Part 1

Week 1: Counting & Permutation

(Basic) Sum Rule of Counting

Set A of n outcomes, set B of m outcomes
No outcome in A is in B, vice versa - when A ∩ B is the empty set
An experiment is performed by drawing one outcome from **either** A or B
There are m + n possible outcomes
Example: servers

Product Rule of Counting

Experiment A of n outcomes, Experiment B of m outcomes
An experiment is performed by drawing one outcome from **both** A and B
For the two experiments together there are mn possible outputs
Example: bits in a byte

- Generalised Product rule of counting = extending the Product Rule of counting to many experiments i.e. for the r experiments together there are $n1 \times n2 \times \cdots \times nr$ possible outputs.
- Generalised Sum Rule of Counting = when A ∩ B is not the empty set. For the experiments together there are $|A| + |B| - |A \cap B|$ possible outputs

Set {A, B} has two permutations {A,B}, {B,A} and one combination {A,B}

Number of permutations, with/without repeated objects

Permutation - Counting the number of ways to generate an ordered (order matters and is counted) subset of size k from a set of n distinguishable objects.

In general, number of permutations of n objects is n! – by direct application of product rule.
E.g the number of ways we can arrange a, b, c = 3!
However when repeated objects are concerned e.g m,o,o there are 3!/2! Ways to arrange the letters

$n! / (n1!n2! \cdots nk!)$

Number of combinations (n choose k)

Combination - • Counting the number of ways to generate an unordered(order does not matter and is not counted) subset of size k from a set of n distinguishable objects

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

E.g. Number of distinct pizzas we can create by selecting 4 toppings from 6 available. 6 choose 4 = 15

Effect of simple constraints on counting e.g. two people must sit together or must not

Add up all possible combinations that take into account the constraint or minus the constraint combination from all possible combinations. NB Test Questions 1,Question 4

Week 2: Axioms of Probability, Conditional Probability and Bayes Theorem

Sample Space = The set of all possible outcomes of an experiment
Random Event = subset of sample space

Sets:
Know set operations: union, intersection, complement and combinations of these
Basic Properties of sets: draw venn diagrams to figure out
DeMorgan's laws:
E and F are events,
$(E \cup F)^c = E^c \cap F^c$
$(E \cap F)^c = E^c \cup F^c$

Axioms of Probability:

Axiom 1: $0 \leq P(E) \leq 1$
Axiom 2: $P(S) = 1$, where S is sample space (set of all possible outcomes)
Axiom 3: If E and F are mutually exclusive ($E \cap F = \varnothing$, they cannot occur at the same time) then $P(E \cup F) = P(E) + P(F)$.

Immediate Consequences/Implications of Axioms incl proof:

$P(E^c) = 1 - P(E)$
since $S = E \cup E^c$ and $E \cap E^c = \varnothing$ then $P(S) = 1 = P(E) + P(E^c)$
$E \subset F$ implies that $P(E) \leq P(F)$
since $F = E \cup (E^c \cap F)$ and $E \cap E^c = \varnothing$
then $P(F) = P(E) + P(E^c \cap F))$
$P(E^c \cap F) \geq 0$ so $P(E) = P(F) - P(E^c \cap F)) \leq P(F)$
$P(E \cup F) = P(E) + P(F) - P(E \cap F)$
$E \cup F = E \cup (E^c \cap F)$ and $E \cap (E^c \cap F) = \varnothing$ (mutually exclusive)
$F = (E \cap F) \cup (E^c \cap F)$, also mutually exclusive
So $P(E \cup F) = P(E) + P(E^c \cap F)$
and $P(F) = P(E \cap F) + P(E^c \cap F)$ i.e. $P(E^c \cap F) = P(F) - P(E \cap F)$
Equally Likely Outcomes
$P(S) = 1$, $P(\text{Heads}) = P(\text{Tails}) = 1$, $2p = 1$, $p = 1/2$

Sampling with and without replacement

Conditional Probability = the probability that event E occurs given that event F has already occurred. Written as $P(E|F)$**.** It is a probability where Sample space is restricted to $S \cap F$ and event space is restricted to $E \cap F$

General definition:

