CS 361 Note Sheet

Tommy Liu

Contents

1. Location and Scale Parameters

1.1. Location Parameters

1.1.1. Mean. Mean represents the center of balance, the average of the data. It is more sensitive to outliers and median.

$$\text{mean}(\{x_i\}) = \mu = \bar{x_i} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

1.1.2. Median & Mode. Median represents center-index value. If the length of not divisible by 2, take the half of the center two values. Median has the following properties:

- $\text{median}(\{x + c\}) = \text{median}(\{x\}) + c$
- $\text{median}(\{kx\}) = k\text{median}(\{x\})$

Mode represents the value that appears most frequently in a dataset. In a histogram, this is the peak, or sometimes we can have multiple peaks.

1.2. Scale Parameters

1.2.1. Standard Deviation. Standard Deviation represents how much the data spreads out with respect to mean. Following is the formula:

$$\text{std}(\{x_i\}) = \sigma^2 = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \text{mean}(\{x_i\}))^2}$$

Compare to variance, the standard deviation is more intuitive because it's in the same units as your data. $\text{std}(\{x_i\})$ has the following properties:

- $\text{std}(\{x_i + c\}) = \text{std}(\{x_i\})$
- $\text{std}(\{kx_i\}) = |k|\text{std}(\{x_i\})$

1.2.2. Variance. Variance represents the average of the squared differences from the mean. In other words, it measures whether the data set is more or less disperse.

$$\text{var}(\{x_i\}) = \sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i - \text{mean}(\{x_i\}))^2$$

And it has the following properties:

- $\text{var}(\{x + c\}) = \text{var}(\{x\})$
- $\text{var}(\{kx\}) = k^2\text{var}(\{x\})$

Note the variance of a constant would remain as 0, since there's difference

1.2.3. Inter-Quantile Range. IQR (Inter-Quantile Range) represents the range between the first quartile (25th percentile) and the third quartile (75th percentile).

$$IQR = Q_3 - Q_1$$

IQR has the following properties:

- $IQR(\{x + c\}) = IQR(\{x\})$
- $IQR(\{kx\}) = |k|IQR(\{x\})$

1.3. Standard Coordinate

Standard coordinates is about transforming data into a standardized scale so that different datasets can be compared directly. This transformation is called standardization. The resulting values are unit-less.

$$\hat{x_i} = \frac{(x_i - \text{mean}(\{x\}))}{\text{std}(\{x\})}$$

Standard coordinates has the following properties:

- $\text{mean}(\{\hat{x}\}) = 0$
- $\text{std}(\{\hat{x}\}) = 1$

For many kinds of data, histograms of these standard coordinates look the same.

- If mean = median, the data is symmetric.
- If mean < median, the data is left-skewed, or negatively skewed
- If mean > median, the data is right-skewed, or positively skewed

2. Correlation

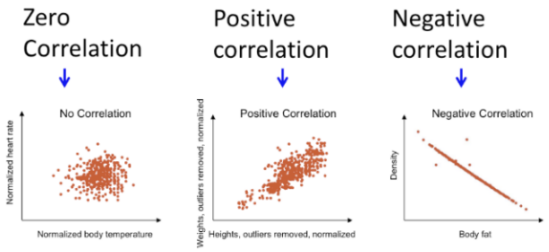
Correlation is a statistical measure that expresses the extent to which two variables are linearly related. Correlation **doesn't mean** causation, it means "tendency".

The correlation coeffecient r is calculated as:

$$r = \text{corr}(\{(x_i, y_i)\}) = \frac{1}{N}\sum_{i=1}^{N}\hat{x_i}\hat{y_i}$$
$$= \text{mean}(\{\hat{x_i}\hat{y_i}\}) = \sum_{i=1}^{N}\frac{\hat{x_i}}{\sqrt{N}}\frac{\hat{y_i}}{\sqrt{N}}$$

where the standardized x, y coordinates $\hat{x_i}$ and $\hat{y_i}$ are calculated as:

$$\hat{x_i} = \frac{x_i - \text{mean}(\{x_i\})}{\text{std}(\{x_i\})},\ \hat{y_i} = \frac{y_i - \text{mean}(\{y_i\})}{\text{std}(\{y_i\})}$$



- $\text{corr}(\{(x_i, y_i)\}) > 0$ shows positive correlation
- $\text{corr}(\{(x_i, y_i)\}) < 0$ shows negative correlation
- $\text{corr}(\{(x_i, y_i)\}) = 0$ shows no correlation

2.1. Properties of Correlation

Correlation value always in the range of $[-1, 1]$

- $\text{corr}(\{(x_i, y_i)\}) = 1 \iff \hat{x_i} = \hat{y_i}$
- $\text{corr}(\{(x_i, y_i)\}) = -1 \iff \hat{x_i} = -\hat{y_i}$

Correlation coefficient is symmetric.

$$\text{corr}(\{x, y\}) = \text{corr}(\{y, x\})$$

Scaling the data can change the sign, but not its absolute value.

$$\text{corr}(\{ax + b, cx + d\}) = \text{sign}(ab) \cdot \text{corr}(\{x, y\})$$

2.2. Prediction

Denote $\hat{y}^p = r + b$ to be a linear predictor, for r is the correlation coefficient

The prediction formula in standard coordinates is:

$$\hat{y_0}^p = r\hat{x_0}, \text{ where } r = \text{corr}(\{(x_i, y_i)\})$$

The prediction formula in original coordinates is:

$$\frac{y_0^p - \text{mean}(\{y_i\})}{\text{std}(\{y_i\})} = r\frac{x_0 - \text{mean}(\{x_i\})}{\text{std}(\{x_i\})}$$

2.3. Errors in Linear Predictor

There is an error term in prediction, which we can denote it as:

$$u_i = \hat{y_i} - \hat{y_i}^p$$

This is the difference between the actual observed values $\hat{y_i}$ and the predicted values $\hat{y_i}^p$ that are obtained using the linear predictor. We want to make mean of error equal to zero, and minimize the variance of error.

Root-mean-sqaure (RMS) prediction error let know how well the linear predictor predicts, i.e., a smaller RMS error indicates a better predictive model.

$$\text{RMS Error} = \sqrt{\text{mean}(\{u_i^2\})} = \sqrt{\text{var}(\{u_i\})} = \sqrt{1 - r^2}$$

3. Probability Fundamentals

- **Outcome** an outcome A is a possible result of a random repeatable experiment
- **Random** uncertain, non-deterministic
- **Sample Space** (Ω) is the set of all possible mutually exclusive outcomes associated with the experiment
- **Event** (E) An event is a subset of the sample space Ω

$$P(E) = \frac{\text{number of outcomes in } E}{\text{total number of outcomes in } \Omega}$$

- **Disjoint** (or **Mutually Exclusive**) if the events can not happen at the same time, they are disjoint or mutually exclusive
- **Independent events** are those where the occurrence of one event does not affect the probability of occurrence of the other event

3.1. Set

Union and intersection operation are commutative:

$$A \cup B = B \cup A, A \cap B = B \cap A$$

Union and intersection operation are associative:

$$(A \cap B) \cap C = A \cap (B \cap C)$$
$$(A \cup B) \cup C = A \cup (B \cup C)$$

Union and intersection operation are distributive:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

Complement:

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

Difference:

$$P(E_1 - E_2) = P(E_1) - P(E_1 \cap E_2)$$

Union:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

3.1.1. DeMorgan's Law.

$$(A \cup B)^C = A^C \cap B^C$$
$$(A \cap B)^C = A^C \cup B^C$$

3.2. Combination & Permutation

We denote c choose k as $\binom{c}{k}$, without regard to the order in which they are chosen, this is called a combination.

Combinations with distinct items:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Combinations with identical items:

$$\binom{n+k-1}{k} = \frac{(n+k-1)!}{k!(n-1)!}$$

Example 1 (from lecture): 4 gifts to 10 friends, each gets 1 gift, how many number of arrangement?
Answer: $\binom{10}{4}$

Permutations, on the other hand, are used when the order of arrangement matters.

Example 2 (from lecture): 10 gifts to 10 friends, each gets 1 gift, how many number of arrangement?
Answer: 10!

Sometimes, if it is not easy to calculate $P(E)$, we can check if the complement $P(E^c)$ is easy to calculate. That is, finding $1 - P(E^c)$ is the same as $P(E)$.

