EECS 126

Albert Ye

October 10, 2023

1 Probability Space

1.1 Definition

Essentially from 70. Events happen with some probability in a larger probability space that contains all events that can happen.

1.2 Axioms of Probability

Proposition 1 (Axioms) 1. (Positivity) $P(\omega > 0)$ for any event ω in probability space Ω.

2. (Totality) In any sample space Ω, $P(\Omega) = 1$.
3. (Additivity) If $A_1, A_2, \ldots, A_n$ are independent, then

$$\sum_{i=1}^{n} A_i = \bigcup_{i=1}^{n} A_i.$$

From just this, we can get some useful information, such as the union bound.

Theorem 2 (Union Bound)

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

The proof is left as an exercise to the student, probably in the homework.

1.3 σ-algebra

Definition 3 (σ-algebra)
Given a sample space Ω, a set $\mathcal{F} \subseteq 2^{\Omega}$ is a σ-algebra if:

1. $\Omega \in \mathcal{F}$
2. If any event A is in $\mathcal{F}$, then its complement $\Omega \setminus A$ is also in $\mathcal{F}$.
3. For countably many events $A_1, A_2, \ldots, A_n, \ldots \in \mathcal{F}$, their union $A = \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

The biggest note is that Ω must be in a σ-algebra in order for any of the axioms of probability to apply.

2 Conditional Probability

2.1 Definition

2.2 Total Probability

2.3 Bayes' Rule

2.4 Continuous Bayes

3 It Depends

3.1 Independence / (Un)correlation

3.2 Conditional Expectation

Notice that $E[X|Y]$ is a random variable, but $E[X|Y = y]$ is a number. We can call $E[X|Y]$ a function $g(Y)$, where then $E[X|Y = y] = g(y)$ is just a value in the function.

3.3 Iterated Expectation

4 Distributions

4.1 Joint Distribution

Definition 4 (Joint Distribution)

A joint distribution $f_{X,Y}(x, y)$

4.2 Marginal Distribution

4.3 Derived Distribution

5 Random Variables

5.1 Discrete

5.1.1 Bernoulli

- PMF: $p_X(k) = \begin{cases} p & k = 1 \\ 1 = p & k = 0 \end{cases}$
- Expected value: p
- Variance: $p(1-p)$.

5.1.2 Binomial

- PMF: $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$ over all $k \in 0, 1, \ldots, n$.
- Expected value: np
- Variance: $np(1-p)$.

Run a Bernoulli test n times, find how many are positive.

5.1.3 Geometric

- PMF: $p_X(k) = (1-p)^{k-1}p$, for $k = 1, 2, \ldots$.
- Expected value: $\frac{1}{p}$
- Variance: $\frac{1-p}{p^2}$.

Here, each trial has a p probability of success, and we want to find the # of trials until one success.

5.1.4 Poisson

- PMF: $p_X(k) = \frac{\lambda^k (e^{-\lambda})}{k!}$.
- Expected value: λ
- Variance: λ

Used to simulate arrivals, I guess. More useful later, with Poisson processes.

5.2 Continuous

5.2.1 Uniform

5.2.2 Exponential

5.2.3 Gaussian

5.2.4 Joint Gaussian

The main tips for Joint Gaussian are to approach it as a sort of vectorized Gaussians over a certain number N of dimensions. Most of the addition / whatever operations in a Gaussian can be remodeled as a Joint Gaussian.

6 Moment Generating Functions

Definition 5

The **moment generating function** (also known as a transform) associated with a RV X, is a function $M_X(s)$ of a scalar parameter s defined by $M_X(s) = E(e^{sX})$.

the simpler notation $M(S)$ can be used whenever the underlying random variable X is clear from context. In more detail, when X is a discrete random variable, the corresponding MGF is given by

$$M(s) = \sum_x e^{sx} p_X(x).$$

Analogously, when continuous, we just replace the summation with an integral to get

$$M(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx.$$

Just an example so that I know what the reference is here:

Example 6 (Discrete Example)

Let

$$p_X(x) = \begin{cases} \frac{1}{2} & x = 2 \\ \frac{1}{6} & x = 3 \\ \frac{1}{3} & x = 5. \end{cases}$$

Then the corresponding transform is

$$M(s) = E(e^{sx}) = \frac{1}{2} + \frac{1}{6}e^{3s} + \frac{1}{3}e^{5s}.$$

Example 7 (Continuous Example)

Let X be an exponential RV with parameter λ:

$$f_X(x) = \lambda e^{-\lambda x} \qquad x \geq 0.$$

Then,

$$\begin{aligned} M(s) &= \lambda \int_0^{\infty} e^{sx} e^{-\lambda x} dx \\ &= \lambda \int_0^{\infty} e^{(s-\lambda)x} dx \\ &= \lambda \left(\frac{e^{(s-\lambda)x}}{s-\lambda} \Big|_0^{\infty} \right. \\ &= \frac{\lambda}{\lambda - s}. \end{aligned}$$

