Lisa Yan
CS109

Handout #5
May 15, 2020

Probability Reference

Based on a handout by Chris Piech and Lisa Yan

Notation

This section maps between math notation used in CS109 and English. Note: “or" is not notation.

0.1 Events

E or F	Capital letters can denote events
A or B	Sometimes they denote sets
$\|E\|$ or $\|A\|$	Size of an event or set
E^C or A^C	Complement of an event or set
EF or AB	Intersection of events or sets
$E \cup F$ or $A \cup B$	Union of events or sets
$P(E)$	The probability of an event E
$P(E\|F)$	The conditional probability of an event E given F
$\binom{n}{k}$	Binomial coefficient
$\binom{n}{r_1, r_2, r_3}$	Multinomial coefficient

0.2 Random Variables

x or y or i	Lower case letters often denote regular variables
X or Y	Capital letters are used to denote random variables
$E[X]$	Expectation of X
$\text{Var}(X)$	Variance of X
$p_X(x)$	Probability mass function (PMF) of X
$p_{X,Y}(x, y)$	Joint probability mass function (PMF) of X and Y
$p_{X\|Y}(x\|y)$	Conditional probability mass function (PMF) of X given Y
$f_X(x)$	Probability density function (PDF) of X
$f_{X,Y}(x, y)$	Joint probability density function (PDF) of X and Y
$f_{X\|Y}(x\|y)$	Conditional probability density function (PDF) of X given Y
$F_X(x)$	Cumulative distribution function (CDF) of X
$F_{X,Y}(x, y)$	Joint cumulative distribution function (CDF) of X and Y
$F_{X\|Y}(x\|y)$	Conditional cumulative distribution function (CDF) of X given Y

$X \sim \mathrm{Ber}(p)$	X is a Bernoulli random variable with parameter p
$X \sim \mathrm{Bin}(n, p)$	X is a Binomial random variable with parameters n, p
$X \sim \mathrm{Poi}(\lambda)$	X is a Poisson random variable with parameter λ
$X \sim \mathrm{Geo}(p)$	X is a Geometric random variable with parameter p
$X \sim \mathrm{NegBin}(r, p)$	X is a Negative Binomial random variable with parameters r, p
$X \sim \mathrm{HypGeo}(n, N, m)$	X is a Hyper Geometric random variable with parameters n, N, m
$X \sim \mathcal{N}(\mu, \sigma^2)$	X is a Gaussian random variable with mean μ and variance σ^2
$X \sim \mathrm{Uni}(a, b)$	X is a Uniform random variable with parameters a, b
$X \sim \mathrm{Exp}(\lambda)$	X is a Exponential random variable with parameter λ
$X \sim \mathrm{Beta}(a, b)$	X is a Beta random variable with parameters a, b

1 Combinatorics

Refer to Lecture Notes 1 and 2 for a more complete summary of combinatorics. We've highlighted the main rules here.

Inclusion-Exclusion Principle: If the outcome of an experiment can either be drawn from set A or set B, and sets A and B may potentially overlap (i.e., it is not guaranteed that $A \cap B = \emptyset$), then the number of outcomes of the experiment is $|A \cup B| = |A| + |B| - |A \cap B|$.

General Principle of Counting: If an experiment has r parts such that part i has n_i outcomes for all $i = 1, \ldots, r$, then the total number of outcomes for the experiment is $\prod_{i=1}^{r} n_i = n_1 \times n_2 \times \cdots \times n_r$

Basic Pigeonhole Principle: For positive integers m and n, if m objects are placed in n buckets, where $m > n$, then at least one bucket must contain at least two objects.

Permutations	Consider the number of ways to order n objects.
n objects are distinct (distinguishable)	$n(n-1)(n-2)\cdots 1 = n!$ ways
n_1 are indistinct (indistinguishable), n_2 are indistinct, …, and n_r are indistinct	$\dfrac{n!}{n_1!n_2!\ldots n_r!}$ ways
Combinations	Consider the number of ways to select groups of objects from a set of n distinguishable objects.
Select r objects	$\dfrac{n!}{r!(n-r)!} = \dbinom{n}{r}$ ways
Select r groups of objects, such that group i has size n_i, and $\sum_{i=1}^{r} n_i = n$	$\dfrac{n!}{n_1!n_2!\cdots n_r!} = \dbinom{n}{n_1, n_2, \ldots, n_r}$ ways
Bucketing	Consider the number of ways to place n objects into r containers.
r distinguishable objects	n^r ways
r indistinguishable objects	$\dfrac{(n+r-1)!}{n!(r-1)!} = \dbinom{n+r-1}{n} = \dbinom{n+r-1}{r-1}$ ways