Example 3 (from lecture): A person may ride a bike on any day of the year equally. What's the probability that he/she rides on a Sunday or on 15th of a month?
Answer: The number of sunday in a year is 52. The number of 15th of a month is 12. Then:

$$P(E_1) = \frac{52}{365}, P(E_2) = \frac{12}{365}$$
$$P(E_1 \cap E_2) = \frac{2}{365}$$
$$P(E) = \frac{52+12-2}{365} = \frac{62}{365} \approx 16.99\%$$

3.2.1. Norms.

Standard set of poker: A "standard" deck of playing cards consists of 52 Cards in each of the 4 suits of Spades, Hearts, Diamonds, and Clubs. Each suit contains 13 cards: Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King. Modern decks also usually include two Jokers.

U.S. currency: A dime is 10 cents, a quarter is 25 cents.

3.3. Conditional & Joint Probability

$P(A \cap B)$ is called Joint Probability, it is the probability of both event A and event B occurring together. It's a measure of the likelihood that both events happen at the same time.

$P(A|B)$ is called Conditional Probability, it is the probability of event A occurring given that event B has already occurred.

They have the following relationship:

$$P(A|B) = \frac{P(A \cap B)}{P(B)},\ P(B) \neq 0$$

Note that whenever we are finding $P(B)$, we need to careful about summing up all the possibility of getting to the event B.

(See slides from lecture 5 & 6 for detailed example)

3.4. Bayers' Rule

Bayer's rule is about updating the probability of a hypothesis as more evidence becomes available.

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

or

$$P(A|B)P(B) = P(B|A)P(A)$$

Chain rule of joint events:

$$\begin{aligned} &P(A \cap B \cap C) \\ =&P(A \cap B \mid C) \cdot P(C) \\ =&P(A \mid B \mid C) \cdot P(B \mid C) \cdot P(C) \end{aligned}$$

3.5. Total Probability

The law of total probability is a fundamental rule relating **marginal probabilities** to **conditional probabilities**.

$$P(A) = \sum_n P(A \cap B_n) = \sum_n P(A|B_n)P(B_n)$$

where the A_m, A_n are disjoint, $A_m \cap A_n = \emptyset$ for $m \neq n$.

Given all of A, find B

$$\begin{aligned} P(B) &= P(A_1 \cap B) + P(A_2 \cap B) + P(A_2 \cap B) \\ &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) \\ &+ P(B|A_3)P(A_3) \end{aligned}$$

Given A and the complement of A, find B

$$\begin{aligned} P(B) &= P(A \cap B) + P(A^C \cap B) \\ &= P(B|A)P(A) + P(B|A^C)P(A^C) \end{aligned}$$

3.6. Independence

Two events are called independent events if either of the following is true:

$$P(A|B) = P(A) \text{ or } P(B|A) = P(B)$$

$$P(A \cap B) = P(A) \times P(B)$$

This implies whether A happened doesn't change the probability of B and vice versa.

3.6.1. Pairwise Independence. A set of events $\{A_1, A_2, \ldots, A_n\}$ is said to be pairwise independent if every pair of events is independent of each other. This means for any two events A_i and A_j where $i \neq j$, the following condition must hold:

$$P(A_i \cap A_j) = P(A_i) \times P(A_j)$$

Pairwise independence does not imply mutual independence among all events.

3.6.2. Mutual Independence. Mutual independence of a collection of events $A_1, A_2, A_3, \ldots, A_n$ is:

$P(A_i|A_jA_k \ldots A_p) = P(A_i)$, where $j, k, \ldots, p \neq i$

3.7. Random Variables

A random variable (RV) is a variable whose value is determined by the outcome of a random process or experiment. We are essentially mapping numbers to events, which is a function.

The values of a random variable can be either **discrete**, which are countable numbers; or **continuous**.

Some properties of RV:

- Random variables have **probability functions**
- Random variable can be **conditioned on events or other random variables**
- Random variables have **averages**

3.7.1. Discrete RV. Discrete RV takes on a countable number of distinct values. Examples include the number of heads in a series of coin flips, or the number of students who score a particular grade in a class.

3.7.2. Continuous RV. Takes on an uncountable number of values. For instance, the exact height of students in a school, or the time it takes for a chemical reaction to complete.

3.8. Probability Distribution Function

The probability distribution of a random variable describes how probabilities are assigned to each of its possible values.

$$\sum_x P(x) = 1$$

For discrete RV, we have **probability mass function** (PMF). Suppose X is a random variable, and the set of outcomes $\{(w_i \in \Omega)$ such that $X(w_i) = x_0\}$ is an event with probability is:

$$P(X = x_0) = \sum_i P(w_i)$$

For continuous RV, we have **probability density function**.

3.9. Cumulative Distribution Function

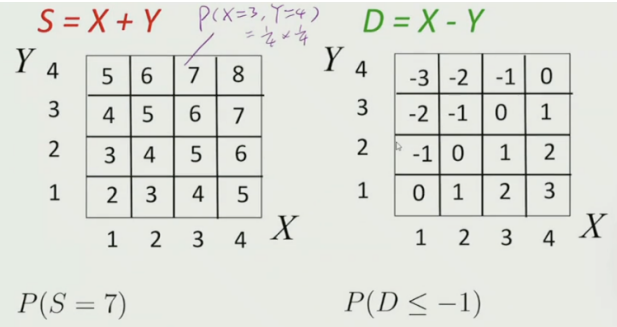
We define the following to be cumulative distribution function of a random variable X:

$$P(X \leq x)$$

where this is a non-decrease function.

3.10. Multiple Random Variable

For a combination of multiple random variable, we can oftenly use a "grid" to visualize the results. For two RVs X and Y, the header of column and row is all the possible values of each, and the grid elements is the results.



3.11. Probability for Random Variables

The conditional probability distribution of X given Y is:

$$P(X = x|Y = y), \text{ where } P(y) \neq 0$$
$$= \frac{P(X = x, X = y)}{P(Y = y)} = \frac{P(X = x \cap X = y)}{P(Y = y)}$$

3.11.1. Joint Probability of RV. The joint probability distribution of two random variables X and Y is:

$$P(X = x \cap Y = y), \text{ such that } \sum_x P(x|y) = 1$$

3.11.2. Marginal Probability of RV. We can recover the individual probability distributions from the joint probability distribution.

$$P(x) = \sum_y P(x, y)$$
$$P(y) = \sum_x P(x, y)$$

Then the sum of the joint probability distribution is:

$$\sum_y \sum_x P(X = x \cap Y = y) = 1$$

3.12. Independence of Random Variables

Random variable X and Y are independent if:

$$P(x, y) = P(x)P(y) \text{ for all } x \text{ and } y$$

$$X \text{ and } Y \text{ are independent } \Rightarrow E[XY] = E[X]E[Y]$$

$$E[XY] = E[X]E[Y] \not\Rightarrow X \text{ and } Y \text{ are independent}$$

3.13. Bayers' Rule for for Random Variables

Bayers rule for events generalizes to random variables:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} = \frac{P(y|x)P(x)}{\sum_x P(y|x)P(x)}$$

where $\sum_x P(y|x)P(x)$ is the total probability.

3.14. Expectation

The expected value of a random variable X is:

$$E[X] = \sum_x xP(X = x)$$

Note that the expected value is a weighted sum of all the values X can take.

3.14.1. Properties of Expected Value. One should note that the expected value of a constant is the constant itself, and that the expectation of a "expected value" is just that "expected value". Here are some properties of expectation:

$$E[kX] = kE[X]$$
$$E[kX + c] = kE[X] + c$$
$$E[X + Y] = E[X] + E[Y]$$

3.14.2. Expectation of a RV function. If f is a function of a random variable X, then $Y = f(X)$ is a random variable as well. Then the expected value of $Y = f(X)$ is:

$$E[Y] = E[f(X)] = \sum_x f(x)P(x)$$

For multiple random variables, we have:

$$E[f(X, Y)] = \sum_x \sum_y f(x, y)P(X = x \cap Y = y)$$

This trick is sometimes called "Blind statistician rule".