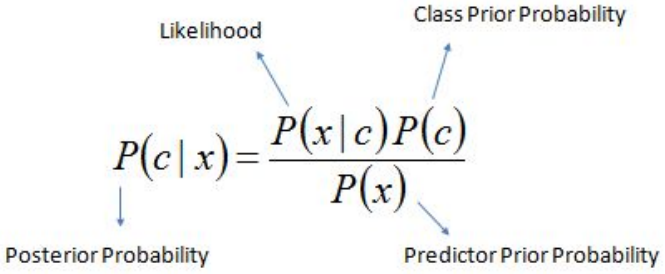
$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

where $P(F) > 0$. Implies

$$P(E \cap F) = P(E|F)P(F)$$

known as the chain rule – its important !
If $P(F) = 0$?

- $P(E|F)$ is undefined
- Can't condition on something that can't happen



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

NB CHAIN RULE - P(E∩F) = P(E|F)P(F)
Conditional prob is a probability (satisfies axioms, incl proof.)
0 ≤ P(E|F) ≤ 1
E ∩ F ⊂ F so P(E ∩ F) ≤ P(F) and P(E|F) = P(E∩F) P(F) ≤ 1
P(S|F) = 1
P(S|F) = P(S∩F) / P(F) = P(F) /P(F) = 1 (chain rule)
If E1, E2 are mutually exclusive events then P(E1 ∪ E2|F) = P(E1|F) + P(E2|F)
P(E1 ∪ E2|F) = P((E1 ∪ E2)∩F) P(F) = P((E1∩F) ∪ (E2∩F)) P(F) = P(E1∩F)+P(E2∩F) / P(F)

Marginalisation incl proof:
Suppose we have mutually exclusive events F1, F2, ..., Fn such that F1 ∪ F2 ∪ ... ∪ Fn = S. Then P(E) = P(E ∩ F1) + P(E ∩ F2) + · · · + P(E ∩ Fn)
Prove by chain rule - see slide 15

NB Special Case: use in HIV/Guilty question
P(E) = P(E|F)P(F) + P(E|F c)P(F c) = P(E|F)P(F) + P(E|F c)(1 − P(F))

Marginalisation example:
Roll two coins. What is the probability that the first coin is heads ?
Event E is first coin heads, F1 is second coin heads, F2 is second coin tails
P(E) = P(E ∩ F1) + P(E ∩ F2) = (1/2 × 1/2) + (1/2 × 1/2) = 1/2

Bayes Rule:

Recall

$$P(E \cap F) = P(E|F)P(F)$$

Clearly, and also

$$P(F \cap E) = P(F|E)P(E)$$

P(E|F) = posterior = updated probability of E after observing F
P(F|E) = likelihood = probability of F given E
P(E) = prior = our guess with no extra info
P(F) = evidence = probability with no extra info

Week 3: Independance

Two events E and F are independent if the order in which they occur doesn't matter. Alternatively, if observing one doesn't affect the other.

Definition: Two events E and F are independent if $P(E \cap F) = P(E)P(F)$

NB three events are independent if they are pairwise independent and triply independent
$P(E \cap F \cap G) = P(E)P(F)P(G)$ and
$P(E \cap F) = P(E)P(F)$,
$P(E \cap G) = P(E)P(G)$,
$P(F \cap G) = P(F)P(G)$,

Caution re fragility of independence assumptions
Multiplying can result in very small probabilities e.g. probability of defaulting on a mortgage

Definition of conditional independence
NB: Independent events can become dependent when we condition on additional information. Also dependent events can become independent.

Two events E and F are called conditionally independent given G if:
$P(E \cap F|G) = P(E|G)P(F|G)$

NB: if two events are independent it does not follow that they are conditionally independent and vice versa

Part 2

Week 4: Random Variables, Bernoulli and Binomial RVs

Random Variable = Effectively mapping every event to a real number

- **Discrete Random Variable =** X can only take on discrete values
- **Continuous Random Variable =** X can take on continuous data

Indicator Random Variable = takes value 1 if event E occurs and 0 if event E does not occur. The sample space, I = {1,0}

Random Variable X can be associated with outcomes of an experiment e.g. X = 0 when outcome is (T,T), X = 1 when outcome is (H,T) or (T, H), X = 2) when outcome is (H, H).

