CSE 312 Notes

Contents

1 Counting

1.1 Sum Rule

If you can choose from one of n choices or one of m choices then the total number of outcomes is $n + m$

1.2 Product Rule

If each outcome is constructed by a sequential process where there are

- n_1 choices for the first step
- n_2 choices for the second step (given the choice for the first step)
- n_k choices for the k^{th} step (given the choice for the previous step)

then the total number of outcomes is $n_1 \times n_2 \times ... \times n_k$

1.3 Power Set

The power set of a set A is the set of all subsets of A, including the empty set and A itself

- $P(A) = \{S \mid S \subseteq A\}$
- $P(\varnothing) = \{\varnothing\}$
- $P(\{x, y\}) = \{\varnothing, \{x\}, \{y\}, \{x, y\}\}$

A set with n elements has 2^n power sets

1.4 Permutations

There are $n!$ ways to order n distinct objects

1.5 Complementary Counting

Let U be a set and S a subset of interest. Let $U \backslash S$ denote the set difference. Then $|U \backslash S| = |U| - |S|$

1.6 ^nP_k Permutations

There are $^nP_k = \frac{n!}{k!}$ ways to *arrange* k out of n distinct objects without repetition

- n permute k

1.7 ^nC_k Combinations

There are $^nC_k = \binom{n}{k} = \frac{n!}{(n-k)! \times k!}$ ways to *choose* k out of n distinct objects without repetition

- n choose k

1.8 Combinatorial Argument/Proof

- Let S be a set of objects
- Show how to count $|S|$ one way, let $|S| = M$
- Show how to count $|S|$ another way, let $|S| = N$
- Then $M = N$

1.9 Binomial Theorem

Let $x, y \in \mathbb{R}$ and $n \in \mathbb{N}$ a positive integer, then

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

- Symmetry in Binomial Coefficients $\binom{n}{k} = \binom{n}{n-k}$
- Pascal's Identity $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$
- Application of Binomial Theorem $\displaystyle\sum_{k=0}^{n} \binom{n}{k} = 2^n$

1.10 Inclusion-Exclusion

If $A_1, A_2, ..., A_N$ are sets, then

$$\begin{aligned}|A_1 \cup A_2 \cup ... \cup A_n| &= \text{singles} - \text{doubles} + \text{triples} - \text{quads} + ... \\ &= (|A_1| + ... + |A_n|) - (|A_1 \cap A_2| + ... + |A_{n-1} \cap A_n|) + ...\end{aligned}$$

1.11 Pigeonhole Principle

If there are n pigeons in $k < n$ holes, then one hole must contain at least $\left\lceil \dfrac{n}{k} \right\rceil$ pigeons

To use the Pigeonhole Principle

1. Identify pigeons
2. Identify pigeonholes
3. Specify how pigeons are assigned to pigeonholes
4. Apply Pigeonhole Principle

1.12 Sleuth's Criterion

For each object constructed, it should be possible to reconstruct the unique sequence of choices that led to it

- If an example has no sequence, then we are undercounting
- If an example has multiple sequences, then we are overcounting

2 Probability

2.1 Sample Space

A sample space Ω is the set of all possible outcomes of an experiment

2.2 Events

An event $E \subseteq \Omega$ is a subset of possible outcomes

- Events E and F are mutually exclusive if $E \cap F = \varnothing$

2.3 Probability Measure

A probability measure is a function $P : \omega \to [0, 1]$ such that

- $\mathbb{P}(\omega) \geq 0$ for all $\omega \in \Omega$
- $\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1$

2.4 Probability Space

A probability space is a pair $(\omega, \mathbb{P})$ where

- ω is a set called the sample space
- $\mathbb{P}$ is the probability measure

If $(\omega, \mathbb{P})$ is a probability space, then for any event $A \in \Omega$ it has probability $\mathbb{P}(A) = \sum_{\omega \in \mathcal{A}} \mathbb{P}(\omega)$