3.15. Variance of Random Variables

The variance of a random variable X is:

$$\text{var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

For standard deviation, we have $\text{std}[X] = \sqrt{\text{var}[X]}$

For random variable X and a constant k:

$$\text{var}[X] \geq 0$$
$$\text{var}[kX] = k^2\text{var}[X]$$

Some other properties:

- $\text{Var}[a] = 0$
- $\text{Var}[aX + b] = a^2 \cdot \text{Var}[X]$
- $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$
- $\text{Var}[X - Y] = \text{Var}[X] + \text{Var}[Y]$

The following holds only if X and Y are independent

$$\begin{aligned}\text{Var}[aX + bY] &= \text{Var}[aX] + \text{Var}[bY] \\ &= a^2 \cdot \text{Var}[X] + b^2 \cdot \text{Var}[Y]\end{aligned}$$

4. Covariance

The covariance of random variables X and Y is:

$$\begin{aligned}\text{cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y] \\ &= \sum_x \sum_y xyP(x, y) - E[X]E[Y]\end{aligned}$$

for X itself, this is:

$$\text{cov}(X, X) = \text{var}[X]$$

4.1. Correlation and Covariance

The covariance has a relationship with correlation:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

When X, Y takes on values with equal probability to generate data sets $\{x, y\}$, the correlation coefficient will be as seen before.

If two events are independent, their covariance is 0 and they are uncorrelated.

4.2. Variance and Covariance

The variance of the sum of two random variables is:

$$\begin{aligned}\text{var}[X + Y] &= \text{var}[X] + \text{var}[Y] + 2\text{cov}(X, Y) \\ &= E[(X + Y)^2] - (E[X + Y])^2\end{aligned}$$

4.3. Properties of Covariance

These are the 3 important properties of covariance, where, **if X, Y are uncorrelated**:

$$E[XY] = E[X]E[Y]$$
$$\text{cov}(X, Y) = 0$$
$$\text{var}[X + Y] = \text{var}[X] + \text{var}[Y]$$

5. (Weak) Law of Large Numbers

Law of Large Numbers (WLLN) states that for a sample mean $\bar{X}$:

$$\lim_{N \to \infty} P(|\bar{X} - E[\bar{X}]| > \epsilon) = 0, \forall \epsilon > 0$$

In other words, it means that:

1. $\bar{X}$ and $E[\bar{X}]$ should be very close to each other
2. $\bar{X}$, $\lim_{N \to \infty} P(|\bar{X} - E[\bar{X}]| < \epsilon) = 1$

Below are the two inequalities that proves this.

5.1. Markov's Inequality

For any random variable X that only takes $x \geq 0$ and constant $a > 0$, we have:

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

For example, if $a = 10E[X]$, then:

$$P(X \geq 10E[X]) \leq \frac{E[X]}{10E[X]}$$

5.2. Chebyshev's Inequality

Chebyshev's Inequality states that the probability that X is greater than k standard deviation away from the mean is small.

For any random variable X and constant $a > 0$:

$$P(|X - E[X]| \geq a) \leq \frac{\text{var}[X]}{a^2}$$

If we let $a = k\sigma$ where $\sigma = \text{std}[X]$:

$$P(|X - E[X]| \geq k\sigma) \leq \frac{1}{k^2}$$

5.2.1. Range of data points. For any set of N data points x_i with a standard deviation σ, Chebyshev's inequality tells us that the number of data points that are at least k standard deviations away from the mean is at most $\frac{N}{k^2}$.

Example 1 (from lecture): Estimate as close as possible, 90% data in a standardize coordinate is within what range?
Answer: This would be meaning 10% data is not in the range, of which we can have $\frac{1}{k^2} = \frac{1}{10}$, solve this we get $k = \sqrt{10} \approx 3.16$

6. All Types of Probability Distributions

6.1. Bernoulli Distribution

It is a binary choice distribution, with which one choice is p, and another choice is $1 - p$; only two outcomes. Its expected value and variance are defined as follows:

$$E[X] = p$$

$$\text{Var}[X] = p(1-p)$$

6.2. Binomial Distribution

It is a series of **bernoulli distribution** experiments with N number of times

Example: In a series of 10 basketball free throws, where a player has a 70% chance of making each shot (Bernoulli trials), what is the probability of making exactly 7 shots in those 10 attempts?

Its probability, expected value, and variance are:

$$P(X = k) = \binom{N}{k} p^k (1-p)^{N-k}$$

$$E[X] = N \cdot p$$

$$\text{Var}[X] = Np(1-p)$$

6.3. Geometric Distribution

For experiments that are waiting for a certain event to occur.

Example: if you are flipping a coin until you get heads (which has a probability of 0.4 of occurring on any given flip), what is the expected number of coin flips until you see the first heads?

Its probability, expected value, and variance are:

$$P(X = k) = (1-p)^{k-1} p$$

$$E[X] = \frac{1}{p}, \ \text{Var}[X] = \frac{1-p}{p^2}$$

where k is the total number of trials, including the success trial.

6.4. Poisson Distribution

Used for modeling the number of events occurring within a fixed interval of time or space when the events are rare and random. Poisson distribution is a discrete distribution.

Example: in a call center, calls arrive at an average rate of 5 calls per minute. What is the probability that exactly 8 calls will be received in a given minute?

Its probability, expected value, and variance are:

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

where λ is the average rate of event occurrences.

$$E[X] = \lambda$$

$$\text{Var}[X] = \lambda$$

6.5. Exponential Distribution

Used to model the time between events in a Poisson process, where events occur at a constant average rate; and the events occur continuously over time at a constant average rate. In Exponential Distribution, we assume that failures form a Poisson process in time, then the time to the next failure is exponentially distributed.

Example: In a manufacturing process, electronic components fail on average every 200 hours. What is the probability that a component will fail within the first 100 hours?

Its probability density function, expected value, and variance are:

$f(x; \lambda) = \lambda e^{-\lambda x}$, where λ is the rate parameter

$$E[X] = \frac{1}{\lambda}$$

$$\text{Var}[X] = \frac{1}{\lambda^2}$$

6.6. Continuous Distribution

This is a general form of continuous distribution. The probability of any single exact value is zero; instead, probabilities are associated with intervals.

Its PDF $p(x)$ and must satisfy two conditions:

- The function must be non-negative for all x, that is $p(x) \geq 0$.
- The total area under the curve of the function and above the x-axis is equal to 1, that is $\int_{-\infty}^{\infty} p(x)dx = 1$.

Its expected value and variance are defined:

$$E[X] = \int_{-\infty}^{\infty} xp(x)dx$$

$$\text{Var}[X] = \int_{-\infty}^{\infty} (x - E[X])^2 p(x)dx$$

6.6.1. Continuous Uniform Distribution. All values in the distribution are equally likely to occur over a specified interval.

> **Example** if a random number between 1 and 10 is chosen uniformly, what is the probability that it is between 3 and 7?

Its probability density function, expected value, and variance are:

$$f(x; a, b) = \frac{1}{b-a} \text{ for } a \leq x \leq b, \text{ and 0 elsewhere}$$

where a and b are the lower and upper bounds of the interval.

$$E[X] = \frac{a+b}{2} = \mu$$

$$\text{Var}[X] = \frac{(b-a)^2}{12}$$

6.7. Normal Distribution

A continuous probability distribution that is often used to model real-world data, such as heights, test scores, and errors.

Its probability density function, expected value, and variance are:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$E[X] = \mu$$

$$\text{Var}[X] = \sigma^2$$

This is general form of the normal distribution, where if $\frac{x-\mu}{\sigma}$ has a standard normal distribution, then the $p(x)$ that has the mean μ and variance σ^2 are also normal distributions.

6.7.1. Standard Normal Distribution. This is a special type of normal distribution, where its mean is $\mu = 0$ and variance is $\sigma^2 = 1$. Its probability density function, expected value, and variance are:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x)^2}{2}}$$

$$E[X] = \mu$$

$$\text{Var}[X] = \sigma^2$$

6.7.2. Properties of Normal Distribution. The **Central Limit Theorem** (CLT) tells us that, for a large enough sample size, the distribution of the sample mean will be approximately normally distributed, no matter the shape of the population distribution.

6.7.3. Approximation with Normal Distribution. By using what CLT implies, we can approximating the Binomial Distribution for a very large number N.

7. Sample Statistics

Sample mean $\bar{X}$ is a random variable.