Events are linked to values of random variables, so can apply ideas for events directly to RVs (chain rule, Bayes, marginalisation)

Probability Mass Function:
A probability is associated with each value that a discrete random variable can take

Cumulative Distribution Function:
F(a) = P(X ≤ a) where a is real-valued.
Use PMF to calculate e.g. F(2) = 3/10 = P(X = 0) + P(X = 1) + P(X = 2) so P(X = 2) = 2/10
Bernoulli Random Variable: when an experiment results in success or failure, X ~ Ber(p).

Binomial Random Variable: when the number of successes or failures is counted i.e. a sum of n bernoulli random variables.

- X is a **Binomial** random variable: $X \sim Bin(n, p)$

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}, \ i = 0, 1, \cdots, n$$

(recall $\binom{n}{i}$ is the number of outcomes with exactly i successes and $n - i$ failures)

Examples:

- number of heads in n coin flips
- number of 1's in randomly generated bit string of length n
- number of packets erased out of a file of n packets

Suppose X ~ Bin(n1, p) and Y ~ Bin(n2, p) (its important than p is the same for both) Then Z = X + Y ~ Bin(n1 + n2, p)
Z = X1 + X2 + · · · + Xn1 + Y1 + Y2 + · · · + Yn2 . All terms are independent, all are Ber(p).
Use in voters question.

Simple stochastic simulation: generating bernoulli and binomial samples in matlab

Week 5: Mean, Variance, Correlation and Conditional Expectation

Definition of expected value = Sometimes we have a number of measurements that we want to summarise by a single value, also referred to as mean or average

Viewing the probability of an event as the frequency with which it occurs when an experiment is repeated many times, the expected value tells us about the overall outcome we can expect.

$$E[X] = \sum_{i=1}^{n} x_i P(X = x_i)$$

I.e. Multiply each value times its respective probability.

Interpretation of expected value in games of chance/reward

NB - if it costs to play the game you must take this into account when calculating reward

Expected value of an indicator RV

Suppose I is the indicator variable for event E (so I = 1 if event E occurs, I = 0 otherwise). Then E[I] = 1 × P(E) + 0 × (1 − P(E)) = P(E).

Expected value of number of iterations of repeated game (coin tossing etc)

Limitations of expected value in games of chance/reward e.g. gamblers ruin

Linearity of expected value incl proof.

E[aX + b] = aE[X] + b

$$\begin{aligned} E[aX + b] &= \sum_{i=1}^{n} (ax_i + b) P(X = x_i) \\ &= \sum_{i=1}^{n} ax_i P(X = x_i) + \sum_{i=1}^{n} bP(X = x_i) \\ &= a \sum_{i=1}^{n} x_i P(X = x_i) + b \sum_{i=1}^{n} P(X = x_i) \\ &= aE[X] + b \end{aligned}$$

Use example = changing currency

Use of linearity in expected value of sums of random variables

E[aX + bY] = aE[X] + bE[Y]

for any two random variables X and Y and constants a and b.
Proof:

$$E[aX + bY] = \sum_x \sum_y (ax + by)P(X = x \text{ and } Y = y)$$
$$= a\sum_x \sum_y xP(X = x \text{ and } Y = y) + b\sum_y \sum_x yP(X = x \text{ and } Y = y)$$
$$\overset{(a)}{=} a\sum_x xP(X = x) + b\sum_y yP(Y = y)$$
$$= aE[X] + bE[Y]$$

(a) Recall marginalising, $\sum_y P(X = x \text{ and } Y = y) = P(X = x)$

Expected value of product of independent RVs, incl proof.