2 Probability

2.1 Definitions, Axioms, and Corollaries

Frequentist definition of probability:

$$P(E) = \lim_{n\to\infty} \frac{n(E)}{n}$$

Axiom 1: $0 \le P(E) \le 1$
Axiom 2: $P(S) = 1$
Axiom 3: If E and F are mutually exclusive ($E \cap F = \emptyset$), then $P(E) + P(F) = P(E \cup F)$

Corollary 1: $P(E^C) = 1 - P(E) \qquad (= P(S) - P(E))$
Corollary 2: $E \subseteq F$, then $P(E) \le P(F)$
Corollary 3: $P(E \cup F) = P(E) + P(F) - P(EF)$ (Inclusion-Exclusion Principle)

General Inclusion-Exclusion Principle:

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{i_1<\cdots<i_r} P(E_{i_1} E_{i_2} \ldots E_{i_r})$$

Define S as a sample space with equally likely outcomes. Then $P(E) = \dfrac{|E|}{|S|}$.

DeMorgan's Laws applied to probability:

$$P((E \cup F)^C) = P(E^C \cap F^C)$$
$$P((E \cap F)^C) = P(E^C \cup F^C)$$

2.2 Conditional Probability

Def. conditional probability	$P(E \mid F) = \dfrac{P(EF)}{P(F)} = \dfrac{P(E \cap F)}{P(F)}$
Chain rule	$P(EF) = P(E \mid F)P(F)$
	$P(E_1 E_2 \ldots E_n) = P(E_1)P(E_2 \mid E_1) \ldots P(E_n \mid E_1 E_2 \ldots E_{n-1})$
Law of Total Probability	$P(F) = P(F \mid E)P(E) + P(F \mid E^C)P(E^C)$
	$P(F) = \sum_{i=1}^{n} P(F \mid E_i)P(E_i)$
Bayes' Theorem	$P(E \mid F) = \dfrac{P(F \mid E)P(E)}{P(F)}$
	$= \dfrac{P(F \mid E)P(E)}{P(F \mid E)P(E) + P(F \mid E^C)P(E^C)}$
	$= \dfrac{P(F \mid E)P(E)}{\sum_i P(F \mid E_i)P(E_i)}$

Conditional paradigm: If we consistent conditionally on an event G, all of the laws of probability still hold.

2.3 Independence

Independence: Two events E and F are independent if and only if $P(EF) = P(E)P(F)$. It can be shown that independence of E and F implies:

- $P(E|F) = P(E)$ and $P(F|E) = P(F)$
- $P(E|F^C) = P(E)$ and $P(F|E^C) = P(F)$

In general, n events $E_1, E_2, \ldots, E_n$ are independent if for every subset with r elements (where $r \leq n$) it holds that:

$$P(E_{i_1}, E_{i_2}, \ldots, E_{i_r}) = P(E_{i_1})P(E_{i_2})\ldots P(E_{i_r})$$

Two events E and F are **conditionally independent** given a third event G holds if $P(EF|G) = P(E|G)P(F|G)$.

3 Random Variables

Discrete Random Variables	
Probability Mass Function (PMF)	$p_X(x)$
PMF must sum to 1	$\sum_x p_X(x) = 1$
Probability with the PMF	$P(X = x) = p_X(x)$
Cumulative Distribution Function (CDF)	$F_X(a) = \sum_{x \leq a} p_X(x)$
Continuous Random Variables	
Probability Density Function (PDF)	$f_X(x)$
PDF must integrate to 1	$\int_{-\infty}^{\infty} f_X(x)dx = 1$
Probability with the PDF	$P(a \leq X \leq b) = \int_a^b f_X(x)dx$
Cumulative Distribution Function (CDF)	$F_X(a) = \int_{-\infty}^{a} f_X(x)dx$

We can compute the probability that the random variable X lies in an interval using the CDF, F_X: $P(a < X \leq b) = F_X(b) - F_X(a)$.

3.1 *Expectation and Variance*

Other names for expectation: mean, average, first moment, expected value.

Definition:	$E[X] = \sum_x x p_X(x)$	X discrete, PMF p_X
	$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$	X continuous, PDF f_X
Linearity of Expectation:	$E[aX + bY + c] = aE[X] + bE[Y] + c$	
Law of the Unconscious Statistician (LOTUS):		
	$E[g(X)] = \sum_x g(x) p_X(x)$	X discrete, PMF p_X
	$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$	X continuous, PDF f_X

Linearity of expectation is often stated as: The expectation of a sum is equal to the sum of expectations.