1. popmean($\{X\}$) is the population (or true) mean
2. popsd($\{X\}$) is the w/rt the whole population
3. $X^{(N)}$ is the mean of N amount of samples

Here are some definition regarding the sample mean, variance, and standard deviation:

$$E[X^{(N)}] = \text{popmean}(\{X\})$$

$$\text{var}[X^{(N)}] = \frac{\text{popsd}(\{X\})^2}{N} = \frac{\text{popvar}(\{X\})}{N}$$

$$\text{std}[X^{(N)}] = \frac{\text{popsd}(\{X\})}{\sqrt{N}}$$

7.1. Standard Unbiased

Standard Unbiased represents for an unbiased estimate of the population standard deviation. It can be write as follows:

$$\text{stdunbiased}(\{x_i\}) = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \text{mean}(\{x_i\}))^2}{N-1}}$$

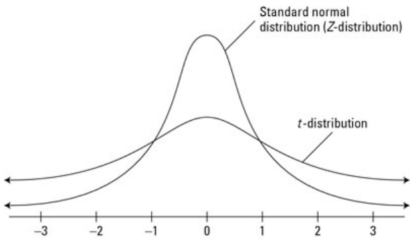
Notice the $N-1$ instead of the N from the formal equation; if N is large, this number is basically the same as the number we compute for population standard deviation.

7.2. Standard Error

Standard error estimate how much variability exists between the sample mean and the population mean.

$$\text{stderr}(\{x_i\}) = \frac{\text{stdunbiased}(\{x_i\})}{\sqrt{N}}$$

7.3. t-distribution & Z-distribution



Small Samples (t-distribution): the t-distribution is wider and has thicker tails than the normal distribution, meaning there's more uncertainty in your estimate of the population mean. The following distribution has a t-distribution with $N-1$ degrees of freedom.

$$T = \frac{\text{mean}(\{x\}) - \text{popmean}(\{X\})}{\text{stderr}(\{x\})}$$

Large Samples (Z-distribution): as your sample size gets bigger, the t-distribution looks more and more like the normal distribution. This is because with more data, your estimate of the population mean gets more reliable. Commonly when $N > 30$ is used a rule of thumb.

$$Z = \frac{\text{mean}(\{x\}) - \text{popmean}(\{X\})}{\text{stderr}(\{x\})}$$

In particular, this becomes a standard normal distribution, where its PDF is $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

7.4. Confidence Interval

A confidence interval gives us a range of values, which is likely to contain the true population parameter (like the mean) with a certain level of confidence.

The confidence interval is given by, which approximating the true population mean:

$$[\text{mean}(\{x\}) - b \times \text{stderr}(\{x\}), \text{mean}(\{x\}) + b \times \text{stderr}(\{x\})]$$

To find the the uncommon factor b, which is b standard errors away from the center mean, we can:

1. Since $\text{CDF}(\mu + b\sigma) = 1 - \alpha$
2. Then $\alpha = (1 - c\%)/2$
3. Then we find b for $\text{CDF}(b) = 1 - \alpha$

Easy confidence intervals

- For about 68% of sample, b is 1
- For about 95% of sample, b is 2
- For about 99% of sample, b is 3

Example A 65% confidence interval means that if you were to take 100 different samples (more than 1 sample, and a sample contains multiple data points) and compute a confidence interval for each sample, then approximately 65 of the 100 confidence intervals will contain the true population parameter. It does not mean that there is a 65% chance that any particular calculated interval contains the population parameter.

7.4.1. Finding CDF. If the n in $CDF(n)$ exceeds to more than 4, we can consider $CDF(n) = 1$.

In general, for a positive Z value n, e.g., $n = -1$:

$$CDF(-n) = 1 - CDF(n)$$

Similarly, to find the probability in a continuous distribution that is not smaller than:

$$P(X > 2) = 1 - P(X \leq 2)$$

7.4.2. Quantile. A quantile is a value that divides a probability distribution into continuous intervals with equal probabilities.

A value y of the quantile of $\frac{1}{3}$ means that, y gives us the probability of $\frac{1}{3}$. Thus we want to find y for:

$$\text{CDF}(y) = \frac{1}{3}$$

7.5. Bootstrap Simulation

We use Bootstrap Simulation to estimate the distribution of a statistic (like the mean or standard deviation) by re-sampling **with replacement** from data.

Draw a Sample → Calculate Stat. → Repeat → Analyze

Sample Size is the size of each bootstrap sample.

Replicates (Number of Bootstrap Samples) is the number of times we repeat the re-sampling process.

7.5.1. Bootstrap errors. The distribution simulated from bootstrapping is called empirical distribution. It is not the true population distribution.

The number of bootstrapping replicates may not be enough, there is a numerical error.

When the statistic is not a well behaving one, such as maximum or minimum of a data set, the bootstrap method may fail to simulate the true distribution

7.6. Hypothesis Test

H_0 is the null hypothesis, H_1 is some hypothesis that justify a probability value, e.g., $P(X > 5) = 55\%$

We are interested in whether to reject a hypothesis H or not, given the data. Define a test statistics x:

$$x = \frac{\text{mean}(\{x\}) - H}{\text{stderr}(\{x\})}$$

If the sample size N is greater than 30, we can assume that x comes from a standard normal distribution. Then if x lies in the rejection region (or the "extreme fraction"), we reject the hypothesis.

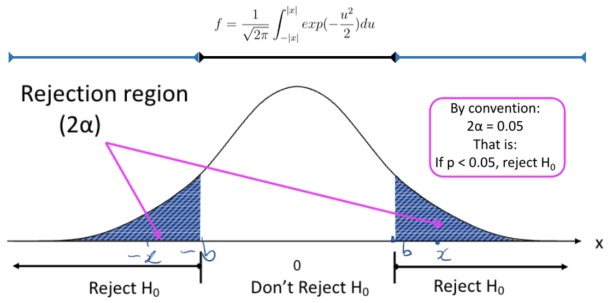
Generally, we can check whether:

$$\text{CDF}(x) < 0.05$$

7.6.1. p-value. The p-value of the a test statistics x can be calculated with the area under curve:

$$p = 1 - f = 1 - \frac{1}{\sqrt{2\pi}} \int_{-|x|}^{|x|} e^{-\frac{u^2}{2}} du$$

If $p < 0.05$, reject H_0



8. Maximum Likelihood Estimation

The maximum likelihood estimation function is:

$$\mathcal{L}(\theta) = P(D \mid \theta) = \prod_{i \in \text{dataset}} P(\mathbf{x}_i \mid \theta)$$

Essentially, we are looking for the value of θ that maximizes the function $P(D|\theta)$. That is, we estimate $\hat{\theta}$:

$$\hat{\theta} = \underset{x}{\text{argmax}} \mathcal{L}(\theta), \text{ or we setup } \frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0$$

Caution: The data must be IID in the estimation, that is, each data item is an independently obtained sample from the same probability distribution.

8.1. MLE with Binomial Model

For a binomial distribution, the MLE of θ can be directly derived from the following formula:

$$\hat{\theta} = \frac{k}{N}$$

8.2. MLE with Geometric Model

For a geometric distribution, this is:

$$\hat{\theta} = \frac{1}{N}$$

8.3. Log Likelihood

When the computation is heavy, we can use log-likelihood to gain better perfomance when estimation. That is:

$$\log(\mathcal{L}(\theta)) = \log(P(\mathcal{D}|\theta)) = \sum_{i \in \text{dataset}} \log(P(d_i|\theta))$$

8.4. MLE with Poisson Model

For an Poisson distribution, this is:

$$\hat{\theta} = \frac{\sum_i x_i}{N}$$

8.5. MLE with Exponential Model

For an exponential distribution, this is:

$$\hat{\theta} = \frac{N}{\sum_i x_i}$$

8.6. MLE with Normal Model

For a normal distribution, this is:

$$\hat{\theta} = \frac{\sum_{i=1}^{N} x_i}{N} = \mu = \text{mean}(\{x\})$$

8.7. Drawbacks of MLE

- Maximizing some likelihood (or even log-likelihood) function is mathematically hard
- If there are few data items, the MLE estimate maybe very unreliable

9. Maximum A Posterior Estimate

From Bayers' Rule, given the unknown parameter θ and the dataset D, we have:

$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Normalizing Constant}}$$

- $P(\theta \mid D)$ is the **posterior probability**
- $P(D \mid \theta)$ is the **likelihood** $\mathcal{L}(\theta)$
- $P(\theta)$ is the **prior probability** of the θ
- $P(D) = \sum_i P(D \mid \theta_i)P(\theta_i)$

where, θ_i are disjoint, and θ is discrete, we can use the above summation to find $P(D)$.

