CS70 Probability Concept Walkthrough

Yuxiang Yang Email:yxyang@berkeley.edu

April 30, 2017

This note is written for the CS70 catchup workshop held at the end of Spring 2017. It's intended to be a guide for students who left behind to catchup with what the course is covering, and is supposed to serve as a summary of key concepts and proofs covered in class.

All the questions I put in this note are what I consider important concepts/proofs for this class. Make sure you are very familiar with them.

What to expect from this note:

- A summary of key probability concepts, formulas
- A collection of commonly-used tricks to solve probability problems

What NOT to expect from this note:

- A thorough walkthrough of all concepts. Please refer to either the official notes or Sinho's probability notes if something in this worksheet feels too brief to read.
- A list of comprehensive topics for any exam. Although I'm a TA this semester, by the time this worksheet is written, I've had no knowledge about any details about the final exam.

Contents

1 Basic Probability Concepts

1.1 Counting

1.1.1 Two basic rules of counting

1. *First rule of counting:* If a decision can be made in k steps, where there are n_1 choices in the first step. For every choice of first step, there are n_2 choices in the second step, etc. Then there are ______ ways to make the decision.

2. *Second rule of counting:* Let A, B be two sets. If there exists a k-to-1 function from A to B, then the relationship between $|A|$ and $|B|$ can be characterized as: ______.

Make sure you know what the two rules of counting means. All fancy counting formulas are derived from these two rules.

1.1.2 Typical of counting problems

1. *k-permutation of n* You choose k people out of n people to line up and take a photo (i.e. order matters), how many possible arrangements do you have?

2. *k-combination of n* You choose k people out of a n people team to attend a conference (i.e. order does not matter), how many possible teams can you form? Which rule of counting are you using?

3. *Stars and bars v1.0* How many **non-negative** integer solutions are there to the following equation, where n and k are integers?

   $$x_1 + x_2 + \cdots + x_k = n$$

   Note that you can also have a balls and bins formulation for this problem: How many ways are there to throw n identical balls into k distinct bins? But I like the above specification first since it's clearer.

4. *Stars and bars v2.0* How many **positive** integer solutions are there to the following equation, where n and k are integers?

   $$x_1 + x_2 + \cdots + x_k = n$$

   Again, a balls and bins formulation would be: how many ways are there to throw n identical balls into k distinct bins, where each bin gets at least 1 ball?

1.1.3 Combinatorial proofs

In combinatorial proofs you seek to tell the same "story of counting" for both sides of equation, and here are some general tips:

1. When you see $\binom{n}{k}$ or similar expressions, try to connect it with some familiar counting problem you've seen, and come up with a "story" for it.
2. When you see multiplication of terms in expression, try to connect it with first rule of counting, i.e. making choice with multiple steps.
3. When you see addition in expression, try to interpret it as "considering different cases". For example, there are 4 Chinese restaurants, 5 Japanese restaurants and 2 fast food nearby, how many ways do I have for lunch?

1.2 Discrete Probability Concepts

1.2.1 Sample Space, Outcome, Event

A *sample space* Ω is the set of all possible *outcomes*ω of a random experiment. Any subset of the sample space is called an *event*.

In a *discrete probability space*, we assign a probability $P(\omega)$ for any $\omega \in \Omega$. In addition, we require:

- $\forall \omega \in \Omega, 0 \leq P(\omega) \leq 1$
- $\sum_{\omega \in \Omega} P(\omega) = 1$

1.2.2 Probability with Multiple Events

1. Joint, conditional and Bayes' rule in one line:

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

   Sometimes you need to do more fancy Bayes rules with multiple events, consider 5.13 in midterm 2.

2. *Independence* State the condition for two events to be independent, both using conditional probability and joint probability. What does independence mean intuitively?

3. *Product rule* Suppose $A_1, A_2 \cdots A_n$ are not necessarily independent. How would you compute$P(A_1 \cap A_2 \cap \cdots A_n)$ from a bunch of conditional probabilities?

4. *Union bound* State the union bound for events $A_1, \cdots A_n$. When does equality hold?

2 Random Variables

This section covers both discrete and continuous random variables.

2.1 Basic Definition and distribution

Definition. A *random variable* X on a sample space Ω is a function $X : \Omega \mapsto \mathbb{R}$ that assigns a real number to each outcome $\omega \in \Omega$.

Note that the above definition holds for both discrete and continuous random variables. The difference is that for discrete random variable, the sample space is finite (or countably infinite, say, for geometric random variables). For continuous random variable, the sample space is infinite. Therefore, the *distribution* for discrete and continuous r.v. is defined differently.

Definition. *Distribution for discrete random variables* Long words short, the distribution of a discrete random variable is a set of values $\{a, P(X = a)\}$, i.e. the value of random variable and associated probability.