Notice, in above examples, that MGF is a **function** of parameter s, and not a number. We can also find MGF's for functions of X:

Proposition 8 (MGF of Linear Function of RV)

Let $Y = aX + b$. Then,

$$M_Y(s) = E(e^{s(aX+b)}) = e^{sb} E(e^{saX}) = e^{sb} M_X(sa).$$

From our previous example, we see that $M_X(s) = \frac{1}{1-s}$ where X is the exponential distribution

6.1 Moments

Now that we've established what a moment generating function is, now it's time to understand what is being generated.

Let's do a generic MGF

$$M(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx.$$

Now, we take the derivative of this.

$$\begin{aligned}\frac{d}{ds}M(s) &= \frac{d}{ds}\int_{-\infty}^{\infty} e^{sx} f_X(x) dx \\ &= \int_{-\infty}^{\infty} \frac{d}{ds} e^{sx} f_X(x) dx \\ &= \int_{-\infty}^{\infty} x e^{sx} f_X(x) dx.\end{aligned}$$

When $s = 0$, we have that this evaluates to $\int_{-\infty}^{\infty} x f_X(x) dx = E(X)$. If we differentiate n times, then we will get

$$\left(\frac{d^n}{ds^n} M(s)\Big|_{s=0}\right. = \int_{-\infty}^{\infty} x^n f_X(x) dx = E(X^n).$$

6.2 Inversion

Proposition 9 (Inversion Property)

The MGF $M_X(s)$ associated with an RV X uniquely determines the CDF of X, assuming that $M_X(s)$ is finite for all s in some interval $[-a, a]$ for positive a.

6.3 Sum of Independent Random Variables

Proposition 10

Addition of independent random variables corresponds to multiplication of transforms.

Proof. Let $Z = X + Y$. $M_Z(s) = E(e^{sZ}) = E(e^{s(X+Y)}) = E(e^{sX}e^{sY})$. Since X, Y are independent, e^{sX} and e^{sY} are independent random variables for any fixed s. Thus, $E(e^{sX}e^{sY}) = E(e^{sX})E(e^{sY}) = M_X(s)M_Y(s)$. □

We can further extend this; if $X_1, \ldots, X_n$ is a collection of independent random variables and $Z = X_1 + \cdots + x_n$, then $M_Z(s) = M_{X_1}(s) \cdots M_{X_n}(s)$.

7 Concentration Inequalities

Theorem 11 (Markov's Inequality)

$$P(X > a) = \frac{E(X)}{a}.$$

Theorem 12 (Chebyshev's Inequality)

$$P(|X - E(X)| > a) = \frac{\mathrm{Var}(X)}{a^2}.$$

Used in lieu of confidence interval tests.

8 Modes of Convergence

8.1 Pointwise

Definition 13 (Pointwise Convergence)

Fix $\omega \in \Omega$, $\{X_n(\omega)\}_{n=1}^{\infty}$ converges **pointwise** if it becomes a real-valued sequence.

Usually, people don't use this because of reasons highlighted in 104.

8.2 Almost Sure

Definition 14 (Almost Sure Convergence)

$\{x_n\}_{n=1}^{\infty}$ converges **almost surely** to X if $P(\{\omega : \omega \in \Omega, \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1$.

This gets rid of ω with probability 0. If you find an ω such that convergence doesn't hold, it's fine as long sa $P(\omega) = 0$.

8.2.1 Checking for Almost Sure Convergence

There are a couple ways to check if some sequence converges almost surely.

8.3 In Probability

This is a weaker bound for convergence than almost sure convergence.

9 Information Theory

9.1 Entropy

First, we define $\mathcal{X}$ as the range of a random variable X over all events in a probability space.

Definition 15 (Entropy)
Given a discrete random variable X and PMF $P_X(x)$, we have **entropy**

$$H(X) = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}.$$

Furthermore, the average amount of surprise is defined as $E\left[\log \frac{1}{P_X(x)}\right]$.

Moreover, some properties of entropy:

1. $H(X) \geq 0$
2. $H(X)$ is
3. $H(X) \leq \log |x|$, achieved when X is uniform on x.

Where x is the range of $X(\omega)$ for all $\omega \in \Omega$.

Definition 16 (Joint Entropy)
Joint entropy $(X, Y) \sim P_{X,Y}$:

$$H(X,Y) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} P_{X,Y}(x,y) \log \frac{1}{P_{X,Y}(x,y)}.$$

Definition 17 (Conditional Entropy)

$$H(Y|X) = \sum_{x \in \mathcal{X}} H(Y|X = x).$$

Next, we observe some properties of joint and conditional entropy.