In MAP, we focus on maximizing $P(D \mid \theta)P(\theta)$, that is, we want to use θ to maximize the posterior $P(\theta \mid D)$:

$$\hat{\theta} = \underset{x}{\text{argmax}}\ P(\theta \mid D)$$

> Note that the following proportional relationship does not include the normalizing constant (the denominator in Bayes' Theorem $P(D)$), which ensures that the posterior probabilities sum to 1. In MAP estimation, this constant is not necessary for the final calculation.
>
> $$P(\theta \mid D) \propto P(D \mid \theta)P(\theta)$$

9.0.1. Drawbacks of MAP.

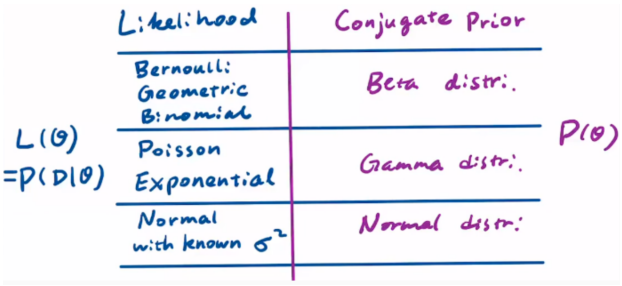
- Maximizing some posteriors $P(\theta \mid D)$ is difficult
- Some choices of prior $P(\theta)$ can overwhelm any data observed
- It's hard to justify a choice of prior

10. Bayesian Posterior

10.1. Conjugacy

For a given likelihood function $P(D \mid \theta)$, a prior $P(\theta)$ is its conjugate prior if it has the following properties:

- $P(\theta)$ belongs to a family of distributions that are expressive
- The posterior $P(\theta \mid D)$ belongs to the same family of distribution as the prior $P(\theta)$
- The posterior $P(\theta \mid D)$ is easy to maximize



10.2. Beta Distribution

A distribution is **Beta distribution** if it has the following PDF:

$$P(\theta) = \begin{cases} K(\alpha, \beta)\theta^{\alpha-1}(1-\theta)^{\beta-1} & 0 \le \theta \le 1, \alpha > 0, \beta > 0 \\ 0 & \text{otherwise} \end{cases}$$

> When $\alpha = \beta = 1$, then $K(1,1) = 1$, and $P(\theta) = \text{Beta}(1,1) = 1$ is continuous uniform distribution.

> The term $K(\alpha, \beta)$ is constant, a function of α and β, defined as:
>
> $$K(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$
>
> Where Γ is the gamma function, defined as $\Gamma(x) = (x-1)!$, and $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$.

The MLE for Beta distrition is:

$$\hat{\theta} = \frac{\alpha - 1 + k}{\alpha + \beta - 2 + N}$$

10.3. Gamma Distribution

A non-negative continuous random variable X has a Gamma distribution if its PDF is:

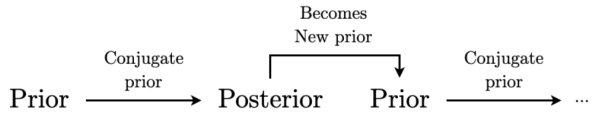
$$P(X = x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$$
$$x > 0, \alpha > 0, \beta > 0$$

A Gamma distribution with parameter α, β has:

$$\mu = \frac{\alpha}{\beta}, \ \ \sigma = \frac{\alpha}{\beta^2}$$

10.4. Update of Bayesian Posterior

Since the posterior is in the same family as the conjugate prior, the posterior can be used as a new prior if more data is observed.



11. Covariance Matrix

Denote the following:
- $\{\mathbf{x}\}$ is a dataset of N number of data items
- $\mathbf{x}_i$, this is a d-dimensional vector
- j-th component of $\mathbf{x}_i$ as $\mathbf{x}_i^{(j)}$
- The matrix for the data set $\{\mathbf{x}\}$ is of $d \times N$ shape

Then the Covariance Matrix is defined as:

$$\text{Covmat}(\{x\}) = \frac{X_c X_c^T}{N}$$

Where:

$$X_c = X - \frac{\sum_i \mathbf{x}_i(j)}{N}$$

Some notes:
- The covariance matrix a $d \times d$ matrix
- The biggest variance is on PC1, where PC1 is the vector that points in the direction of the maximum variance in the data.

> Covariance is the un-normalized correlation. Correlation is the normalized covariance with a positive value as factor. For a data set with dimension d, the number of possible co-variances is, choose 2 out of d, i.e., $\binom{d}{2}$.

11.1. Symmetricity

- Covariance matrix is symmetric

$$\text{cov}(\{x\}; j, k) = \text{cov}(\{x\}; k, j)$$

- The covariance matrix is positive and semi-definite; all of its eigenvalues are greater or equal to 0
- Covariance matrix is diagonalizable

11.2. Diagonal & Off-Diagonal Elements

- Each diagonal value is the variance for each feature, it is also the corresponding eigenvalue, and that for each $\sigma^2 \leq 0$
- If non-diagonal value is 0, means no correlations

11.3. Diagonalization

Suppose M is the covariance matrix, U is a orthogonal matrix, and D is the diagonalizable matrix:

$$D = U^{-1} M U$$

we can solve for M by:

$$M = UDU^{-1}, \ \text{ or } M = UDU^T \text{ if } D \text{ is orthonormal}$$

Note that D contains the eigenvalues of M, to find any PC_i, identify the i-th column of U.

11.4. Mean Squared Error (MSE)

(Lecture 19 p. 8) MSE is a measure of the difference between values predicted by a model and the values actually observed:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Given a diagonalized matrix (consists of eigenvalues) of a covariance matrix, suppose the eigenvalues are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$.

The percentage of variance covered by λ_i is:

$$\frac{\lambda_i}{\sum_n \lambda_n}$$

And the MSE of removing λ_i is λ_i

11.5. Other Properties

$$\text{Covmat}(\{x\} + c) = \text{Covmat}(\{x\}) + c$$

$$\text{Covmat}(\{Ax\}) = A \ \text{Covmat}(\{x\}) A^T$$

12. Classification

Classification is a kind of supervised learning.

12.1. Classifier

We use **class confusion matrix** for training and testing, and use **cross-validation** to prevent **overfitting**.

12.1.1. Class Confusion Matrix. A "confusion matrix" is a table or matrix used in classification tasks in machine learning and statistics to visualize the performance of an algorithm. It is for **supervised learning**.

The confusion matrix is typically arranged as follows:

	Predicted Pos.	**Predicted Neg.**
Actual Pos.	TP	FN
Actual Neg.	FP	TN

- True Positives (TP): The number of positive instances correctly predicted as positive.
- False Positives (FP): The number of negative instances incorrectly predicted as positive (also known as Type I error).
- True Negatives (TN): The number of negative instances correctly predicted as negative.
- False Negatives (FN): The number of positive instances incorrectly predicted as negative (also known as Type II error).

$$\begin{bmatrix} \mathrm{TP}^{1,1} & \mathrm{FP}^{1,2} & \mathrm{FP}^{1,3} \\ \mathrm{FP}^{2,1} & \mathrm{TP}^{2,2} & \mathrm{FP}^{2,3} \\ \mathrm{FP}^{3,1} & \mathrm{FP}^{3,2} & \mathrm{TP}^{3,3} \end{bmatrix}$$

Above is an example 3×3 matrix representation, for entries like $TP^{1,1}$ indicates that it's the count of True Positives for Class 1 predicted as Class 1, and $FP^{1,2}$ indicates False Positives for Class 1 predicted as Class 2, and so on.

Accuracy of the confusion matrix is given by:

$$\frac{\text{Diagonal elements}}{\text{Sum of all matrix elements}}$$

Precision for each class i is given by:

$$\frac{\mathrm{TP}^{i,i}}{\mathrm{TP}^{i,i} + \mathrm{FP}^{\forall j \neq i,i}}$$

12.1.2. Cross-validation. We use k-fold (oftenly $k = 5$) to split the data into two parts. For $k = 5$, we have 20% for testing and 80% for training.

