Probability Theory Review

STATS 415: Data Mining and Machine Learning

University of Michigan

Fall 2022

Outline

Elements of Probability

Random Variables

Definition of probability space

- **Sample space** Ω: The set of all the outcomes of a random experiment.
- **Event space** $\mathcal{F}$: A set whose elements $A \in \mathcal{F}$ (called events) are subsets of Ω (i.e., $A \subseteq \Omega$).
- **Probability measure**: A function $P : \mathcal{F} \to \mathbf{R}$ that satisfies the following properties:
  - Non-negativity: $P(A) \geq 0$, for all $A \in \mathcal{F}$
  - Completeness: $P(\Omega) = 1$
  - Countable Additivity: If $A_1, A_2, \ldots$ are disjoint events (i.e., $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P\left(A_i\right)$$

Properties of probability

- If $A \subseteq B \implies P(A) \leq P(B)$.
- $P(A \cap B) \leq \min(P(A), P(B))$
- $P\left(A^c\right) \triangleq P(\Omega \backslash A) = 1 - P(A)$
- $P(A \cup B) \leq P(A) + P(B)$ This property is known as the union bound.
- If $A_1, \ldots, A_k$ are a set of disjoint events such that $\bigcup_{i=1}^k A_i = \Omega$, then $\sum_{i=1}^k P\left(A_k\right) = 1.$ This property is known as the Law of Total Probability.

Conditional probability and independence

- Conditional probability:

$$P(A \mid B) \triangleq \frac{P(A \cap B)}{P(B)}$$

- Independence: Two events are called independent if and only if $P(A \cap B) = P(A) * P(B)$

- Mutually Independence: In general we say that $A_1, \ldots, A_k$ are mutually independent if for any subset $S \subseteq \{1, 2, \ldots, k\}$, we have

$$P\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} P\left(A_i\right)$$

Law of total probability and Bayes' theorem

- **Law of total probability**: Theorem. Suppose $A_1, \ldots, A_n$ are disjoint events, and event B satisfies $B \subseteq \bigcup_{i=1}^{n} A_i$, then

$$P(B) = \sum_{i=1}^{n} P\left(A_i\right) P\left(B \mid A_i\right)$$

- **Bayes' theorem**: Theorem. Suppose $A_1, \ldots, A_n$ are disjoint events, and event B satisfies $B \subset \bigcup_{i=1}^{n} A_i$. Then if $P(B) > 0$, it is the case that

$$P\left(A_j \mid B\right) = \frac{P\left(A_j\right) P\left(B \mid A_j\right)}{\sum_{i=1}^{n} P\left(A_i\right) P\left(B \mid A_i\right)}.$$

Outline

Definition and examples

- **Random variable**: a random variable X is a function

$$X : \Omega \longrightarrow \mathbf{R}$$

Such that for all "nice" subsets $A \subseteq \mathbf{R}$ we have

$$\{\omega \in \Omega | X(\omega) \in A\} \in \mathcal{F}$$

In words, we can calculate the probability that the random variable X is on the subset A.

Definition and examples

- **Ex 1.** Consider an experiment in which we flip 10. coins, and we want to know the number of coins that come up heads

  we might have $\omega_0 = \langle H, H, T, H, T, H, H, T, T, T \rangle \in \Omega$

  In our experiment above, suppose that $X(\omega)$ is the number of heads which occur in the sequence of tosses ω. Then $X(\omega_0) = 5$. Note that $X(\omega_0)$ can take only a finite (Countable) number of values 0,1,...,10, so it is known as a **discrete random variable**. Here, the probability of the set associated with a random variable X taking on some specific value k is:

$$P(X = k) := P(\{\omega : X(\omega) = k\})$$

Definition and examples

- **Ex 2.** Suppose that $X(\omega)$ is a random variable indicating the amount of time it takes for a radioactive particle to decay. In this case, $X(\omega)$ takes on a infinite (Uncountable) number of possible values, so it is called a **continuous random variable**. In this case we are interested in the probability of intervals of time.

$$P(a \leq X \leq b) := P(\{\omega : a \leq X(\omega) \leq b\})$$

Cumulative distribution functions

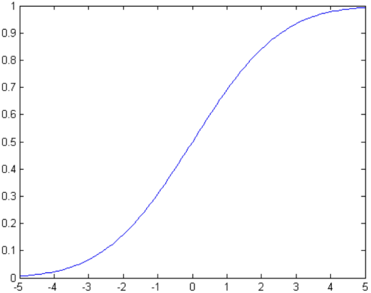
A cumulative distribution function (CDF) is a function $F_X : \mathbf{R} \to [0,1]$ which specifies a probability measure as,

$$F_X(x) \triangleq P(X \leq x).$$

By using this function one can calculate the probability of any event in $\mathcal{F}$. [3] Figure 1 shows a sample CDF function. A CDF function satisfies the following properties.

- $0 \leq F_X(x) \leq 1$
- $\lim_{x\to-\infty} F_X(x) = 0$.
- $\lim_{x\to\infty} F_X(x) = 1$.
- $x \leq y \implies F_X(x) \leq F_X(y)$.

Cumulative distribution functions



Probability mass functions

If X is a **discrete random variable**, we can define the **Probability mass function** $p_X : \Omega \to \mathbf{R}$:

$$p_X(x) \triangleq P(X = x)$$

A PMF function satisfies the following properties.

- $0 \leq p_X(x) \leq 1$.
- $\sum_{x \in \mathrm{Val}(X)} p_X(x) = 1$.
- $\sum_{x \in A} p_X(x) = P(X \in A)$.

Probability density functions

For some continuous random variables, the cumulative distribution function $F_X(x)$ is differentiable everywhere. In these cases, we define the Probability Density Function (PDF) as the derivative of the CDF, i.e.,

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}$$

We can interpret $f_X(x)$ as $P(x \leq X \leq x + \Delta x) \approx f_X(x)\Delta x$
A PDF function satisfies the following properties.

- $f_X(x) \geq 0$.
- $\int_{-\infty}^{\infty} f_X(x) = 1$.
- $\int_{x \in A} f_X(x)dx = P(X \in A)$.

Expectation

Suppose that $g : \mathbf{R} \longrightarrow \mathbf{R}$ is an arbitrary function. We define the expectation or expected value of $g(X)$ as

- **discrete random variable**

$$E[g(X)] \triangleq \sum_{x \in \mathrm{Val}(X)} g(x) p_X(x)$$

- **continuous random variable**

$$E[g(X)] \triangleq \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Expectation

Expectation satisfies the following properties:

- $E[a] = a$ for any constant $a \in \mathbf{R}$.
- $E[af(X)] = aE[f(X)]$ for any constant $a \in \mathbf{R}$.
- $E[f(X) + g(X)] = E[f(X)] + E[g(X)]$. This property is known as the linearity of expectation.
- $E[1_{\{X \in A\}}] = P(X \in A)$.

Variance

The variance of a random variable X is a measure of how concentrated the distribution of a random variable X is around its mean. Formally, the variance of a random variable X is defined as

$$\mathrm{Var}[X] \triangleq E\left[(X - E(X))^2\right]$$

We note the following properties of the variance.

- $\mathrm{Var}[a] = 0$ for any constant $a \in \mathbf{R}$.
- $\mathrm{Var}[af(X)] = a^2\,\mathrm{Var}[f(X)]$ for any constant $a \in \mathbf{R}$.

Some common distributions

- Bernoulli
- Binomial
- Geometric
- Poisson
- Uniform
- Exponential
- Normal