2.5 Uniform Probability Space

A uniform probability space is a pair $(\Omega, \mathbb{P})$ such that $\mathbb{P}(x) = \frac{1}{|\Omega|}$ for all $x \in \Omega$

If $(\omega, \mathbb{P})$ is a uniform probability space, then for any event $E \in \Omega$ it has probability $\mathbb{P}(E) = \frac{|E|}{|\Omega|}$

2.6 Axioms of Probability

1. Non-negativity: $\mathbb{P}(E) \geq 0$
2. Normalization: $\mathbb{P}(\Omega) = 1$
3. Countable Additivity: If E and F are mutually exclusive, then $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$

Corollaries of the axioms

1. Complementation: $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$
2. Monotonicity: If $E \subseteq F$, then $\mathbb{P}(E) \leq \mathbb{P}(F)$
3. Inclusion-Exclusion: $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$

2.7 Conditional Probability

The conditional probability of event A given an event B occurred is $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

- We can rearrange the equation such that $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B)$
- If A and B are independent events, then $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(A) \times \mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A)$

2.8 Bayes' Theorem

The probability of an event A, based on prior knowledge of conditions related to the event is $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}$

2.9 Partitions

Non-empty events $E_1, E_2, ..., E_n$ partition the sample space Ω if

$$E_1 \cup E_2 \cup ... \cup E_n = \bigcup_{i=1}^{n} E_i = \Omega$$

- The union of partitions cover the sample space
- The intersection of partitions is the null set

2.10 Law of Total Probability

If events $E_1, E_2, ..., E_n$ partition the sample space Ω, then for any event F

$$\mathbb{P}(F) = \mathbb{P}(F \cap E_1) + \mathbb{P}(F \cap E_2) + ... + \mathbb{P}(F \cap E_n) = \sum_{i=1}^{n} \mathbb{P}(F \cap E_i)$$

2.11 Chain Rule

$$\mathbb{P}(A_1 \cap ... \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2 \mid A_1) \cdot \mathbb{P}(A_3 \mid A_1 \cap A_2) \cdot ... \cdot \mathbb{P}(A_n \mid A_1 \cap ... \cap A_{n-1})$$

2.12 Independence

Two events A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cap \mathbb{P}(B)$

- If $\mathbb{P}(A) \neq 0$, then $\mathbb{P}(B \mid A) = \mathbb{P}(B)$
- If $\mathbb{P}(B) \neq 0$, then $\mathbb{P}(A \mid B) = \mathbb{P}(A)$
- Independent events with non-zero probabilities are never mutually exclusive

2.13 Conditional Independence

Two events A and B are independent conditioned on C if $\mathbb{P}(C) \neq 0$ and $\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid C) \cdot \mathbb{P}(B \mid C)$

- If $\mathbb{P}(A \cap C) \neq 0$, then $\mathbb{P}(B \mid A \cap C) = \mathbb{P}(B \mid C)$
- If $\mathbb{P}(B \cap C) \neq 0$, then $\mathbb{P}(A \mid B \cap C) = \mathbb{P}(A \mid C)$

3 Discrete Random Variables

3.1 Discrete Random Variables

A discrete random variable for a probability space $(\Omega, \mathbb{P})$ is a function $X : \Omega \to \mathbb{R}$

- Discrete random variables partition the sample space
  - Every event must have a probability
  - Every event has exactly one probability

3.2 Probability Mass Function (PMF)

The probability mass function of a discrete random variable $X : \Omega \to \mathbb{R}$ specifies, for any real number x, the probability that $X = x$

$$\mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = x\})$$

where $\{\omega \in \Omega \mid X(\omega) = x\}$ is the event space

3.3 Cumulative Distribution Function (CDF)

The cumulative distribution function of a random variable $X : \Omega \to \mathbb{R}$ specifies, for any real number x, the probability that $X \leq x$

$$F_X(x) = \mathbb{P}(X \leq x)$$

3.4 Expectation

Given a discrete random variable $X : \Omega \to \mathbb{R}$, the expectation or expected value of X is