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

Definition of **variance**: $\mathrm{Var}(X) = E[(X - E[X])^2]$.

- Most often computed as $\mathrm{Var}(X) = E[X^2] - (E[X])^2$.
- Note: $\mathrm{Var}(X) \geq 0$.
- Standard deviation is defined as $\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$. Note: $\mathrm{SD}(X) \geq 0$.

3.2 *Common Discrete Distributions*

If a random variable follows a particular distribution we use the $\sim$ symbol to represent that the type of the random variable and pass in the appropriate parameters. For example if X follows a Normal distribution with mean 5 and variance 4 we write $X \sim \mathcal{N}(5, 4)$.

All probability mass functions (PMFs) are 0 outside the support.

Bernoulli Random Variable. $X \sim \mathrm{Ber}(p)$

An indicator variable that takes on the value 1 (“success”) or 0. Often the variable is defined to be 1 if an underlying event has occured, 0 otherwise.

PMF:	$p_X(k) = \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = 0 \end{cases}$	Support:	$\{0, 1\}$
$E[X]$:	p	$\mathrm{Var}(X)$:	$p(1 - p)$
Parameter:	p: The probability that X is 1		

Note: Sometimes in Machine learning algorithms, a differentiable version of the PMF is used: $p^k(1 - p)^{1-k}$. We will talk about this later.

Binomial Random Variable. $X \sim \text{Bin}(n, p)$
A variable that represents the number of successes in a fixed number of independent trials. The probability of success must be the same for each trial.

PMF:	$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$	Support:	$\{0, 1, 2, \ldots, n\}$
$E[X]$:	np	Var(X):	$np(1-p)$
Parameters:	n: the number of trials		
	p: the probability of success of each trial		

Note: $\text{Bin}(1, p) = \text{Ber}(p)$.

Poisson Random Variable. $X \sim \text{Poi}(\lambda)$
The number of events occurring in a fixed interval of time or space if these events occur independently with a constant average rate.

PMF:	$p_X(k) = \dfrac{\lambda^k e^{-\lambda}}{k!}, k \geq 0$	Support:	$\{0, 1, 2, \ldots\}$
$E[X]$:	λ	Var(X):	λ
Parameter:	λ: the average number of events per fixed interval.		

Note: The Poisson RV is the number of events in an interval of time. The Exponential RV is a continuous RV that models the time until the next event occurs. They have the same parameter, λ.

Note 2: The Poisson can approximate the Binomial when λ is "moderate" (in this class, defined as $n > 20$ and $p < 0.05$ or $n > 100$ and $p < 0.1$) when the trials are mildly dependent, or even when the probability of success varies slightly between trials.

Geometric Random Variable. $X \sim \text{Geo}(p)$
The number of independent Bernoulli trials until the first success. The probability of success must be the same for each trial.

PMF:	$p_X(k) = (1-p)^{k-1} p$	Support:	$\{1, 2, \ldots\}$
$E[X]$:	$\dfrac{1}{p}$	Var(X):	$\dfrac{1-p}{p^2}$
Parameter:	p: the probability of success of each trial		

Negative Binomial Random Variable. $X \sim \text{NegBin}(r, p)$
The number of independent Bernoulli trials until the r-th success. The probability of success must be the same for each trial.

PMF:	$p_X(k) = \binom{k-1}{r-1}(1-p)^{k-r}p^r$	Support:	$\{r, r+1, \ldots\}$
$E[X]$:	$\frac{r}{p}$	Var(X):	$\frac{r(1-p)}{p^2}$
Parameters:	r: the total number of successes to obtain		
	p: the probability of success of each trial		

Note: $\text{NegBin}(1, p) = \text{Geo}(p)$.

3.3 *Common Continuous Distributions*

All probability density functions (PDFs) are 0 outside the support.

Uniform Random Variable. $X \sim \text{Uni}(a, b)$

PDF:	$f_X(x) = \frac{1}{b-a}$	Support:	$a \le x \le b$
$E[X]$:	$\frac{a+b}{2}$	Var(X):	$\frac{(b-a)^2}{12}$

Exponential Random Variable. $X \sim \text{Exp}(\lambda)$
The waiting time until an event occurs when events occur independently with a constant average rate.

