook [2]
- `http://orion.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`
- Available on Piazza

A linear transformation of a random variable is also a random variable. Random variable $Y = \mathbf{a}^\top \mathbf{X} + b$ is a scalar variable, while the random variable $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$ is a vector variable ($\mathbf{A}$ is a matrix).

- $\mathbb{E}[\mathbf{a}^\top \mathbf{X} + b] = \mathbf{a}^\top \mu + b$

$$
\begin{aligned}
\mathbb{E}[\mathbf{a}^\top \mathbf{X} + b] &= \mathbb{E}[\mathbf{a}^\top \mathbf{X}] + b \\
&= \mathbf{a}^\top \mathbb{E}[\mathbf{X}] + b \\
&= \mathbf{a}^\top \mu + b
\end{aligned}
$$

- $var[\mathbf{a}^\top \mathbf{X} + b] = \mathbf{a}^\top \mathbf{\Sigma} \mathbf{a}$
- $\mathbb{E}[Y] = \mathbb{E}(\mathbf{A}\mathbf{X} + \mathbf{b}) = \mathbf{A}\mu + \mathbf{b}$
- $cov(\mathbf{A}\mathbf{X} + \mathbf{b}) = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top$

General Transformations

- $f()$ is **not linear**
- Example: X - discrete

$$Y = f(X) = \begin{cases} 1 & \text{if } X \text{ is even} \\ 0 & \text{if } X \text{ is odd} \end{cases}$$

Approximate Methods

- Generate N samples from distribution for X
- For each sample, $x_i, i \in [1, N]$, compute $f(x_i)$
- Use empirical distribution as *approximate* true distribution

Approximate Expectation

$$\mathbb{E}[f(X)] = \int f(x)p(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

Computing the distribution of a function of a random variable is not alway s analytically possible. Monte Carlo approximation allows a *rough* estimation of the distribution through sampling of the variable and applying the function on the samples.

Obviously, this approach does not let us compute the pdf of $f(X)$ but it allows computing approximate statistics for $f(X)$, such as the mean, variance, etc.

- $\mathbb{E}[Y] \approx \bar{y} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$
- $var[Y] \approx \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - \bar{y})^2$
- $P(Y \leq y) = \frac{1}{N}\#\{f(x_i) \leq y\}$

8 Information Theory - Introduction

- Quantifying uncertainty of a random variable

Entropy

- $\mathbb{H}(X)$ or $\mathbb{H}(p)$

$$\mathbb{H}(X) \triangleq -\sum_{k=1}^{K} p(X=k) \log_2 p(X=k)$$

Let us consider a discrete random variable X that takes 2 values with equal probability (0.5). In this case the entropy of X will be:

$$\mathbb{H}(X) = -(0.5 \log_2(0.5) + 0.5 \log_2(0.5)) = 1$$

In another example, let X take first value with 0.9 probability and the second value with 0.1 probability.

$$\mathbb{H}(X) = -(0.9 \log_2(0.9) + 0.1 \log_2(0.1)) = 0.4690$$

- Variable with maximum entropy: Uniform distribution
- Lowest entropy: Zero

A uniformly distributed discrete variable has the highest entropy since every possibility is equally likely. A *delta-function* which puts all mass on one possibility has the lowest (or 0) uncertainty.

KL Divergence

- **Kullback-Leibler Divergence** (or *KL Divergence* or *relative entropy*): Measuring the dissimilarity of two probability distributions

$$\begin{aligned}
\mathbb{KL}(p||q) &\triangleq \sum_{k=1}^{K} p(k) \log \frac{p_k}{q_k} \\
&= \sum_{k} p(k) \log p(k) - \sum_{k} p(k) \log q(k) \\
&= -\mathbb{H}(p) + \mathbb{H}(p, q)
\end{aligned}$$

- $\mathbb{H}(p, q)$ is the *cross-entropy*
- KL-divergence is asymmetric
- *Important fact:* $\mathbb{H}(p, q) \geq \mathbb{H}(p)$

Cross-entropy is the average number of bits needed to encode data coming from a source with distribution p when use distribution q to define our codebook.

Mutual Information

- What does learning about one variable X tell us about another, Y?
  - Correlation?

Mutual Information

$$\mathbb{I}(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

- $\mathbb{I}(X;Y) = \mathbb{I}(Y;X)$
- $\mathbb{I}(X;Y) \geq 0$, equality iff $X \perp Y$

Correlation only captures linear relationships. Mutual information calculates how similar the joint distribution is to the factored distribution. The mutual information is zero iff the variables are independent.

References

[1] E. Jaynes and G. Bretthorst. *Probability Theory: The Logic of Science.* Cambridge University Press Cambridge:, 2003.

[2] K. B. Petersen and M. S. Pedersen. The matrix cookbook, nov 2012. Version 20121115.

[3] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference (Springer Texts in Statistics).* Springer, Oct. 2004.