E[XY] = E[X]E[Y]

- Take two **independent** random variables X and Y
- $E[XY] = E[X]E[Y]$
- Proof:

$$E[XY] = \sum_x \sum_y xyP(X = x \text{ and } Y = y)$$
$$= \sum_x \sum_y xyP(X = x)P(Y = y)$$
$$= \sum_x xP(X = x)\sum_y yP(Y = y)$$
$$= E[X]E[Y]$$

Definition of variance = Variance is a summary value (a statistic) that quantifies “spread”
Let X be a random variable with mean μ.
The variance of X is Var(X) = E[(X − μ)^2]
Variance is mean squared distance of X from the mean μ

Var(X) ≥ 0
Standard deviation is square root of variance p Var(X)

Alternative expression for discrete random variables : Var(X) = E[X^2] − (E[X])^2

Proof:

$$\begin{aligned} Var(X) &= \sum_{i=1}^{n}(x_i-\mu)^2 p(x_i) \\ &= \sum_{i=1}^{n}(x_i^2 - 2x_i\mu + \mu^2)p(x_i) \\ &= \sum_{i=1}^{n} x_i^2 p(x_i) - 2\sum_{i=1}^{n} x_i p(x_i)\mu + \mu^2 \sum_{i=1}^{n} p(x_i) \\ &= E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - \mu^2 \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

var(aX+b)=a^2var(X), incl proof.
Unlike expectation, variance is not linear. Instead we have the above ^. Observe that b does not affect variance.

Proof:

$$\begin{aligned} Var(aX+b) &= E[(aX+b)^2] - E[aX+b]^2 \\ &= E[a^2X^2 + 2abX + b^2] - (aE[X] + b)^2 \\ &= a^2E[X^2] + 2abE[X] + b^2 - a^2E[X]^2 - 2abE[X] - b^2 \\ &= a^2E[X^2] - a^2E[X]^2 \\ &= a^2(E[X^2] - E[X]^2) = a^2Var(X) \end{aligned}$$

(recall $E[aX+b] = aE[X] + b$).

Variance of sum of independent RVs, incl proof.

Var(X + Y) = Var(X) + Var(Y).

$$\begin{aligned} Var(X+Y) &= E[(X+Y)^2] - E[X+Y]^2 \\ &= E[X^2 + 2XY + Y^2] - (E[X] + E[Y])^2 \\ &= E[X^2] + 2E[XY] + E[Y^2] - E[X]^2 - 2E[X]E[Y] - E[Y]^2 \\ &= E[X^2] - E[X]^2 + E[Y^2] - E[Y]^2 + 2E[XY] - 2E[X]E[Y] \\ &= E[X^2] - E[X]^2 + E[Y^2] - E[Y]^2 \\ &= Var(X) + Var(Y) \end{aligned}$$

(recall $E[XY] = E[X]E[Y]$ when X and Y are independent)

- Bernoulli random variable, $X \sim Ber(p)$:

$$E[X] = p$$
$$Var(X) = E[X^2] - (E[X])^2 = p - p^2 = p(1-p)$$

- Binomial random variable, $X \sim Bin(n, p)$. Sum of n $Ber(p)$ independent random variables so:

$$E[X] = np$$
$$Var(X) = np(1-p)$$

Definition of joint PMF
P(X = x and Y = y)

PMF of X given Y = y is P(X = x|Y = y) = P((X=x and Y =y) / P(Y =y)

Definition of covariance
Say X and Y are random variables with expected values μX and μY .
The covariance of X and Y is defined as:
Cov(X, Y) = E[(X − μX)(Y − μY)] = E[XY] − E[X]E[Y]

Recall when X and Y are independent then E[XY] = E[X]E[Y], so Cov(X, Y) = 0. But Cov(X, Y) = 0 does not imply that X and Y are independent

NB - Cov(X, X) = Var(X)

Definition of correlation

The correlation is another example of a summary statistic. It indicates the strength of a linear relationship between X and Y .

Correlation says nothing about the slope of line (other than its sign).
When relationship between X and Y is not roughly linear, correlation coefficient tells us almost nothing

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Correlation varies between -1 and 1
If X = Y then corr(X, Y) = 1. If X = −Y then corr(X, Y) = −1.

Recall when X and Y are independent then E[XY] = E[X]E[Y], so corr(X, Y) = 0. But corr(X, Y) = 0 does not imply that X and Y are independent.

Correlation vs causality
Correlation does not imply causality

Conditional expectation

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

Conditional expectation can be used to make predictions
Linearity of conditional expectation, incl proof.
Marginalisation and conditional expectation, incl proof
Use of conditional expectation in random sums

Part 3

Week 8: Inequalities, Sample Mean, Weak law of large numbers and CLT

When we don't know the probability distribution but we know mean, variance or non-negativity (we know it's positive) we can use inequalities to say something about the distribution (not precise but still important especially if we collect more and more measurements - **LLN - law of large numbers**).