Proposition 18 1. (Chain Rule)

$$H(X,Y)H(X) + H(Y|X) = H(Y) + H(X|Y).$$

2. (Conditioning Reduces Entropy)

$$H(Y|X) \leq H(Y).$$

3.

$$H(X,Y) \leq H(X) + H(Y).$$

9.2 Mutual Information

Created by a Bob Fano, who argued more important than entropy.

Definition 19 (Mutual Information)

We define $I(X,Y)$ as the **mutual information** between X and Y, such that

$$\begin{aligned} I(X:Y) &= H(X) - H(X|Y) \geq 0 \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(Y) - H(Y|X). \end{aligned}$$

We can think of $I(X,X) = H(X)$ as well.

Definition 20 (Kullback-Leibler Divergence)

We can also call this **relative entropy**.

$$D(P \parallel Q) = \sum_{x \in \mathcal{X}} P(X) \log \frac{P(x)}{Q(x)} \geq 0.$$

We can see that the mutual information can further be reduced to

$$\begin{aligned} I(X:Y) &= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} P_{X,Y}(x,y) \log \frac{P_{X,Y}(x,y)}{P_X(x) P_Y(y)} \\ &= D(P_{X,Y} \parallel P_X \otimes P_Y), \end{aligned}$$

where we define $P_X \otimes P_Y$ as the cross product.

9.3 Source Coding

Let $X_1, X_2, \ldots, X_n$ be a string of symbols or binary code or etc. in a file. We want to convert this into some compressed $b(X_1, X_2, \ldots X_n)$.

Theorem 21

We assume $X_1, X_2, \ldots, X_n$ are i.i.d as X.

1. There exists a source code such that

$$\lim_{n \to \infty} E\left[\frac{1}{n} |b(x_1, \cdots, x_n)|\right] \leq H(X) + \epsilon$$

   for any $\epsilon > 0$.

2. Conversely, no source code can achieve an average length less than $H(X)$ bits per symbol.

10 Markov Chains

Definition 22 (Markov Chain)

$\{X_n\}_{n \in \mathbb{N}}$ is a discrete-time Markov Chain (DTMC) on state space $\mathcal{X}$ if it satisfies the Markov property: For all positive integers n and feasible sequence of states $x_0, x_1, x_2, \ldots, x_{n+1} \in \mathcal{X}$;

$$\Pr(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = P(X_{n+1} = x_{n+1}, X_n = x_n).$$

We further denote P as the transition probability matrix, which is done by taking the row statistic of $\mathcal{X}$.

10.1 Distributions

Denote distribution of X_n as Π_n. Then, $\Pi_n = \Pi_0 P^n$. We have a **stationarity distribution** $\Pi = \Pi \cdot P$, and this is also called the balance equation.

10.2 Recurrence and Transience

For $x \in \mathcal{X}$, we define $T_x = \min\{n \in \mathbb{N}, X_n = n\}$ as the hitting time of x, and $T_x^+ = \min\{n \in \mathbb{Z}_+ X_n = n\}$.

T_x determines the first time that a Markov chain reaches a certain state, and T_x^+ calculates the same thing except ignoring trivial (initial) cases.

Now, some notation. Let $\Pr_x(A) = \Pr(A|X_0 = x)$ and $E_x[Z] = E[Z|X_0 = x]$. This is probability and expectation given an initial state in the Markov chain. Furthermore, let $\rho_{x,y} = \Pr_x(T_y^+ < \infty)$, $\rho_x = \rho_{x,x}$.

Definition 23

State x is **recurrent** if $\rho_x = 1$, **transient** otherwise.

A recurrent state essentially means that a state in a Markov chain will certainly be reached again.

Proposition 24

Denote $N_x = \sum_{n\in\mathbb{N}} \mathbb{I}(X_n = x)$. Then,

1. If x is recurrent, then $N_x = \infty$ almost surely.
2. If x is transient then $E_x[N_x] = \frac{\rho(x)}{1-\rho(x)}$.

10.3 Classification of States

Definition 25 (Communicating Class)

We say x communicates with y if $\rho_{x,y} > 0$ and $p_{y,x} > 0$.

A **communicating class** is a maximal set of states which communicate with each other.

Definition 26

Markov Chain is **irreducible** if it consists of only a single communicating class.

The class property is a property that's necessarily shared by all members of class. Anyways, now time to start applying the many definitions we've just made:

Theorem 27

Recurence and transience are class properties.

Are we not going over the proof for this?

Proposition 28

Every finite state irreducible chain is recurrent.

Proof. Basically prove that one of the states must be recurrent using the fact that there are finite states, and then use the above theorem to see that this is a class property. □