36-700: Probability and Mathematical Statistics **Spring 2019**

Lecture 1: Basics of Probability

Lecturer: Jing Lei

1.1 Introduction

The first part of our adventure is a highly selective review of probability theory, focusing especially on things that are most useful in statistics.

1.1.1 Sample spaces and events

A typical way to go about defining things is to suppose that we conduct an experiment. An experiment is a measurement of a random (stochastic) process.

Our measurements take values in some set Ω: this is the sample space. The sample space effectively defines all possible outcomes of our measurement.

Examples:

- Suppose I toss a coin: in this case the sample space $\Omega = \{H, T\}$.
- If I measure the reaction time to some stimulus the sample space $\Omega = (0, \infty)$.
- Suppose I toss a coin twice: what is the sample space?

An event is some subset of $A \subseteq \Omega$, i.e., it is a subset of possible outcomes of our experiment. We say that an event A occurs if the outcome of our experiment lies in the set A.

1.1.2 A quick aside on set-theoretic notation

Here is some basic notation for set operations.

- (Subset) $A \subseteq B$ means that all elements in A are also in B.
- (Complement) A^c: elements that are not in A.
- (Empty set) $\Omega^c = \emptyset$.
- (Union) $A \cup B$: elements that are either in A or in B, or both.

- (Intersection) $A \cap B$: elements that are both in A and B. Sometimes we use AB for brevity.
- (Set difference) $A \backslash B = A \cap (B^c)$: elements that are in A but not in B.
- (Symmetric difference) $A \triangle B = (A \backslash B) \cup (B \backslash A)$: elements that are in A or B, but not both.
- (Cardinality) $|A|$ denotes the number of elements in A.

Exercise: Prove that $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$.

1.1.3 Probability

A probability distribution is a mapping from events to real numbers that satisfies certain axioms. We denote this mapping by $\mathbb{P} : A \mapsto \mathbb{R}$. The axioms are:

1. **Non-negativity:** $\mathbb{P}(A) \geq 0, \quad \forall\ A \subseteq \Omega$.
2. **Unity of Ω:** $\mathbb{P}(\Omega) = 1$.
3. **Countable additivity:** For a collection $A_1, A_2, \ldots,$ of disjoint sets we must have that,

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

We can use these axioms to show several useful and intuitive properties of probability distributions:

- $\mathbb{P}(\emptyset) = 0$.
- $A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$.
- $0 \leq \mathbb{P}(A) \leq 1$.
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

All of these properties can be understood via a Venn diagram.

Example: Suppose I toss a fair coin twice, and denote by H_1 the event that the first coin lands heads, and H_2 the event that the second coin lands heads. Calculate $\mathbb{P}(H_1 \cup H_2)$.

We can use the above formula:

$$\begin{aligned}\mathbb{P}(H_1 \cup H_2) &= \mathbb{P}(H_1) + \mathbb{P}(H_2) - \mathbb{P}(H_1 \cap H_2) \\ &= 0.5 + 0.5 - 0.25 = 0.75.\end{aligned}$$

Exercise: Derive an analogous formula for $\mathbb{P}(\cup_{i=1}^n A_i)$.

1.1.4 Counting and the uniform distribution on discrete sets

Suppose we toss a die twice. There are 36 possible outcomes: $\Omega = \{(t_1, t_2) : t_1, t_2 = 1, 2, 3, 4, 5, 6\}$. If the die is fair then each outcome is equally likely. This is an example of a uniform distribution on a discrete set.

The general rule of calculating probability of an event under the uniform distribution in finite sample space is

$$\boxed{\mathbb{P}(A) = \frac{|A|}{|\Omega|}.}$$

For example, let A be the event that the sum of two tosses being less than five. Then $A = \{(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)\}$. Thus $\mathbb{P}(A) = 6/36 = 1/6$.

Example: There are two black balls and three white balls in a bag. Two balls are randomly drawn, without replacement, from the bag. What is the probability of the two balls have different color? What is the probability if the balls are drawn with replacement?

When drawing without replacement, the sample space Ω has cardinality $|\Omega| = 5 \times 4 = 20$, and the event A has cardinality $|A| = 2 \times 3 + 3 \times 2 = 12$ (first white and second black, or first black second white). So $\mathbb{P}(A) = 12/20 = 0.6$.

When drawing with replacement, the sample space Ω has cardinality $5 \times 5 = 25$, and A still has cardinality 12. Then $\mathbb{P}(A) = 12/25 = 0.48$.

More generally, calculating probabilities under the uniform distribution on a discrete set is based on counting.

Even more generally, probabilities under non-uniform distributions can be calculated by adding the probabilities of the individual singleton events.

Exercise: Suppose we toss an unfair die twice, where the number "1" has chance 2/7 while numbers "2" through "6" are equally likely, each with probability 1/7. What's the probability that the sum of two tosses being less than five?