12.1.3. Entropy. The Shannon entropy is the measure of uncertainty for a general distribution. If class i contains a fraction $P(i)$ of the data, we need $\log_2 \frac{1}{P(i)}$ bits for that class. Then the entropy $H(D)$ of a data set is defined as the **weighted mean** of entropy for every class, and $H(D)$ is given by:

$$\begin{aligned} H(D) &= \sum_{i=1}^{c} P(i) \log_2 \frac{1}{P(i)} \\ &= \sum_{i=1}^{c} (-P(i) \log_2(P(i))) \end{aligned}$$

12.1.4. Choosing Classifiers.

Some criteria to consider in choosing the classifier:

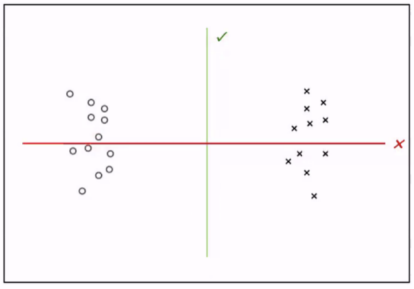
- Accuracy
- Speed, such as testing and classifying
- Flexibility, such as variety of data
- Interpretation, such as decision boundary
- Scaling effect

12.1.5. Problem with Classifiers.

- It may not do well in the real data that was used in training, i.e., it's possible that the training data is "bad"
- **Over-fitting** model is too complex, performs well on the training data but poorly on unseen or test data.
- **Under-fitting** model is too simple to capture the underlying structure of the data, performs poorly on new, unseen data.

12.2. Decision Tree

A tree-structured model that represents decisions and their possible consequences, including chances of event occurrence, resource costs, and utility.



The steps for decision tree are:

1. Choose a dimension or feature, and a "split"
2. Split the training data into left-child (D_L) and right-child (D_R) subsets
3. Repeat the first two steps above recursively on each child
4. Stop the recursion based on the some conditions
5. Label the leavs with class labels

12.3. Random Forest

A Random Forest operates by constructing a multitude of decision trees at training time. We build the random forest by training each decision tree on a random subset with replacement from the training data and subset of features are also randomly selected ("Bagging").

12.4. Random Forest v. Decision Tree

Random Forest generally has higher Accuracy, but could resulting in over-fitting issue. While Decision Trees has better interpretability and simple to implement, but may be biased with imbalanced data or can also have over-fitting issue.

12.5. Support Vector Machine

We want to find a hyperplane in an N-dimensional space (N is the number of features) that distinctly classifies the data points.

12.5.1. Decision Boundary. SVM uses a hyperplane as its decision boundary, the decision boundary is given by:

$$a_1x^{(1)} + a_2x^{(2)} + \cdots + a_dx^{(d)} + b = 0$$

$$\text{or } \text{sign}(\mathbf{a}^T\mathbf{x} + b) = 0$$

12.5.2. Loss Function.

We define a hinge loss function to be:

$$\max(0, 1 - y_i(\mathbf{a}^T\mathbf{x}_i + b))$$

With a regularization penalty ($||\mathbf{a}||^2 = \mathbf{a}^T\mathbf{a}$), the training error cost is then:

$$S(\mathbf{a}, b) = \left[\frac{1}{N}\sum_{i=1}^{N}\max(0, 1 - y_i(\mathbf{a}^T\mathbf{x}_i + b))\right] + \lambda(\frac{\mathbf{a}^T\mathbf{a}}{2})$$

where λ is the **regularization parameter**, it trade off between these two objectives.

12.5.3. Convex Function. A function $f(x)$ is called convex on an interval if the line segment between any two points on the graph of the function lies above or on the graph.

12.5.4. Gradient Descent.

The loss (Q) in GD is defined as:

$$f(\vec{\mathbf{a}}) = \frac{1}{k}\sum_{j=1}^{k} Q(\vec{\mathbf{a}}, j) + \text{penalty}$$

$$\vec{\mathbf{a}}_{n+1} = \vec{\mathbf{a}}_n - \eta\nabla f(\vec{\mathbf{a}}_n)$$

12.5.5. Stochastic Gradient Descent.

The loss (Q) in SGD is defined as:

$$f(\vec{\mathbf{a}}) = \frac{1}{m}\sum_{j=1}^{m} Q(\vec{\mathbf{a}}, i) + \text{penalty}$$

$$\vec{\mathbf{a}}_{n+1} = \vec{\mathbf{a}}_n - \eta\nabla g(\vec{\mathbf{a}}_n)$$

where g is a function $g(\vec{\mathbf{a}}) = f(\vec{\mathbf{a}}) + z$, z is some noise.

12.6. Naive Bayes Classifer

This is a probabilistic machine learning model used for classification tasks. The "naive" assumption in Naive Bayes is the **conditional independence of features** given the class label.

The training is a MAP estimator of class variable y given the data x:

$$\underset{y}{\text{argmin}}\ P(x|y)P(y) = \underset{y}{\text{argmin}}\ \left[\prod_{j=1}^{d} P(x^{(j)}|y)\right] P(y)$$

and we can also use log likelihood to solve for this heavy production.

(Lecture 21 p. 40) We can use either (1) Normal distribution, (2) Poisson distribution, or (3) Bernoulli distribution to model $P(x^{(j)}|y)$:

12.6.1. Poisson Dist. Model.

Find the λ_{MLE} for each $P(x^{(j)}|y = y_i)$:

$$\lambda_{\text{MLE}} = \frac{\sum_i x_i^{(j)}}{N_{\text{total number of } y = y_i}}$$

Then $P(x^{(j)}|y)$ is found by plugging into the PDF of Poisson distribution.

12.6.2. Normal Dist. Model.

Find the μ_{MLE} and σ_{MLE} for each $P(x^{(j)}|y = y_i)$

Then $P(x^{(j)}|y)$ is found by plugging into the PDF of Normal distribution.

12.6.3. Advantages.

- Simple and easy to implement.
- Efficient on large datasets.
- Performs well with categorical input features.

12.6.4. Disadvantages.

- Relies on assumption of feature independence.
- May not perform well with correlated features.
- Less effective with continuous features unless they are Gaussian distributed.

13. Linear Regression

Linear Regression is a kind of supervised learning. Unlike classifier that deals with categorical value data (as label), linear regression deals with real value data.

Suppose the dataset $\{(\mathbf{x}, y)\}$ consists of N labeled items $(\mathbf{x}_i, y_i)$. If we represent the dataset as a table, then:

$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	y
1	2	0
2	3	2
3	6	5

- **explanatory variables** $\mathbf{x}^{(j)}$ is the d columns representing $\{\mathbf{x}\}$
- **dependent variable** is the numerical colm. y

13.1. Linear Model

We want to model y as a linear function of $\mathbf{x}^{(j)}$, with some randomness. This is given by:

$$\begin{aligned} y &= \mathbf{x}^{(1)}\beta_1 + \mathbf{x}^{(2)}\beta_2 + \cdots + \mathbf{x}^{(d)}\beta_d + \xi \\ &= \mathbf{x}^T\beta + \xi \end{aligned}$$

where ξ (pronounce as "ksigh") is a zero-mean random variable that represents model error. β is the d-dimensional vector of coefficients that we train on.