Use - stock market - limited information, forecasting required.
E.g. P(X-2) <= 0.5

Markov's Inequality
What is the probability that the value of r.v. X is "far" from its mean ?
A generic answer for non negative RV X is markov's inequality:

$$P(X \geq a) \leq \frac{E(X)}{a} \text{ for all } a > 0$$

Proof:

- Let indicator $I_a(X) = 1$ if $X \geq a$ and $I_a(X) = 0$ otherwise. Then $aI_a(X) \leq X$ i.e. $I_a(X) \leq \frac{X}{a}$.
- $E(I_a(X)) \leq E(\frac{X}{a}) = \frac{E(X)}{a}$
- $E(I_a(X)) = P(X \geq a) \leq \frac{E(X)}{a}$

NB E[indicator random variable] = probability of the event
E[constant] = constant

Chebyshev's Inequality - *built on Markov's*

Suppose X is a random variable with mean $E(X) = \mu$ and variance $var(X) = \sigma^2$. Then

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2} \text{ for all } k > 0$$

Proof:

- Since $(X - \mu)^2$ is a non-negative random variable we can apply Markov's inequality with $a = k^2$ to get

$$P((X - \mu)^2 \geq k^2) \leq \frac{E((X - \mu)^2)}{k^2} = \frac{\sigma^2}{k^2}$$

- Note that $(X - \mu)^2 \geq k^2 \iff |X - \mu| \geq k$, so

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

'Spread' of random variables around mean is linked to the variance.
Can be applied using a 'spread' which is a multiple of standard deviation e.g 3σ rather than a constant.

- Applying Chebyshev's inequality $P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$ with $k = n\sigma$ gives:

$$P(|X - \mu| \geq n\sigma) \leq \frac{1}{n^2}$$

- With $n = 3$ then $P(|X - \mu| \geq 3\sigma) \leq \frac{1}{9} = 0.11$.
- This holds even when distribution is not Gaussian, so can be quite handy (if conservative).

k	Min. % within *k* standard deviations of mean	Max. % beyond *k* standard deviations from mean
1	0%	100%
$\sqrt{2}$	50%	50%
1.5	55.56%	44.44%
2	75%	25%
3	88.8889%	11.1111%
4	93.75%	6.25%
5	96%	4%
6	97.2222%	2.7778%
7	97.9592%	2.0408%
8	98.4375%	1.5625%
9	98.7654%	1.2346%
10	99%	1%

Less than 1,2 and 3 standard deviations mean 0->75->93.75 vs 68->95->99.7 empirical rule. Chebyshev is clearly weaker.

NB
Markov uses a value, a. Chebyshev uses a distance from the mean, k.
Both provide an upper bound
Markov uses mean/expected value, Chebyshev uses mean/expected value and variance
Chebyshev is more accurate than markov

Distribution of Sample Mean:
It's binomial - e.g. tossing a coin 5 times, Bin(5,1/2), counting number of successes and failures.

X is called the "sample mean" or the "empirical mean".
X is a random variable.

Random variable $\bar{X} = \frac{1}{N}\sum_{k=1}^{N} X_k$.

- Suppose the X_k are **independent and identically distributed** (i.i.d)
- Each X_k has mean $E(X_k) = \mu$ and variance $Var(X_k) = \sigma^2$.

Then we can calculate the mean of $\bar{X}$ as:

$$E(\bar{X}) = E(\frac{1}{N}\sum_{k=1}^{N} X_k) = \frac{1}{N}\sum_{k=1}^{N} E(X_k) = \mu$$

NB: recall linearity of expectation: $E(X + Y) = E(X) + E(Y)$ and $E(aX) = aE[X]$

Sample mean is an unbiased estimator of μ since E[X] = μ

We can calculate the variance of $\bar{X}$ as:

$$var(\bar{X}) = var(\frac{1}{N}\sum_{k=1}^{N} X_k) = \frac{1}{N^2} var(\sum_{k=1}^{N} X_k)$$
$$= \frac{1}{N^2}\sum_{k=1}^{N} var(X_k) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$$

NB: recall $Var(aX) = a^2 Var(X)$ and $Var(X + Y) = Var(X) + Var(Y)$ when X, Y independent.

