EECS126 Course Notes

Anmol Parande

Spring 2021 - Professor Thomas Courtade

Disclaimer: These notes reflect 126 when I took the course (Spring 2021). They may not accurately reflect current course content, so use at your own risk. If you find any typos, errors, etc, please raise an issue on the GitHub repository.

Contents

1 Introduction to Probability

Definition 1 *A probability space is a triple $(\Omega, \mathcal{F}, P)$ where Ω is a set of objects called the sample space, $\mathcal{F}$ is a family of subsets of Ω called events, and the probability measure $P : \mathcal{F} \to [0, 1]$.*

One key assumption we make is that $\mathcal{F}$ is a σ-algebra containing Ω, meaning that countably many complements, unions, and intersections of events in $\mathcal{F}$ are also events in $\mathcal{F}$. The probability measure P must obey **Kolmogorov's Axioms**.

1. $\forall A \in \mathcal{F},\ P(A) \geq 0$
2. $P(\Omega) = 1$
3. If $A_1, A_2, \dots \in \mathcal{F}$ and $\forall i \neq j,\ A_i \bigcap A_j = \emptyset$, then $P\left(\bigcup_{i \geq 1} A_i\right) = \sum_{i \geq 1} P(A_i)$

We choose Ω and $\mathcal{F}$ to model problems in a way that makes our calculations easy.

Theorem 1

$$P(A^c) = 1 - P(A)$$

Theorem 2 (Inclusion-Exclusion Principle)

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \left(\sum_{1 \leq i_1 < \cdots < i_k \leq n} P(A_{i_1} \cap \cdots \cap A_{i_k}) \right)$$

Theorem 3 (Law of Total Probability) *If $A_1, A_2, \cdots$ partition Ω (i.e A_i are disjoint and $\cup A_i = \Omega$), then for event B,*

$$P(B) = \sum_i P(B \cap A_i)$$

1.1 Conditional Probability

Definition 2 *If B is an event with $P(B) > 0$, then the conditional probability of A given B is*

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

Intuitively, conditional probabilty is the probability of event A given that event B has occurred. In terms of probability spaces, it is as if we have taken $(\Omega, \mathcal{F}, P)$ and now have a probabilty measure $P(\cdot|C)$ belonging to the space $(\Omega, \mathcal{F}, P(\cdot|C))$.

Theorem 4 (Bayes Theorem)

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

1.2 Independence

Definition 3 *Events A and B are independent if $P(A \cap B) = P(A)P(B)$*

If $P(B) > 0$, then A, B are independent if and only if $P(A|B) = P(A)$. In other words, knowing B occurred gave no extra information about A.

Definition 4 *If A, B, C with $P(C) > 0$ satisfy $P(A \cap B|C) = P(A|C)P(B|C)$, then A and B are conditionally independent given C.*

Conditional independence is a special case of independence where A and B are not necessarily independent in the original probability space which has the measure P, but are independent in the new probability space conditioned on C with the measure $P(\cdot|C)$.

2 Discrete Probability

Definition 5 *A random variable is a function $X : \Omega \to \mathbb{R}$ with the property $\forall \alpha \in \mathbb{R},\ \{\omega \in \Omega :\ X(\omega) \leq \alpha\} \in \mathcal{F}$.*

The condition in definition 5 is necessary to compute $P(X \leq \alpha),\ \forall \alpha \in \mathbb{R}$. This requirement also let us compute $P(X \in B)$ for most sets by leveraging the fact that $\mathcal{F}$ is closed under complements, unions, and intersections. For example, we can also compute $P(X > \alpha)$ and $P(\alpha < X \leq \beta)$. In this sense, the property binds the probability space to the random variable.

definition 5 also implies that random variables satisfy particular algebraic properties. For example, if X, Y are random variables, then so are $X+Y$, XY, X^p, $\lim_{n\to\infty} X_n$, etc.

Definition 6 *A discrete random variable is a random variable whose codomain is countable.*

Definition 7 *The probability mass function (or distribution) of a random variable X is the frequency with which X takes on different values.*

$$p_X : \mathcal{X} \to [0,1] \text{ where } \mathcal{X} = \mathit{range}(X), \qquad p_X(x) = \mathit{Pr}\{X = x\}.$$

Note that $\sum_{x\in\mathcal{X}} p_X(x) = 1$ since $\bigcap_{x\in\mathcal{X}}\{w : X(w) = x\} = \Omega$.

Definition 8 *If X and Y are random variables on a common probability space $(\Omega, \mathcal{F}, P)$, then the joint pmf describes the frequencies of joint outcomes.*

$$p_{XY}(x,y) = \mathit{Pr}\{X = x, Y = y\}$$

Definition 9 *The marginal distribution of a joint PMF is the PMF is the distribution of a single random variable.*

$$p_X(x) = \sum_y p_{XY}(x, Y = y)$$

Although random variables are defined based on a probability space, it is often most natural to model problems without explicitly specifying the probability space. This works so long as we specify the random variables and their distribution in a "consistent" way. This is formalized by the so-called Kolmogorov Extension Theorem but can largely be ignored.

Definition 10 *Two random variables X and Y are independent if $p_{XY}(x,y) = p_X(x)p_Y(y)$.*

Just like independent, we can extend the notion of conditional probability to random variables.

Definition 11 *For a discrete random variable, the conditional PMF is given by*

$$p_{X|Y}(x,y) = \frac{p_{XY}(x,y)}{p_Y(y)} = \frac{P(\{X = x\} \cap \{Y = y\})}{P(\{Y = y\})} = P(X = x | Y = y).$$

The interpretation is the same: given the value of random variable Y, what is the distribution of X.