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_d \end{bmatrix}, \mathbf{x}^T = \begin{bmatrix} \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \cdots & \mathbf{x}^{(d)} \end{bmatrix}$$

13.2. Training

We define a set new variables:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \cdots \\ \mathbf{y}_N \end{bmatrix}, X = \begin{bmatrix} \mathbf{x}_1^T \\ \cdots \\ \mathbf{x}_N^T \end{bmatrix}, \mathbf{e} = \begin{bmatrix} \xi_1 \\ \cdots \\ \xi_N \end{bmatrix}$$

Given a training dataset $\{(\mathbf{x}, y)\}$, we want to fit a model $y = \mathbf{x}^T\beta + \xi$. Then to train the model, we need to choose β that makes $\mathbf{e}$ to be as small as possible, in the matrix equation:

$$\mathbf{y} = X \cdot \beta + \mathbf{e}$$

To make $\mathbf{e}$ to be as small as possible, we aim to **minimize** $||\mathbf{e}||^2$, which is:

$$||\mathbf{e}||^2 = ||\mathbf{y} - X\beta||^2 = (\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta)$$

We can differentiating with respect to β:

$$\frac{\mathrm{d}||\mathbf{e}||^2}{\mathrm{d}\beta} = X^TX\beta - X^T\mathbf{y} = 0$$

Finally, if X^TX is invertible, the **least sqaures estimate** of the coefficient is:

$$\hat{\beta} = (X^TX)^{-1}X^T\mathbf{y}$$

13.3. Loss Function

We define the **Least Square Loss function** $Q_j(\beta)$ as:

$$||\mathbf{e}||^2 = f(\beta) = \sum_{j=1}^{k} Q_j(\beta) = \sum_{i=1}(\mathbf{x}_i^T\beta - \mathbf{y}_i)^2$$

13.4. Prediction

If we train the model coefficients $\hat{\beta}$, we can predict y_0^p from $\mathbf{x}_0$, which is defined as:

$$y_0^p = \mathbf{x}_0^T\hat{\beta}$$

13.5. R-square evaluation

We can use this evaluation metric to evaluate our model, which is given by:

$$R^2 = \frac{\text{var}(\{\mathbf{x}_i^T\hat{\beta}\})}{\text{var}(\{y_i\})} = \frac{\text{var}(\{\mathbf{x}_i^T\hat{\beta}\})}{\text{var}(\{\mathbf{x}_i^T\hat{\beta}\}) + \text{var}(\mathbf{e})}$$

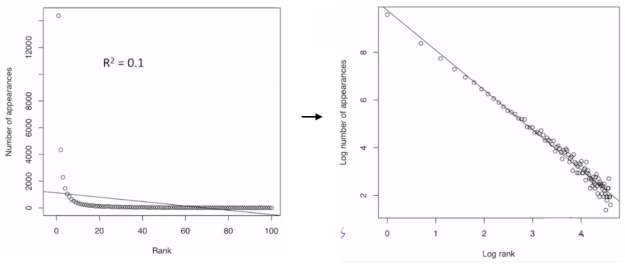
where $0 \leq R^2 \leq 1$, the larger R^2 value the better fit.

13.6. Working with non-linear relationship

To capture these non-linear relationships while still using linear regression, we can apply transformations to the data first. Below are the common methods of doing this:

13.6.1. Zipf's Law. Take the Log.

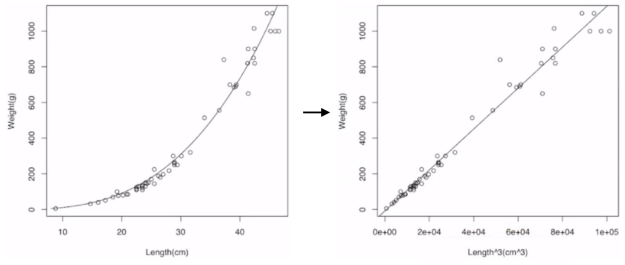
Zipf's law is an empirical law that is commonly observed in natural languages, stating that the frequency of any word is inversely proportional to its rank in the frequency table. In many cases, the relationship between the rank and frequency of words follows a non-linear pattern.



To linearize this relationship for linear regression analysis, we can apply **logarithmic transformation**.

13.6.2. Cubic. Squared Power.

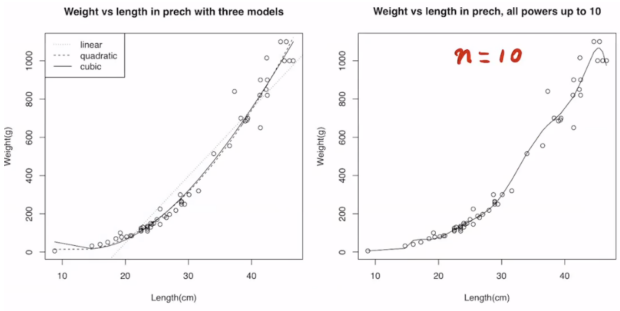
If the relationship between the dependent and independent variables is polynomial (e.g., quadratic, cubic), applying a power transformation can linearize the relationship.



This transformation turns a polynomial relationship into a linear one in terms of the transformed variable, allowing for the use of linear regression.

13.6.3. Over-fitting Issue. One issue with applying the transformation on non-linear relationship and do linear regression, is that the resulting fitting line might make no sense

To avoid this, there are two methods:



1. **Validation** use a validation set to choose the transformed explanatory variables. But the hard part is that the number of combination is exponential in the number of variables.
2. **Regularization** Impose a penalty on complexity of the model during the training, and encourage smaller coefficients.
   That is, we can yield a **regularized least squares estimation** of the coefficients to be:

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T \mathbf{y}$$

where the penalty is added with a weight parameter λ and $\lambda > 0$

14. Unsupervised Learning

Unsupervised learning means **knowledge discovery from feature vectors without labels**. Unsupervised learning can be coupled with supervised learning.

- Discovering latent factors Such as using PCA as the dimensional reduction method, where we use eigenvalues of covariant matrix.
- Discovering clusters Such as k-means.
- Discovering graph structure (Gaussian graph)
- Matrix completion

14.1. Types of Clustering

14.1.1. By Input Type.

- **Similarity-based clustering** indicates that the input is $N \times N$ similarity or distance matrix
- **Feature-based clustering** indicates that the input is $N \times D$ feature matrix

14.1.2. By Output Type.

- **Hierarchical clustering** (HC), we can either do a top-down (divisive) clustering, or bottom-up (agglomerative) clustering

- **Flat clustering**, we have mixture models, K-means clustering, or spectral clustering

14.2. Hierarchical Clustering

14.2.1. Divisive Clustering (Top-down). The main idea is that it starts with all data points in a single cluster and iteratively splits the cluster into smaller clusters until each data point is in its own cluster or a stop criterion is met.

14.2.2. Agglomerative Clustering (Bottom-up). The main idea is that it starts with each data point as a separate cluster and iteratively merges them into larger clusters based on a certain similarity or distance metric. A "dendrogram" is oftenly generated.

14.3. K-means Clustering

k is the number of cluster we want. The steps are:

1. Pick a value k as the number of clusters
2. Select k random cluster centers
3. Iterate until convergence, where we assign each data to the nearest center, and update the center within cluster

Example (from lecture): given a data set $\{0, 2, 4, 6, 24, 26\}$, initialize the k-means clustering algorithm with 2 clusters centers $c_1 = 3, c_2 = 4$.

The 1st-iteration would be:

$c_1 = 3$	$c_2 = 4$
$L = \{0, 2\}$	$R = \{4, 6, 24, 26\}$
$\text{mean}(L) = 1 = c_1$	$\text{mean}(R) = 15 = c_2$

The 2nd-iteration would be:

$c_1 = 1$	$c_2 = 15$
$L = \{0, 2, 4, 6\}$	$R = \{24, 26\}$
$\text{mean}(L) = 3$	$\text{mean}(R) = 25$

Note that it's a minimization of a cost function ϕ:

$$\begin{aligned}\phi(\delta, \mathbf{c}) &= \sum_{i,j} \delta_{i,j}[(\mathbf{x}_i - \mathbf{c}_j)^T(\mathbf{x}_i - \mathbf{c}_j)] \\ &= \sum_{i}^{N}\sum_{j}^{k} \delta_{i,j}||\mathbf{x}_i - \mathbf{c}_j||^2\end{aligned}$$

14.3.1. Choosing k value. Choose based on the knowledge from dataset, or other certain cost functions.

14.3.2. Problem with K-means. K-means clustering is sensitive to outliers.

14.3.3. Vector Quantization. (VQ)

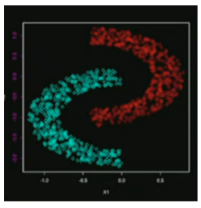
It's about partition of a large set of points (vectors) into groups having approximately the same number of points closest to them. The goal is to minimize the distance between points in a cluster and the central point of that cluster. The process is as follows:

1. **Identify Clusters**: clusters are identified for a data set, each with its own histogram reflecting the frequency of data points in that cluster.
2. **Create New Dataset**: A new dataset is then formulated, where the identified features become the new data points, and their respective counts in the clusters are used as their values.
3. **Utilize Clusters for Prediction**: With this restructured dataset, we can employ the pre-existing cluster patterns and frequencies to estimate or predict new frequencies.
4. **Label Prediction**: By leveraging the labels from the initial data, it is possible to predict corresponding labels for previously unknown data points within this new dataset.