For a continuous random variable X, we cannot assign a non-zero probability for any specific value of X. Instead, we define PDF and CDF.

Definition. *Distribution for continuous random variables*

What are the requirements for a function to be a valid PDF? CDF? How do you convert between PDF and CDF?

Let X be a continuous random variable, how do you find $P(X \leq x)$? $P(X \geq x)$? $P(a \leq X \leq b)$? Express them both in terms of PDF ($f_X(x)$) and CDF ($F_X(x)$).

2.2 Joint Distribution, Independence

1. Discrete Case: Joint PMF: $P(X = x, Y = y)$

   How would you find the marginal distribution $P(X = x)$ from the joint distribution $P(X = x, Y = y)$?

What does it mean for two discrete random variables to be independent?

2. Continuous Case
   Joint PDF: $f_{X,Y}(x,y)$
   Joint CDF: $F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(x,y)dydx = P(X \leq x, Y \leq y)$
   Express $P(a \leq X \leq b, c \leq Y \leq d)$ as a function of both the joint PDF $f_{X,Y}(x,y)$.

   How would you get the marginal distribution $f_X(x)$ from the joint PDF $f_{X,Y}(x,y)$?

   What does it mean for two continuous random variables to be independent?

2.3 Expectation

2.3.1 Definition

Discrete case: $E[X] = \sum_x x \cdot P(X = x)$
Continuous case: $E[X] = \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x$

2.3.2 Expectation of function of r.v

Discrete case: $E[g(X)] = \sum_x g(x) \cdot P(X = x)$
Continuous case: $E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\mathrm{d}x$
Protip: The expectation of any term, is the sum/integral of the value of that term times the probability (or density) that the term takes that value.

2.3.3 Properties

What is linearity of expectation? Does it hold for both discrete and continuous random variables?

2.4 Variance

2.4.1 Definition

Use linearity of expectation to justify the following:

$$Var(X) = E[(X - E[X])^2] = E[X^2] - E^2[X]$$

Note that this definition works for both continuous and discrete random variables.

2.4.2 Properties

(Yes/No) Is the variance always positive?
(Yes/No) Could the variance ever be zero? In what case?

Prove $Var(cX) = c^2 Var(X)$.

When does it hold that $Var(X + Y) = Var(X) + Var(Y)$? Justify your reasoning.

2.5 Covariance

2.5.1 Definition

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

2.5.2 Properties

Does X and Y being independent implies $Cov(X, Y) = 0$? Does zero covariance imply independence? Prove or give counterexample.

What does a positive covariance imply? This could act as a good sanity check when you're doing problems.

Prove the following properties about covariance.

1. $Cov(X,X) = Var(X)$

2. $Cov(X, aY + b) = a \cdot Cov(X,Y)$

3. $Cov(X, Y + Z) = Cov(X,Y) + Cov(X,Z)$ Note that this property allows you to further apply the magic of indicator random variables! e.g. If $X = X_1 + \cdots + X_n, Y = Y_1 + \cdots Y_m$, then you would have $Cov(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} Cov(X_i, Y_j)$

4. $Var(X+Y) = Var(X) + Var(Y) + 2cov(X,Y)$

2.6 Conditional Expectation

2.6.1 Definition

Discrete Case: $E[X|Y=y] = \sum_x x \cdot P(X = x \mid Y = y)$
Continuous Case: $E[X|Y=y] = \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y)\mathrm{d}x$

2.6.2 Properties

In general, is $E[Y|X]$ a constant or a function? If it's a function, what is the "input" of this function?

Prove the following properties about conditional probability:

1. Factoring known values: $E[h(X) \cdot Y|X] = h(X)E[Y|X]$

2. If X and Y are independent, $E[Y|X] = E[Y]$

3. Law of iterated expectation: $E[Y] = E[E[Y|X]]$

For more properties about covariance and conditional expectation, please refer to the section about estimation below.

2.7 Common Random Variables

1. Make sure you are familiar with *Binomial, Geometric, Poisson, Exponential* random variables, what underlying process they model, their PMF/PDF, expectation and variance. Also it's helpful to identify the kind of random variables that the problem is describing, as that would save a lot of time.

   (a) *Binomial* (with parameter p): You flip a biased coin n times with head probability p, what's the number of heads that shows up? Or, more abstractly (and more generally), you try something n times, each trial succeeds with probability p and is independent of other trials. How many successes are you going to see?

   (b) *Geometric* (with parameter p): You flip a biased coin with head probability p over and over again until the head comes up for the first time. How many flips will you need? Or, you keep trying something until first success. Each trial succeeds with probability p and is independent of other trials. How many trials will you need?

   (c) *Poisson* Think of this as the continuous version of Binomial. See page 7 of note 19 for a detailed proof.

   (d) *Exponential* Think of this as the continuous version of Geometric. You still count the amount of time to wait until first success, except that things are happening in continuous time. See page 6 of note 20 for a detailed proof.