2.1 Properties of Discrete Random Variables

2.1.1 Expectation

Definition 12 *The expectation of a discrete random variable describes the center of a distribution and is given by*

$$\mathbb{E}\left[X\right] = \sum_{x \in \mathcal{X}} x p_X(x)$$

provided the series exists.

Expectation has several useful properties. If we want to compute the expectation of a function of a random variable, then we can use the law of the unconscious statisitician.

Theorem 5 (Law of the Unconscious Statistician)

$$\mathbb{E}\left[g(X)\right] = \sum_{x \in \mathcal{X}} g(x) p_X(x)$$

Another useful property is its linearity.

$$\mathbb{E}\left[aX + bY\right] = a\mathbb{E}\left[X\right] + b\mathbb{E}\left[Y\right], \ \forall a, b \in \mathbb{R}.$$

For expectations where it is complicated to compute $p_X(x)$, we can use the tail-sum formula.

Theorem 6 (Tail Sum) *For a non-negative integer random variable,*

$$\mathbb{E}\left[X\right] = \sum_{k=1}^{\infty} Pr\left\{X \geq k\right\}.$$

When two random variables are independent, expectation has some additional properties.

Theorem 7 *If X and Y are independent, then*

$$\mathbb{E}\left[XY\right] = \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right].$$

2.1.2 Variance

Definition 13 *The variance of a discrete random variable X describes its spread around the expectation and is given by*

$$Var\left(X\right) = \mathbb{E}\left[(X_{\mathbb{E}}\left[X\right])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2.$$

Theorem 8 *When two random variables X and Y are independent, then*

$$Var\left(X+Y\right) = Var\left(X\right) + Var\left(Y\right).$$

2.1.3 Covariance and Correlation

Definition 14 *The covariance of two random variables describes how much they depend on each other and is given by*

$$Cov\left(X,Y\right) = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])(Y - \mathbb{E}\left[Y\right])\right].$$

If $Cov\left(X,Y\right) = 0$ then X and Y are uncorrelated.

Definition 15 *The correlation coefficient gives a single number which describes how random variables are correlated.*

$$\rho(X,Y) = \frac{Cov\left(X,Y\right)}{\sqrt{Var\left(X\right)}\sqrt{Var\left(Y\right)}}.$$

Note that $-1 \leq \rho \leq 1$.

2.2 Common Discrete Distributions

Definition 16 *X is uniformly distributed when each value of X has equal probability.*

$$X \sim \textit{Uniform}(\{1,2,\cdots,n\}) \implies p_X(x) = \begin{cases} \frac{1}{n} & x = 1,2,\cdots,n, \\ 0 & \textit{else.} \end{cases}$$

Definition 17 *X is a Bernoulli random variable if it is either 0 or 1 with $p_X(1) = p$.*

$$X \sim \textit{Bernoulli}(p) \implies p_X(x) = \begin{cases} 1-p & x=0, \\ p & x=1, \\ 0 & \textit{else.} \end{cases}$$

$$\mathbb{E}\left[X\right] = p \qquad Var\left(X\right) = (1-p)p$$

Bernoulli random variables are good for modeling things like a coin flip where there is a probability of success. Bernoulli random variables are frequently used as indicator random variables $\mathbb{1}_A$ where

$$\mathbb{1}_A = \begin{cases} 1 & \text{if A occurs,} \\ 0 & \text{else.} \end{cases}$$

When paired with the linearity of expectation, this can be a powerful method of computing the expectation of something.

Definition 18 *X is a Binomial random variable when*

$$X \sim \textit{Binomial}(n,p) \implies p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, \cdots, n \\ 0 & \textit{else.} \end{cases}$$

$$\mathbb{E}\left[X\right] = np \qquad Var\left(X\right) = np(1-p)$$

A binomial random variable can be thought of as the number of successes in n trials. In other words,

$$X \sim \text{Binomial}(n,p) \implies X = \sum_{i=1}^{n} X_i, \quad X_i \sim \text{Bernoulli}(p).$$

By construction, if $X \sim \text{Binomial}(n,p)$ and $Y \sim \text{Binomial}(m,p)$ are independent, then $X + Y \sim \text{Binomial}(m+n,p)$.

Definition 19 *A Geometric random variable is distributed as*

$$X \sim \textit{Geom}(p) \implies p_X(x) = \begin{cases} p(1-p)^{x-1} & x = 1, 2, \cdots \\ 0 & \textit{else.} \end{cases}$$

$$\mathbb{E}\left[X\right] = \frac{1}{p} \qquad Var\left(X\right) = \frac{1-p}{p^2}$$

Geometric random variables are useful for modeling the number of trials required before the first success. In other words,

$$X \sim \text{Geom}(p) \implies X = \min\{k \geq 1 : X_k = 1\} \text{ where } X_i \sim \text{Bernoulli}(p).$$

A useful property of geometric random variables is that they are memoryless:

$$\Pr\{X = K + M | X > k\} = \Pr\{X = M\}.$$

Definition 20 *A Poisson random variable is distributed as*

$$X \sim Poisson(\lambda) \implies p_X(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & x = 0, 1, \cdots \\ 0 & \textit{else.} \end{cases}$$

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

Poisson random variables are good for modeling the number of arrivals in a given interval. Suppose you take a given time interval and divide it into n chunks where the probability of arrival in chunk i is $X_i \sim \text{Bernoulli}(p_n)$. Then the total number of arrivals $X_n = \sum_{i=1}^n X_i$ is distributed as a Binomial random variable with expectation $np_n = \lambda$. As we increase n to infinity but keep λ fixed, we arrive at the poisson distribution.

A useful fact about Poisson random variables is that if $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ are independent, then $X + Y \sim \text{Poisson}(\lambda + \mu)$.