1.2 Conditional probability

Definition 1.1 (Conditional probability) *If $\mathbb{P}(B) > 0$, the conditional probability of A given B is*

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}\,. \tag{1.1}$$

Example: Consider tossing a fair die. Let A be the event that the result is an odd number, and $B = \{1, 2, 3\}$. Then $P(A|B) = 2/3$, and $P(A) = 1/2$. In this example A and B are not independent.

Remark: In general $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$.

The chain rule: A simple re-writing of the above expression yields the so-called chain rule:

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

More generally,

$$\mathbb{P}(A_1 \cap A_2 \cap A_3 \ldots) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_2, A_1)\ldots.$$

1.3 Independence of Events

Independence roughly asks the question of whether one event provides any information about another. For example, if we toss a fair coin twice, let H_i be the event that the ith toss is head ($i = 1, 2$), then intuitively knowing if the event H_1 occurred or not does not provide any information about H_2. The formal definition of independence is

Definition 1.2 (Independence) *Two events A and B are called* independent *if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)\,. \tag{1.2}$$

A set of events A_j ($j \in I$) are called mutually independent *if*

$$\mathbb{P}\left(\bigcap_{j\in J} A_j\right) = \prod_{j\in J} \mathbb{P}(A_j)\,,$$

for any finite subset J of I.

Conditional probability gives another interpretation of independence: A and B are independent if the unconditional probability is the same as the conditional probability.

Example: We can formally verify the coin toss example: $P(A_1) = P(A_2) = 1/2$, and $P(A_1A_2) = 1/4$.

Example: To see a less obvious example, consider tossing a fair die once. Let $A = \{1, 2, 3, 4\}$, $B = \{2, 4, 6\}$. Then $AB = \{2, 4\}$. $\mathbb{P}(A) = 2/3$, $\mathbb{P}(B) = 1/2$, and $\mathbb{P}(AB) = 1/3 = \mathbb{P}(A)\mathbb{P}(B)$. Then we conclude that A and B are independent.

Here are some simple facts about independence.

1. Ω is independent of any other event. The same holds for $\emptyset$.
2. If A, B are disjoint, both having positive probability, then A and B cannot be independent.
3. If A and B are independent, then A^c and B are also independent.

When combined with other properties of probability, independence can sometimes allow very easy calculation of the probability of certain events.

Example: Consider tossing a fair coin, what is the probability of at least one head in the first 10 tosses?

Let A be the event of at least one head in 10 tosses. Then A^c is the event of no heads in 10 tosses, or equivalently, all 10 tosses being tail. Therefore $\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - 2^{-10}$.

1.4 Bayes' Rule

Roughly Bayes rule allows us to calculate the probability of $B|A$ from the probability of $A|B$. As a preliminary we need the following:

Theorem 1.3 (Law of total probability) *Let A_1, ..., A_k be a partition of Ω. The for any B,*

$$\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i).$$

Proof: The claim follows by observing that $A_i \cap B$ $(i = 1, ..., k)$ forms a partition of B, and $\mathbb{P}(A_i \cap B) = \mathbb{P}(B|A_i)\mathbb{P}(A_i)$. ∎

The law of total probability is a combination of additivity and conditional probability. It leads to the very useful Bayes' theorem.

Theorem 1.4 (Bayes' Rule) *Let A_1, ..., A_k be a partition of Ω. Then*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}.$$

The proof is simple. The numerator is just $\mathbb{P}(A_iB)$ and the denominator is $\mathbb{P}(B)$.

This is useful when $\mathbb{P}(A_i|B)$ is not obvious to calculate but $\mathbb{P}(B|A_i)$ and $\mathbb{P}(A_i)$ are easy to find. A typical application is classification.

Example: Suppose there are three types of emails: A_1 ="spam", A_2 ="low priority", A_3="high priority". Based on previous experience, $\mathbb{P}(A_1) = 0.85$, $\mathbb{P}(A_2) = 0.1$, $\mathbb{P}(A_3) = 0.05$. Let B be the event that an email contains the word "free", then $\mathbb{P}(B|A_1) = 0.9$, $\mathbb{P}(B|A_2) = 0.1$, $\mathbb{P}(B|A_3) = 0.1$. Now a new coming email contains the word "free", what is the probability that it is spam?

Answer:

$$\begin{aligned}\mathbb{P}(A_1|B) =& \frac{\mathbb{P}(B|A_1)\mathbb{P}(A_1)}{\mathbb{P}(B|A_1)\mathbb{P}(A_1) + \mathbb{P}(B|A_2)\mathbb{P}(A_2) + \mathbb{P}(B|A_3)\mathbb{P}(A_3)} \\ =& \frac{0.85 \times 0.9}{0.85 \times 0.9 + 0.1 \times 0.1 + 0.05 \times 0.1} \approx 0.98\,.\end{aligned}$$