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) \cdot \mathbb{P}(\omega) \text{ or equivalently } \mathbb{E}[X] = \sum_{x \in \Omega_X} x \cdot \mathbb{P}(X = x)$$

3.5 Linearity of Expectation

For any two random variables X and Y

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

- Linearity of expectations applies for both independent and dependent variables

For any random variables $X_1, X_2, ..., X_n$ and real numbers $a_1, a_2, ..., a_n \in \mathbb{R}$

$$\mathbb{E}[a_1X_1 + a_2X_2 + ... + a_nX_n] = a_1\mathbb{E}[X_1] + a_2\mathbb{E}[X_2] + ... + a_n\mathbb{E}[X_n]$$

3.6 Law of the Unconscious Statistician

Given a discrete real variable $X : \Omega \to \mathbb{R}$, the expectation or expected value of $Y = g(X)$ is

$$\mathbb{E}[Y] = \sum_{\omega \in \Omega} g(X(\omega)) \cdot \mathbb{P}(\omega)$$

or equivalently

$$\mathbb{E}[Y] = \sum_{\omega \in \Omega_X} g(x) \cdot \mathbb{P}(X = x)$$

or equivalently

$$\mathbb{E}[Y] = \sum_{y \in \Omega_y} y \cdot \mathbb{P}(Y = y)$$

3.7 Variance

The variance of a discrete real variable X is $\mathrm{Var}(X) = \sum_{x \in X} \mathbb{P}_X(x) \cdot (x - \mathbb{E}[X])^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

- $\mathrm{Var}(a \cdot X + b) = a^2 \cdot \mathrm{Var}(X)$
- $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$
- $\mathrm{Var}(X) = \mathrm{Var}(-X)$
- If X and Y are independent, then $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

3.8 Standard Deviation

The standard deviation of a discrete real variable X is $\sigma(X) = \sqrt{\mathrm{Var}(X)}$

3.9 Independent Random Variables

Two random variables X, Y are mutually independent if for all x, y

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$$

3.10 Discrete Uniform Random Variables

A discrete random variable X is equally likely to take any integer value between integers a and b inclusive, denoted $X \sim \mathrm{Unif}(a, b)$

- $\mathbb{P}(X = x) = \frac{1}{b-a+1}$
- $\mathbb{E}[X] = \frac{a+b}{2}$
- $\mathrm{Var}(X) = \frac{(b-a)(b-a+2)}{12}$

3.11 Bernoulli Random Variables

A Bernoulli random variable X takes value 1 with probability p, and value 0 with probability $1-p$, denoted $X \sim \mathrm{Ber}(p)$

- $\mathbb{P}(X = x) = \begin{cases} p & x = 1 \\ 1-p & x = 0 \\ 0 & \text{otherwise} \end{cases}$
- $\mathbb{E}[X] = p$
- $\mathrm{Var}(X) = p(1-p)$

3.12 Binomial Random Variables

A binomial random variable X is the number of successes in n independent random variables $Y_i \sim \mathrm{Ber}(p)$ where $X = \displaystyle\sum_{i=1}^{n} Y_i$, denoted $X \sim \mathrm{Bin}(n, p)$

- $\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$
- $\mathbb{E}[X] = np$
- $\mathrm{Var}(X) = np(1-p)$

3.13 Geometric Random Variables

A geometric random variable X models the number of independent trials $Y_i \sim \mathrm{Ber}(p)$ before seeing the first success, denoted $X \sim \mathrm{Geo}(p)$

- $\mathbb{P}(X = x) = (1-p)^{x-1} p$
- $\mathbb{E}[X] = \frac{1}{p}$
- $\mathrm{Var}(X) = \frac{1-p}{p^2}$

3.14 Negative Binomial Random Variables

A negative binomial random variable X models the number of independent trials $Y_i \sim \mathrm{Ber}(p)$ before seeing the r^{th} success. $X = \displaystyle\sum_{i=1}^{r} Z_i$ where $Z_i \sim \mathrm{Geo}(p)$, denoted $X \sim \mathrm{NegBin}(r, p)$