- As N increases, the variance of $\bar{X}$ falls.
- $Var(NX) = N^2 Var(X)$ for random variable X.
- But when add together **independent** random variables $X_1 + X_2 + \cdots$ the variance is only $NVar(X)$ rather than $N^2 Var(X)$
- This is due to **statistical multiplexing**. Small and large values of X_i tend to cancel out for large N.

Weak Law for Large Numbers:

Sample mean concentrates around mean μ as N increases

Consider N independent identically distributed (i.i.d) random variables $X_1, \cdots X_N$ each with mean μ and variance σ^2. Let $\bar{X} = \frac{1}{N}\sum_{k=1}^{N} X_k$. For any $\epsilon > 0$:

$$P(|\bar{X} - \mu| \geq \epsilon) \to 0 \text{ as } N \to \infty$$

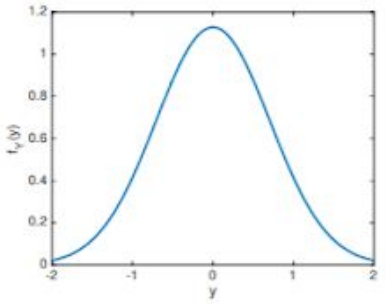
That is, $\bar{X}$ **concentrates** around the mean μ as N increases.
Proof:

- $E(\bar{X}) = E(\frac{1}{N}\sum_{k=1}^{N} X_k) = \frac{1}{N}\sum_{k=1}^{N} E(X_k) = \mu$
- $var(\bar{X}) = var(\frac{1}{N}\sum_{k=1}^{N} X_k) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$
- By Chebyshev's inequality: $P(|\bar{X} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{N\epsilon^2}$

Central Limit theorem:
Curve becomes bell shaped as N increases

Normal or Gaussian function defines CLT

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



$\mu = 0,\ \sigma = 1$

General challenge - how big should N be?

Week 9: Confidence Intervals, Continuous Random Variables

e.g. $p - 0.05 \leq Y \leq p + 0.05$ with probability at least 0.95.

Given a random variable X a confidence interval [a,b] is an interval such that $P(a \leq X \leq b) \geq c$ where c is a target probability e.g. 0.95 or 0.99.

Three ways to calculate:

Chebyshev	Mean and variance required, chebyshev allows us to select a and b such that $P(a \leq X \leq b) \geq c$. A and b are often larger than necessary as chebyshev provides an upper bound (>=) for the distance from the mean. Works for all N. Not an approximation.
CLT	Allows us to approximate the distribution of X as normal (1σ, 2σ, 3σ confidence intervals) provided X is the sum of sufficiently many independent and identically distributed random variables. Main difficulty is in the "sufficiently many" part, so CLT confidence bounds may be overly optimistic i.e select interval [a,b] to be smaller than it ought to be. CLT confidence intervals are approximations whose accuracy depends on the size of N, Chebyshev confidence intervals are real confidence intervals. Gives full distribution of sample mean.
Bootstrapping	Bootstrapping uses multiple measurements of X through **resampling** to estimate its distribution and then from this to estimate confidence interval [a,b]. This requires sufficiently many observations of X that the estimate of its distribution is accurate, but does not require that the distribution is Normal. Again, confidence interval is only an approximation whose accuracy depends on the size of N. Gives full distribution of sample mean. Need access to all N measurements. "Quick and dirty method!"

Example: Running Time of New Algorithm

Suppose we have an algorithm to test. We run it N times and measure the time to complete, gives measurements $X_1, \cdots, X_N$.