PDF:	$f_X(x) = \lambda e^{-\lambda x}$	Support:	$x \ge 0$
$E[X]$:	$\frac{1}{\lambda}$	Var(X):	$\frac{1}{\lambda^2}$
CDF:	$F_X(x) = 1 - e^{-\lambda x}$		

Note: The Exponential RV models the time until the next event occurs. The Poisson RV is a discrete RV that models the number of events in an interval of time. They have the same parameter, λ.

Note: The Exponential RV is memoryless, in that the time you wait until the first success is distributed as an Exponential RV, independent of the amount of time you have waited so far.

Normal (Gaussian) Random Variable. $X \sim \mathcal{N}(\mu, \sigma^2)$

PDF:	$f_X(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}$	Support:	$-\infty < x < \infty$
$E[X]$:	μ	Var(X):	σ^2

Note: When $\mu = 0$ and $\sigma^2 = 1$ ("zero mean, unit variance"), X is called a Standard Normal with CDF Φ.

Note 2: The Normal can approximate a Binomial with larger variance (in this class, defined as $np(1-p) > 10$. All trials must be independent. This approximation comes from the Central Limit Theorem.

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with CDF F_X. The following properties hold:

Linearity:	$aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$
Standard Normal:	$Z = \frac{X-\mu}{\sigma}$ is the Standard Normal with CDF Φ. Therefore $F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$.

4 Joint Distributions

	Jointly Continuous X, Y	**Jointly Continuous** X, Y
Joint PMF	$p_{X,Y}(x,y) = P(X = x, Y = y)$	–
Joint PDF	–	$f_{X,Y}(x,y)$
Joint CDF	$F_{X,Y}(x,y) = P(X \le x, Y \le y)$	
Marginal distributions	$p_X(a) = \sum_y p_{X,Y}(a,y)$	$f_X(a) = \int_{-\infty}^{\infty} f_{X,Y}(a,y)dy$
	$p_Y(b) = \sum_x p_{X,Y}(x,b)$	$f_Y(b) = \int_{-\infty}^{\infty} f_{X,Y}(x,b)dx$
Conditional distributions	$p_{X\|Y}(x\|y) = \dfrac{p_{X,Y}(x,y)}{p_Y(y)}$	$f_{X\|Y}(x\|y) = \dfrac{f_{X,Y}(x,y)}{f_Y(y)}$
Independence	$p_{X,Y}(x,y) = p_X(x)p_Y(y)$	$f_{X,Y}(x,y) = f_X(x)f_Y(y)$
	$p_{X\|Y}(x\|y) = p_X(x)$	$f_{X\|Y}(x\|y) = f_X(x)$
Bayes' Theorem	$p_{X\|Y}(x\|y) = \dfrac{p_{Y\|X}(y\|x)p_X(x)}{p_Y(y)}$	$f_{X\|Y}(x\|y) = \dfrac{f_{Y\|X}(y\|x)f_X(x)}{f_Y(y)}$

We can compute the probability involving two jointly distributed random variables X and Y using their joint CDF, $F_{X,Y}$: $P(a_1 < X \le a_2, b_1 < Y \le b_2) = F_{X,Y}(a_2, b_2) - F_{X,Y}(a_1, b_2) - F_{X,Y}(a_2, b_1) + F_{X,Y}(a_1, b_1)$.

In general, n random variables $X_1, X_2, \dots, X_n$ are independent if for all $x_1, x_2, \dots, x_n$:

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i) \qquad \text{(jointly discrete)}$$

$$P(X_1 \leq x_1, X_2 = x_2, \dots, X_n \leq_n) = \prod_{i=1}^{n} P(X_i \leq x_i) \qquad \text{(jointly continuous)}$$

n variables $X_1, X_2, \dots, Xn$ are independent and identically distributed (i.i.d.) random variables if they are independent and have the same PMF (if discrete) or PDF (if continuous).

4.1 Independent Sums of Random Variables

If X and Y are independent, then

$$P(X + Y = n) = \sum_k P(X = k)P(Y = n - k) \qquad (X, Y \text{ jointly discrete})$$

$$f_{X+Y}(\alpha) = \int_{-\infty}^{\infty} f_X(x) f_Y(\alpha - x) dx \qquad (X, Y \text{ jointly continuous})$$