- $\mathbb{P}(X = x) = \binom{x-1}{r-1} (1-p)^{x-r} p^r$
- $\mathbb{E}[X] = \frac{r}{p}$
- $\mathrm{Var}(X) = \frac{r(1-p)}{p^2}$

3.15 Hypergeometric Random Variables

A hypergeometric random variable X measures the number of white balls you draw when you draw n balls uniformly at random from a total of N of which K are white and the rest are black, denoted $X \sim \mathrm{HypGeo}(N, K, n)$

- $\mathbb{P}(X = x) = \dfrac{\dbinom{K}{x}\dbinom{N-K}{n-x}}{\dbinom{N}{n}}$
- $\mathbb{E}[X] = n\frac{K}{N}$
- $\mathrm{Var}(X) = n\frac{K(N-K)(N-n)}{N^2(N-1)}$

3.16 Poisson Random Variables

A Poisson random variable X is the actual number of events happening given events happen independently at an average rate of λ per unit time, denoted $X \sim \mathrm{Poi}(\lambda)$

- $\mathbb{P}(X = x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}$
- $\mathbb{E}[X] = \lambda$
- $\mathrm{Var}(X) = \lambda$

The Poisson random variable $\mathrm{Poi}(np)$ well approximates the binomial random variable $\mathrm{Bin}(n, p)$ when n is large, p is small, and np is moderate

3.17 Sum of Independent Poisson Random Variables

Let $X \sim \mathrm{Poi}(\lambda_1)$ and $Y \sim \mathrm{Poi}(\lambda_2)$ such that $\lambda = \lambda_1 + \lambda_2$. Let $Z = X + Y$

- $\mathbb{P}(Z = z) = \dfrac{e^{-(\lambda_1+\lambda_2)}}{z!}(\lambda_1 + \lambda_2)^z$
- $\mathbb{E}[Z] = \lambda_1 + \lambda_2$

4 Continuous Random Variables

4.1 Probability Density Function (PDF)

A probability density function $f_X : \mathbb{R} \to \mathbb{R}$ represents a continuous random variable X

- $f_X(x) \geq 0$ for all $x \in \mathbb{R}$
- $\int_{-\infty}^{+\infty} f_X(x)\ dx = 1$
- $f_X(x)$ may be greater than 1

4.2 Cumulative Distribution Function (CDF)

The cumulative distribution of a continuous random variable X specifies, for any real number x, the probability that $X \leq x$

$$F_X(a) = \mathbb{P}(X \leq a) = \int_{-\infty}^{a} f_X(x)\ dx$$

4.3 Expectation

Given a continuous random variable X, the expectation or expected value of X is

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} f_X(x) \cdot x\ dx$$

4.4 Variance

The variance of a continuous random variable X is

$$\mathrm{Var}(X) = \int_{-\infty}^{+\infty} f_X(x) \cdot (x - \mathbb{E}[X])^2\ dx = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

4.5 Continuous Uniform Random Variables

A continuous uniform random variable X is denoted $X \sim \mathrm{Unif}(a, b)$

- $f_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$
- $\mathbb{E}[X] = \frac{b+a}{2}$
- $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$

4.6 Exponential Distribution

An exponential random variable X models the waiting time before the next event occurs given that λ events occur per unit time, denoted $X \sim \text{Exp}(\lambda)$

- $f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$
- $F_X(x) = 1 - e^{-\lambda x}$ for $x \geq 0$
- $\mathbb{E}[X] = \frac{1}{\lambda}$
- $\text{Var}(X) = \frac{1}{\lambda^2}$

4.7 Memoryless Random Variables

A random variable is memoryless if for all $s, t > 0$, $\mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t)$