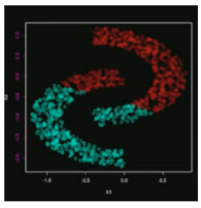
For VQ, no matter how noisy the data is, we can always uniform the features as clusters, and use them together as classifiers. This can be applied to different size of the images.

14.4. Spectral Clustering

14.4.1. Two-moon Problem. Other than being sensitive to outliers, K-means also fails in the Two-moon problem. This is due to the non-convex and complex shape of the data distribution.



Correct



Incorrect with K-means

Spectral clustering is an alternative approach using graph representation

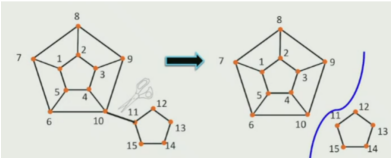
14.4.2. Graph and Adjacency Matrix. Each data point is a node in the graph, and edges between nodes represent some form of similarity or connection between the data points. For example, the edge may be 1 if it's connect, and 0 if it's not. The strength of the connection is often quantified by weights assigned to the edges.

Consider this adjacency matrix:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

$A_{1,2} = 1$ since because node 1 and 2 are connected

14.4.3. Graph Cutting. It's about finding the minimum cut of the graph with least energy, that is, by cutting the fewest edge, we are able to get two subgraphs (or disconnected components.)



14.4.4. Laplacian Matrix. We want to transform the adjacency matrix W into a Laplacian Matrix L, that is, getting:

$$L_{\text{Laplacian Matrix}} = D - W_{\text{Adjacency Matrix}}$$

And D is derived from W:

$$D_{ij} = \begin{cases} \sum_k \omega_{ik} & i = j \\ 0 & i \neq j \end{cases}$$

In other words, the diagonal elements of D is the sum of the same row values in W.

Consider the following adjacency matrix and its D:

$$W = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}, D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

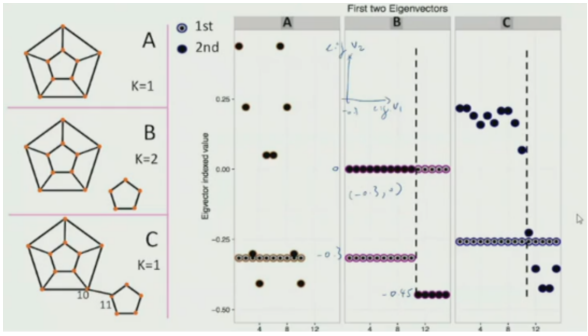
Note that L is a zero-sum matrix with symmetric property, and the weights are non-negative. Two other important properties:

1. All $\lambda_i \geq 0$, and there is at least one eigenvalue $\lambda_0 = 0$.
2. The number of 0-valued eigenvalues λ_i is equal to the number of clusters, or connected graphs in the graph.

L has a quadratic form, which is a **energy function**:

$$f'Lf = \frac{1}{2}\sum_{ij}(f_i f_j)^2 \geq 0, f \neq \vec{0}$$

14.4.5. Eigenvalue Distribution. We can observe the eigenvalues of Laplacian matrix to find the minimum cut. In the following picture, horizontal axis is the vertice index, and vertical axis is the eigenvalue indexed value, where we compare the first two eigenvectors. Coupled with K-means, we can find the boundary in the graph, where in graph C, the dash line indicate the desired edge



15. Markov Chain

A Markov Chain is a stochastic model describing a sequence of possible events where the probability of each event depends only on the state attained in the previous event. It is about the **conditional probability with matrix**. We have the following:

$$P(X_{n+1}|X_n, X_{n-1}, \ldots, X_1) = P(X_{n+1}|X_n) = f(n)$$

Example We can find the total probability of $P(X_3)$:

$$P(X_3) = \sum_{X_2, X_1} P(X_3|X_2, X_1)P(X_2, X_1)$$

with just this:

$$P(X_3) = \sum_{X_2} P(X_3|X_2)P(X_2)$$

15.1. Transition Probability Matrix

It is about the **probability of transitioning from one state to another**.

For a Markov Chain with states $S = \{s_1, s_2, \ldots, s_n\}$, the transition probability from state s_i to state s_j is denoted as P_{ij}. The transition probabilities for all state pairs form a transition probability matrix P, where each element P_{ij} represents the probability of moving from state i to state j.

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

Constructing the transition matrix: determine the states, then draw state diagram with directed edges as the probability. Transition matrix does not contain the initial probability.

15.1.1. Properties.

- Each entry satisfy the following:

$$P_{ij} = P(X_t = j | X_{t-1} = i), \text{ where } P_{ij} \geq 0$$

- Each row (outgoing edge of a node) sums to 1.
- Each element represents a conditional probability and is less than or equal to 1.

15.1.2. Probability Distributions over States.

Let π be a row vector containing the probability distribution over all the finite discrete states at $t = 0$, we have:

$$\pi_i = P(X_0 = i)$$

We call π is the prior probability. Then let $P^{(t)}$ be a row vector containing the probability distribution over states at time point t, we have:

$$P^{(t)} = P(X_i = i)$$

We can use the initial state distribution of π to compute future state distributions using the transition matrix. For example, the probability distribution after one step is $P^{(1)} = \pi P$, after two steps is $P^{(2)} = P^{(1)} P$. We ca derive a propagation for between time step 0 to 1 here:

$$\begin{aligned} P_j^{(1)} &= P(X_1 = J) \\ &= \sum_i P(X_1 = j, X_0 = i) \\ &= \sum_i P(X_1 = j | X_0 = i) P(X_0 = i) \\ &= \sum_i P_{ij} \pi_i \end{aligned}$$

Weather Example Suppose each row and col is in the order of "Sunny", "Rainy", and "Snowy", below is a simple transition matrix:

$$P = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.4 & 0.1 & 0.5 \end{pmatrix}$$

Then we can denote the 0.7 entry as:

$$P(X_{t+1} = \text{Sunny} | X_t = \text{Sunny}) = 0.7$$

And the initial distribution π is what occurred at the moment. So if today is sunny, then $\pi = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$

15.1.3. Irreducibility.

If one state can be reached from any other state in the graph, the Markov chain is called **irreducible**, i.e., single chain.

15.2. Stationary and Non-stationary

The stationary distribution $\mathbf{s}$ has the property:

$$\mathbf{s}P = \mathbf{s}$$

where $\mathbf{s}$ is a row eigenvector of P with eigenvalue 1.

Stationary Markov Chain is a markov chain if its transition probabilities do not change over time, i.e., the state distribution is constant, or, the probability of being in any given state is the same at every time step.

To solve a stationary markov chain, that is, solving $\mathbf{s}$ for $\mathbf{s}P = \mathbf{s}$, we have:

$$\begin{aligned} (\mathbf{s}P)^T &= \mathbf{s}^T \\ P^T \mathbf{s}^T &= \mathbf{s}^T \\ A\mathbf{u} &= \mathbf{u} && (A = P^T, \mathbf{u} = \mathbf{s}^T) \\ A\mathbf{u} &= 1 \cdot \mathbf{u} && (\text{Use } \lambda = 1) \\ (A - I)\mathbf{u} &= 0 \end{aligned}$$

Solve for $\mathbf{u}$

Non-stationary Markov Chain is a markov chain where the transition probabilities can change over time. It's typically one of the following situation:

1. **Periodic**, not stable and not irreducible
2. **Absorbing**, stable bu not irreducible

16. Appendix

16.1. Common Derivatives

$f(x)$	$f'(x)$
a^x	$a^x \ln(a)$
e^{ax}	ae^{ax}
$\ln(x)$	$\frac{1}{x}$
$\log_a(x)$	$\frac{1}{x \ln a}$

16.2. Logarithm Rules

- $\ln(e^{ax}) = ax$
- $\log_a(a) = 1$
- $\log_a(1) = 0$
- $\log_a(x^n) = n \log_a(x)$
- $\log_b(c) = \frac{\log_a(c)}{\log_a(b)}$
- $\log_b(a) = \frac{1}{\log_a(b)}$
- $\log_a(\frac{1}{b}) = -\log_a(b)$
- $\log_{a^m}(a^n) = \frac{n}{m}, m \neq 0$