Common Sums of Independent Random Variables

Independent X, Y	**Distribution of** $X + Y$
$X \sim \text{Bin}(n_1, p),\ \ Y \sim \text{Bin}(n_2, p)$	$\text{Bin}(n_1 + n_2, p)$
$X \sim \text{Poi}(\lambda_1),\ \ Y \sim \text{Poi}(\lambda_2)$	$\text{Poi}(\lambda_1 + \lambda_2)$
$X \sim \text{Uni}(0, 1),\ \ Y \sim \text{Uni}(0, 1)$	$f_{X+Y}(\alpha) = \begin{cases} \alpha & 0 \leq \alpha \leq 1 \\ 2 - \alpha & 1 < \alpha \leq 2 \\ 0 & \text{otherwise} \end{cases}$
$X \sim \mathcal{N}(\mu_1 \sigma_1^2),\ \ Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$	$\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
Independent $X_1, X_2, \dots, X_n$	**Distribution of** $\sum_{i=1}^{n} X_i$
$X_i \sim \text{Bin}(n_i, p)$ for $i = 1, \dots, n$	$\text{Bin}(\sum_{i=1}^{n} n_i, p)$
$X_i \sim \text{Poi}(\lambda_i)$ for $i = 1, \dots, n$	$\text{Poi}(\sum_{i=1}^{n} \lambda_i)$
$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1, \dots, n$	$\mathcal{N}(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)$

4.2 Statistics of multiple RVs

Law of The Unconscious Statistician, extended to $g(X, Y)$, a function of two jointly distributed random variables:

$$E[g(X, Y)] = \sum_x \sum_y g(x, y) p_{X,Y}(x, y) \qquad (X, Y \text{ jointly discrete})$$

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dy dx \qquad (X, Y \text{ jointly continuous})$$

Conditional expectation of X given $Y = y$:

$$E[X|Y = y] = \sum_x xP(X = x|Y = y) = \sum_x xp_{X|Y}(x|y) \qquad (X, Y \text{ jointly discrete})$$

$$E[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx \qquad (X, Y \text{ jointly continuous})$$

Law of Total Expectation:

$$\begin{aligned} E[X] &= E[E[X|Y]] \\ &= \sum_y E[X|Y = y]P(Y = y) && (Y \text{ discrete}) \\ &= \int_{-\infty}^{\infty} E[X|Y = y]f_Y(y)dy && (Y \text{ continuous, density } f_Y(y)) \end{aligned}$$

Definition of **covariance**: $\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$.

- Most often computed as $\text{Cov}(X) = E[XY] - E[X]E[Y]$.
- Correlation of X and Y: $\rho(X, Y) = \dfrac{\text{Cov}(X, Y)}{SD(X)SD(Y)}$
- Relation to variance: $\text{Var}(X) = \text{Cov}(X, X)$
- Symmetry: $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- Non-linear: $\text{Cov}(aX + b, Y) = a\text{Cov}(X, Y)$
- Covariance of sums: $\text{Cov}(\sum_i X_i, \sum_j Y_j) = \sum_i \sum_j \text{Cov}(X_i, Y_j)$

Variance of sums:

$$\text{Var}(X + Y) = \text{Var}(X) + 2 \cdot \text{Cov}(X, Y) + \text{Var}(Y)$$

Independence of two random variables X and Y implies

- $E[XY] = E[X]E[Y]$ (the converse is not necessarily true), and therefore
- $\text{Cov}(X, Y) = 0$ and $\rho(X, Y) = 0$, and furthermore
- $\text{Var}(X, Y) = \text{Var}(X) + \text{Var}(Y)$.

4.3 Common joint distributions

Multinomial Distribution

A distribution that models the counts of outcomes $i = 1, 2, \ldots, m$, respectively, in a fixed number of independent trials, where each trial results in one of m outcomes.

Joint PMF:	$P(X_1 = c_1, X_2 = c_2, \ldots, X_m = c_m) = \binom{n}{c_1, c_2, \ldots, c_m} p_1^{c_1} p_2^{c_2} \cdots p_m^{c_m}$
Support:	$\sum_{i=1}^{m} c_i = n$, where c_i is a non-negative integer for $i = 1, \ldots, m$
Parameters:	n: the total number of trials
	$p_1, p_2, \ldots, p_m$: the probabilities of m outcomes, where p_i is the probability of outcome i and $\sum_{i=1}^{m} p_i = 1$.

Bivariate Normal (Gaussian) Distribution. $\boldsymbol{X} = (X_1, X_2) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Joint PDF:

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)}$$

Support:	$-\infty < x < \infty$	
Parameters:	$\boldsymbol{\mu} = (\mu_1, \mu_2)$: mean vector	
	$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$: covariance matrix	
Marginal distributions:	$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$	$X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$

Note: $\rho\sigma_1\sigma_2 = \mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1)$. When $\rho = 0$, X_1 and X_2 are independent.