- $X \sim \text{Exp}(\lambda)$ is a memoryless random variable

4.8 Normal Distribution

A normal random variable X with parameters $\mu \in \mathbb{R}$ and $\sigma \geq 0$ is denoted $X \sim \mathcal{N}(\mu, \sigma^2)$

- $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- $\mathbb{E}[X] = \mu$
- $\text{Var}(X) = \sigma^2$

Properties of the normal distribution

- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$
- $\text{Var}(aX + b) = a^2\text{Var}(X)$
- If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ where X and Y are independent, then
  $$aX + bY + c \sim \mathcal{N}(a\mu_x + b\mu_y + c,\ a^2\sigma_X^2 + b^2\sigma_Y^2)$$

4.9 Standard Unit Normal Distribution

The standard unit normal distribution Z is a normal random variable with parameters $\mu = 0$ and $\sigma^2 = 1$, denoted $Z \sim \mathcal{N}(0, 1)$

- $\mathbb{P}(Z \leq z) = \mathbb{P}(-z \leq Z) = \Phi(z)$
- $\mathbb{P}(z \leq Z) = \mathbb{P}(Z \leq -z) = 1 - \Phi(z)$

4.10 Standardizing Normal Distributions

Given a normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, the CDF of X is given by

$$\mathbb{P}\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

where $Z = \frac{X-\mu}{\sigma}$

4.11 Central Limit Theorem

Let $S_n = X_1 + ... + X_n$, where $X_1, ..., X_n$ are independent and identically distributed (iid) random variables each with expectation μ and variance σ^2

- $\mathbb{E}[S_n] = n\mu$
- $\mathrm{Var}(S_n) = n\sigma^2$

The CDF of $Y_n = \dfrac{S_n - n\mu}{\sigma\sqrt{n}}$ converges to the CDF of the standard unit normal $\mathcal{N}(0, 1)$

- $\mathbb{E}[Y_n] = 0$
- $\mathrm{Var}(Y_n) = 1$

Alternately, the CDF of $\bar{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ converges to the CDF of normal variable $\mathcal{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$

- $\mathbb{E}\left[\bar{X}\right] = \mu$
- $\mathrm{Var}\left(\bar{X}\right) = \dfrac{\sigma^2}{n}$

4.12 Continuity Correction

To estimate the probability that a discrete random variable lands in the integer interval $[a, b]$, compute the probability that the continuous approximation lands in the interval $\left[a - \frac{1}{2}, b + \frac{1}{2}\right]$

4.13 Minimum of IID Random Variables

If $Y_1, ..., Y_m$ are iid continuous uniform random variables $\mathrm{Unif}(0, 1)$, then $\mathbb{E}[\min(Y_1, ..., Y_m)] = \frac{1}{m+1}$

- Let $val = \min(Y_1, ..., Y_m)$. Then $m = \frac{1}{\mathbb{E}[val]} - 1$

4.14 Discrete Counting

Suppose we have an unknown number of iid random variables $Y_1, ..., Y_m$ and k independent hash functions $h_i : U \to [0, 1]$. Let $val_i = \min(h_i(Y_1), ..., h_i(Y_m))$. Then

$$\mathbb{E}[val] \approx \frac{1}{k}\sum_{i=1}^{k} val_i \text{ such that } m \approx \frac{1}{\frac{1}{k}\sum_{i=1}^{k} val_i} - 1$$

5 Joint Distributions

5.1 Joint Probability Mass Function

Let X and Y be discrete random variables. The joint probability mass function of X and Y is

$$p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$$

5.2 Joint Range

Let X and Y be discrete random variables. The joint range of $p_{X,Y}$ is

$$\Omega_{X,Y} = \{(x, y) \mid p_{X,Y}(x, y) > 0\}$$

where $(x, y) \subseteq \Omega_X \times \Omega_Y$

5.3 Joint Distributions of Independent Variables

Let X and Y be discrete random variables. X and Y are independent if and only if

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$$

for all $(x, y) \in \Omega_X \times \Omega_Y$

5.4 Marginal Probability Mass Function

Let X and Y be discrete random variables with joint probability mass function $p_{X,Y}(x, y)$. The marginal probability mass function of X is