2. Make sure you know what the *memoryless* property of geometric and exponential random variable means, and how it could be used to simplify a complicated process. Example problem: Discussion13A: problem 2.

2.8 General tips/tricks about random variables

It's common to have a problem describing a complicated process, and ask you to do some calculation of random variables associated to it. Here are some general tricks in approaching such problems:

2.8.1 Indicator random variables

(Related Problems: Discussion 9a, 9b, 11a)

This is probably the most important important tool in discrete random variables. Use indicator random variables to break a big process down, and go back to whatever you're looking for by adding things back. Note that indicator can not only be used to compute expectation, it works for variance and covariance as well.

Also, don't be afraid to multiply indicators! In calculating the variance/covariance, it's common to compute $E[X_iY_j]$, where X_i, Y_j are indicators. Note that $X_iY_j = 1$ iff they're both one, and that means the events that both indicators direct on are happening.

2.8.2 Go from CDF to PDF

(Related Problem: Discussion 13A: 1, 3)

When you want to know the distribution of some random variable Y as a function of a bunch of random variables $X_1 \cdots X_n$, it's helpful to start by finding the CDF $F_Y(y) = P(Y \leq y)$, then differentiate to get PDF. A good example would be $Y = \max(X_1 \cdots X_n)$, or $Y = \min(X_1 \cdots X_n)$.

2.8.3 From conditional expectation to expectation

(Related Problem: Discussion11b: 2,3; Homework12: 6,7)

Sometimes the expectation of a random variable is hard to find. However, conditioning on some other random variable, the conditional expectation is just a model we're all familiar with. In this case, law of iterated expectation comes in handy to convert from conditional expectation to expectation.

2.8.4 Just sum/integrate over all cases

(Related Problem: HW9/5,6; Midterm2 5.11; Dis13A: 3, HW13: 7)

Sometimes, to find the probability of some event / some random variable taking some value, it doesn't hurt just to sum over all possible cases in which such event would occur. e.g. Let X, Y be non-negative integer-valued random variables, $P(X \leq Y) = \sum_{x=0}^{\infty} \sum_{y=x}^{\infty} P(X = x, Y = y)$, or $P(X + Y = k) = \sum_{x=0}^{k} P(X = x, Y = k - x)$.

Note that this works for continuous random variables as well.

3 Applications of probability

3.1 Inequalities and Confidence Intervals

3.1.1 Markov's Inequality, Chebyshev's Inequality

What is Markov's inequality? Does it hold for *every* random variable?

What is Chebyshev's inequality? How to derive it from Markov's inequality?

3.1.2 Confidence Intervals

Related problems: HW10: 8; Dis10A

Confidence intervals derived by Chebyshev's inequality is always an interval centered around the mean (expectation) of a random variable. The ______ (wider/narrower) the interval is, the greater the probability is for the value to be in this interval.

For any problem about confidence interval, all you need to do is to tune the *width* of the confidence interval such that the probability of staying in the interval is high, or the probability of staying out of the interval is low. Make sure you translate the problem statement to this level before doing any computations.

3.2 Estimation

Think about estimation in general as the following story: You have two random variables X and Y and you know their joint distribution. Now, someone gives you the value of X, and you want to give the best guess of Y based on this information. For any estimation function $G(X)$, we compute the estimation error as $E[(Y - G(X))^2]$.

1. What is projection property? How would you use projection property to prove that letting $G^*(X) = E[Y|X]$ minimizes $E[(Y - G(X))^2]$?

   Related Problem: HW12: 3;

2. What is the formula for LLSE? In what case would $L[Y \mid X]$ just be a constant? In that case, is X and Y independent?

3. If $E[Y|X]$ is linear in X, what does that tell you about the LLSE $L(Y|X)$? Can the LLSE ever be a better estimation than MMSE?

   Related Problem: HW12:4

3.3 Markov Chain

3.3.1 Structure of Markov Chain

What is Markov Property? (HW13 Q2)

You need 3 things to define a Markov chain: a finite set of states X, transition probability $P(i, j) = P(X_n = j|X_{n-1} = i)$ and an initial distribution π_0. Make sure you know these 3 things mean.

Given an initial probability distribution, make sure you know how to calculate the probability distribution in the next timestep.

3.3.2 Invariant distribution and hitting time

What does it mean for a Markov Chain to be irreducible? Aperiodic? Give examples of markov chains that (a) is *not* irreducible (b) is irreducible but *not* aperiodic (c) irreducible and aperiodic

Does every markov chain has a unique invariant distribution? How to compute the invariant distribution of a Markov Chain?

How do you compute the hitting time, i.e. the expected number of timesteps until you first enter a certain state?

Make sure you know how to model things as a Markov chain.

Related Problem: Dis12A/B; HW13