- Mean running time is $\mu = 1$, variance is $\sigma^2 = 4$
- How many trials do we need to make so that the measured sample mean running time is within 0.5s of the mean μ with 95% probability ? $P(|X - \mu| \geq 0.5) \leq 0.05$ where $X = \frac{1}{N}\sum_{k=1}^{N} X_k$
- CLT tells us that $X \sim N(\mu, \frac{\sigma^2}{N})$ for large N. Normal distribution satisfies the "68-95-99.7 rule".

$$P(-\sigma \leq X - \mu \leq \sigma) \approx 0.68$$
$$P(-2\sigma \leq X - \mu \leq 2\sigma) \approx 0.95$$
$$P(-3\sigma \leq X - \mu \leq 3\sigma) \approx 0.997$$

So we need $2\sigma = 2\sqrt{\frac{\sigma^2}{N}} = 0.5$ i.e. $N \geq 64$.

chebyshev uses upper bound >= 0.5 this means that we are getting the probability that sample mean running time distance from mean running time is equal to or bigger than 0.5 - when what we really want <= 0.5 i.e. 'within 0.5'. This can be overcome using the axiom of probability **P(E c) = 1 − P(E)**. P(|X - u| < 0.5) = 1 - P(|X-u| >= 0.5) but now we're not including the probability that distance from mean is equal to 0.5…. Trade offs!

CLT uses mean and variance proved by the distribution of sample mean (see above in notes) Questions often ask for how big N should be to get certain CI e.g. voting

Continuous Random Variables:
E.g. travel time to work, temperature of a room.

CDF makes sense for both continuous and discrete RVs
FY (y) := P(Y ≤ y)
NB - CDF always starts at zero and rises to one, it never decreases.

Note:
P(Y ≤ b) = P(Y ≤ a) + P(a < Y ≤ b)
i.e. FY (b) = FY (a) + P(a < Y ≤ b)
Therefore; P(a < Y ≤ b) = FY (b) − FY (a) … Used below in PDF

Area under a curve:

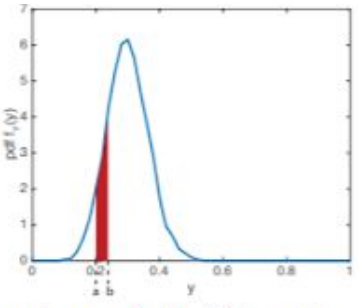
Write the area under curve between a and b as $\int_a^b f(y)dy$
Think of integral as the sum of areas of rectangles each of width h as $h \to 0$. Integral symbol $\int$ is supposed to be suggestive of a sum.
Can think of dy as h (infinitesimally small).

Probability Density Function
A function of a continuous random variable, **whose integral** across an interval gives the probability that the value of the variable lies within the same interval.
PDF is not a probability, it's integral is! I.e. the area under the PDF is.

- For a continuous-valued random variable Y there exists a function $f_Y(y) \geq 0$ such that:

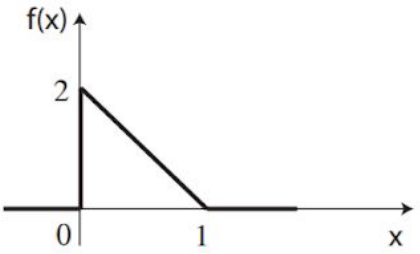
$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt$$

$$\begin{aligned} P(a < Y \leq b) &= F_Y(b) - F_Y(a) \\ &= \int_{-\infty}^{b} f_Y(t)dt - \int_{-\infty}^{a} f_Y(t)dt \\ &= \int_{a}^{b} f_Y(t)dt \end{aligned}$$



NB - CDF is the integral of PDF/ PDF is the derivative of the CDF!
Therefore in the above example FY(b) - FY(a) you could use CDF if possible.
NB - CDF has a limit of 1 as it is a probability, therefore the total area under the PDF is equal to 1. In other words, the chances of the outcome being in the total interval of possibilities is 100%.

NB - you don't always need integration! E.g. Suppose continuous random variable X has pdf as shown:

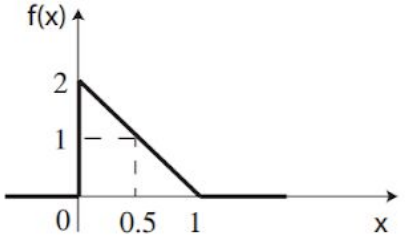


What is the probability P(0≤X≤1) ? What is the probability P(0≤X≤0.5) ?

Solution

(i) The area under the triangle is 1 (the area of the square with side of length 1 is 1, and the area of the triangle is half of that square). Therefore P(0≤X≤1)=1

(ii) Split the area of f(x) between 0≤x≤0.5 into rectangle and a triangle as follows:



The area of the rectangle is 1x0.5 = 0.5. The area of the smaller triangle is 0.25. So the total area is 0.75 and P(0≤X≤0.5)=0.75.