$$p_X(x) = \sum_{y \in \Omega_Y} p_{X,Y}(x, y)$$

Similarly, the marginal probability mass function of Y is

$$p_Y(y) = \sum_{x \in \Omega_X} p_{X,Y}(x, y)$$

5.5 Additional Notes on Joint Distributions

	Discrete Random Variables	Continuous Random Variables
Joint PMF/PDF	$p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$	$f_{X,Y} \neq \mathbb{P}(X = x, Y = y)$
Joint range/support	$\{(x, y) \in \Omega_X \times \Omega_Y \mid p_{X,Y}(x, y) > 0\}$	$\{(x, y) \in \Omega_X \times \Omega_Y \mid f_{X,Y}(x, y) > 0\}$
Joint CDF	$F_{X,Y}(x, y) = \sum_{t \leq x} \sum_{s \leq y} p_{X,Y}(t, s)$	$F_{X,Y}(s, t) = \int_{-\infty}^{s} \int_{-\infty}^{t} f_{X,Y}(x, y)\, dy\, dx$
Normalization	$\sum_{x,y} p_{X,Y}(x, y) = 1$	$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy\, dx = 1$
Marginal PMF/PDF	$p_X(x) = \sum_{y \in \Omega_y} p_{X,Y}(x, y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$
Expectation	$\mathbb{E}[g(X, Y)] = \sum_{x,y} g(x, y) \cdot p_{X,Y}(x, y)$	$\mathbb{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) \cdot f_{X,Y}(x, y)\, dy\, dx$

6 Discrete and Continuous Random Variables

6.1 Conditional Expectation

Let X be a discrete random variable. Then the conditional expectation of X given event $Y = y$ is

$$\mathbb{E}[X \mid Y = y] = \sum_{x \in \Omega_X} x \cdot \mathbb{P}(X = x \mid Y = y)$$

6.2 Law of Total Probability for Discrete Variables

Let E be an event and let Y be a discrete random variable that takes values $\{1, 2, .., n\}$. Then

$$\mathbb{P}(E) = \sum_{i=1}^{n} \mathbb{P}(E \mid Y = i) \cdot \mathbb{P}(Y = i)$$

6.3 Law of Total Probability for Continuous Variables

Let E be an event and let Y be a continuous random variable. Then

$$\mathbb{P}(E) = \int_{-\infty}^{+\infty} \mathbb{P}(E \mid Y = y) \cdot f_Y(y)\ dy$$

6.4 Law of Total Expectation for Discrete Variables

Let X be a random variable and let Y be a discrete random variable that takes values $\{1, 2, ..., n\}$. Then

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X \mid Y = i] \cdot \mathbb{P}(Y = i)$$

6.5 Law of Total Expectation for Continuous Variables

Let X be a random variable and let Y be a continuous random variable. Then

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} \mathbb{E}[X \mid Y = y] \cdot f_Y(y)\ dy$$

7 Covariance

7.1 Covariance

The covariance of two random variables X and Y is

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

- If X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$
- If $X = Y$, then $\mathrm{Cov}(X,Y) = \mathrm{Var}(X) = \mathrm{Var}(Y)$

7.2 Covariance Matrix

The covariance matrix of a set of n random variables $X_1, ..., X_n$ is defined as

$$\Sigma = \begin{pmatrix} \mathrm{Cov}(X_1,X_1) & ... & \mathrm{Cov}(X_1,X_n) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n,X_1) & ... & \mathrm{Cov}(X_n,X_n) \end{pmatrix}$$

7.3 Multivariable Gaussian Distributions

A multivariable Gaussian distribution with parameters $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n\times n}$ is defined as

$$f(x \mid \mu, \Sigma) = \frac{1}{\sqrt{|2\pi\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

where $x \in \mathbb{R}^n$

8 Maximum Likelihood Estimation

8.1 Probability vs Likelihood

A probability function $\mathbb{P}(x \mid \theta)$ is a function with input being an event x for some fixed probability model with parameter θ

A likelihood function $\mathcal{L}(x \mid \theta)$ is a function with input being the parameter θ of the probability model for some fixed dataset x

8.2 Likelihood of Discrete Observations

The likelihood of independent discrete observations $x_1, ..., x_n$ is

$$\mathcal{L}(x_1, ..., x_n \mid \theta) = \prod_{i=1}^{n} \mathbb{P}(x_i \mid \theta)$$

8.3 Likelihood of Continuous Observations

The likelihood of independent continuous observations $x_1, ..., x_n$ is

$$\mathcal{L}(x_1, ..., x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

8.4 Log-Likelihood of Observations

The log-likelihood of independent observations $x_1, ..., x_n$ is

$$\mathcal{LL}(x_1, ..., x_n \mid \theta) = \ln \mathcal{L}(x_1, ..., x_n \mid \theta) = \sum_{i=1}^{n} \ln \mathbb{P}(x_i \mid \theta)$$

8.5 Maximum Likelihood Estimation

Given data $x_1, ..., x_n$, find $\hat{\theta}$ of model such that $\mathcal{L}(x_1, ..., x_n \mid \hat{\theta})$ is maximized

$$\hat{\theta} = \underset{\theta}{\arg\max}\ \mathcal{L}(x_1, ..., x_n \mid \theta)$$

To calculate $\hat{\theta}$

- Define the likelihood $\mathcal{L}(x_1, ..., x_n \mid \theta)$
- Compute the log-likelihood $\ln \mathcal{L}(x_1, ..., x_n \mid \theta)$
- Compute the first order derivative $\frac{d}{d\theta} \ln \mathcal{L}(x_1, ..., x_n \mid \theta)$
- Solve for $\hat{\theta}$ by determining the points with zero gradient
  - Ideally we want to calculate the second order derivative to verify that the point represents a maxima
  - For the purposes of this class, we can assume this point always represents a maxima

8.6 Likelihood of Observations From Gaussian Distribution

The likelihood of independent continuous observations $x_1, ..., x_n$ from Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ is

$$\mathcal{L}(x_1, ..., x_n \mid \theta_1, \theta_2) = \left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\theta_1)^2}{2\theta_2}}$$

where $\theta_1 = \mu$ and $\theta_2 = \sigma^2$

8.7 Log-Likelihood of Observations From Gaussian Distribution

The log-likelihood of independent observations $x_1, ..., x_n$ from Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ is

$$\mathcal{LL}(x_1, ..., x_n \mid \theta_1, \theta_2) = \ln \mathcal{L}(x_1, ..., x_n \mid \theta_1, \theta_2) = -n \cdot \frac{\ln(2\pi\theta_2)}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta_1)^2}{2\theta_2}$$

8.8 Maximum Likelihood Estimation for Observations From Gaussian Distribution

Given data $x_1, ..., x_n$ from Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, find $\hat{\theta}$ of model such that $\mathcal{L}(x_1, ..., x_n \mid \hat{\theta})$ is maximized

$$\hat{\theta_1}, \hat{\theta_2} = \underset{\theta_1, \theta_2}{\arg\max}\ \mathcal{L}(x_1, ..., x_n \mid \theta_1, \theta_2)$$

To calculate $\hat{\theta}_1, \hat{\theta}_2$

- Define the likelihood $\mathcal{L}(x_1, ..., x_n \mid \theta_1, \theta_2)$
- Compute the log-likelihood $\ln \mathcal{L}(x_1, ..., x_n \mid \theta_1, \theta_2)$
- Compute the first order partial derivatives
  - $\frac{\partial}{\partial \theta_1} \ln \mathcal{L}(x_1, ..., x_n \mid \theta_1, \theta_2) = \frac{1}{\theta_2} \sum_{i=1}^{n} (x_i - \theta_1)$
  - $\frac{\partial}{\partial \theta_2} \ln \mathcal{L}(x_1, ..., x_n \mid \theta_1, \theta_2) = -\frac{n}{2\theta_2} + \frac{1}{2{\theta_2}^2} \sum_{i=1}^{n} (x_i - \hat{\theta}_1)^2$
- Solve for $\hat{\theta}_1, \hat{\theta}_2$ by determining the points with zero gradient
  - $\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} x_i$
  - $\hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\theta}_1)^2$

8.9 Unbiased Estimators

An estimation $\hat{\theta}$ of parameter θ is an unbiased estimator if $\mathbb{E}[\hat{\theta}_n] = \theta$ for all $n \geq 1$

8.10 Consistent Estimators

An estimator $\hat{\theta}$ of parameter θ is consistent if $\lim\limits_{n\to\infty} \mathbb{E}[\hat{\theta}_n] = \theta$

- A consistent estimator is not necessarily unbiased
- Maximum likelihood estimators are always consistent

9 Markov Chains

9.1 Discrete-Time Stochastic Process

A discrete-time stochastic process is a sequence of random variables $X^{(0)}, X^{(1)}, X^{(2)}, ...$ where $X^{(t)}$ is the state at time t

9.2 Markov Chain

Markov chains are probabilistic finite automaton whose next state depends only on the current state and not on the history

9.3 Transition Probability Matrix

Given a Markov chain with n states, the transition probability matrix is an $(n \times n)$ matrix

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix}$$

where the element p_{ij} denotes the probability that the next state is j, given that the current state is i

- The row sum of a transition probability matrix is equal to 1

9.4 Probability Distribution of States

Given a Markov chain at time t with n states $1, ..., n$ and transition probability matrix

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} \qquad \text{where } p_{ij} = \mathbb{P}\left(X^{(t+1)} = j \mid X^{(t)} = i\right)$$

Let $q^{(t)} = \left({q_1}^{(t)}, ..., {q_n}^{(t)}\right)$ represent the probability distribution vector of the state at time $t+1$, where ${q_i}^{(t)} = \mathbb{P}(X^{(t)} = i)$. Then

$$q^{(t+1)} = q^{(t)} P$$

$$q^{(t+1)} = q^{(0)} P^{t+1}$$

- P^t converges to P as t goes to infinity
  - Values in the same column converge to the same value

9.5 Stationary Distributions

A stationary distribution $\pi = (\pi_1, ..., \pi_n)$ is a probability distribution of states where

$$\left({\pi_1}^{(t+1)}, ..., {\pi_n}^{(t+1)}\right) = \left({\pi_1}^{(t)}, ..., {\pi_n}^{(t)}\right)$$

that is, the probability distribution $\left({\pi_1}^{(\tau)}, ..., {\pi_n}^{(\tau)}\right)$ of the state no longer changes for all time $\tau \geq t$

9.6 Fundamental Theorem of Markov Chains

If a Markov chain is irreducible and aperiodic, then it has a unique stationary distribution

- A Markov chain is irreducible if for every state there exists a positive probability path to every other state
- A Markov chain is aperiodic if the time taken for the chain to loop back along different paths to a node are co-prime

9.7 Finite Markov Chains

Finite Markov chains are defined by a set of states and a transition probability matrix

- Consists of n states $1, ..., n$
- The state at time t is denoted by $X^{(t)}$
- Transition matrix P has dimension $(n \times n)$
- Has Markov property, where state at time t depends only on state at time $t - 1$

9.8 Markov's Inequality

Let X be a random variable taking only non-negative values. Then for any $t > 0$

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

10 Applications of Probability

10.1 Bloom Filters

- State
  - Stores information about a set of elements
- Behavior
  - `add(element)` adds a new element to the bloom filter
  - `contains(element)` returns true if the element is in the bloom filter, returns false otherwise
    * Prone to false positives
    * If returns false, then element is definitely not in the bloom filter
    * If returns true, then element is possibly in the bloom filter