Uniform Random Variables PDF:

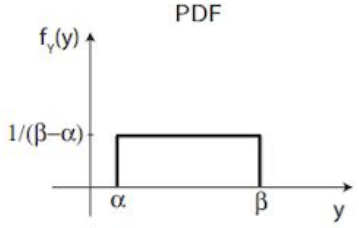
"Since X is continuous valued it is only intervals such as 0≤X≤0.25 that have non-zero probability."

Intervals are of the same length in the distribution

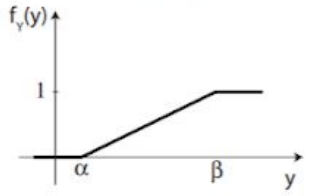
Y is a **uniform random variable** when it has PDF:

$$f_Y(y) = \begin{cases} \frac{1}{\beta-\alpha} & \text{when } \alpha \le y \le \beta \\ 0 & \text{otherwise} \end{cases}$$

PDF



CDF



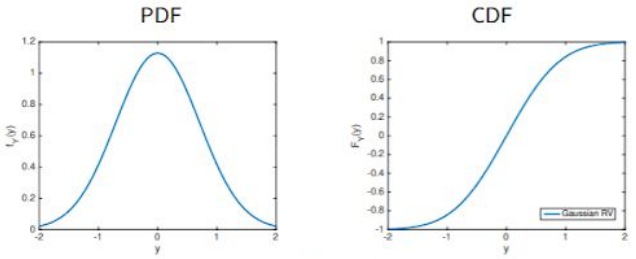
- For $\alpha \le a \le b \le \beta$: $P(a \le Y \le b) = \frac{b-a}{\beta-\alpha}$
- rand() function in Matlab.
- A bus arrives at a stop every 10 minutes. You turn up at the stop at a time selected uniformly at random during the day and wait for 5 minutes. What is the probability that the bus turns up ?

The height of the curve is 1/b-a or 0 otherwise in the PDF

Normal Random Variable PDF:

Y is a **Normal random variable** $Y \sim N(\mu, \sigma^2)$ when it has PDF:

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



$\mu = 0, \sigma = 1$

- $E[Y] = \mu$, $Var(Y) = \sigma^2$
- Symmetric about μ and defined for all real-valued x
- A Normal RV is also often called a **Gaussian random variable** and the Normal distribution referred to as the Gaussian distribution.

we can think of fX (x)dx as the probability that X takes a value between x and x + dx.

NB Expectation and Variance:

For discrete RV X	For continuous RV X
$E[X] = \sum_x xP(X = x)$	$E[X] = \int_{-\infty}^{\infty} xf_X(x)dx$
$E[X^n] = \sum_x x^n P(X = x)$	$E[X] = \int_{-\infty}^{\infty} x^n f_X(x)dx$

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

E[aX + b] = aE[X] + b
Var(X) = E[(X − μ)^2] = E[X^2] − (E[X])^2
Var(aX + b) = a^2 Var(X) **… all still apply**
Joint CDF
When X and Y are independent then:
P(X ≤ x and Y ≤ y) = P(X ≤ x)P(Y ≤ y) i.e. FXY (x, y) = FX (x)FY (y)

For continuous:

$$F_{XY}(x, y) = \int_{\infty}^{x} \int_{-\infty}^{y} f_{XY}(u, v)du\ dv$$

Conditional PDF:

Suppose X and Y are two continuous random variables with joint PDF $f_{XY}(x, y)$. Define conditional PDF:

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

the chain rule also holds for PDFs:

$$f_{XY}(x, y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x)$$

and so we have Bayes Rule for PDFs:

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)}$$

Independence in PDFs

Suppose X and Y are two continuous random variables with joint PDF $f_{XY}(x, y)$. Then X are Y are independent when:

$$f_{XY}(x, y) = f_X(x)f_Y(y)$$

Why ?

$$\begin{aligned} P(X \leq x \text{ and } Y \leq y) &= \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(u, v)dudv \\ &= \int_{-\infty}^{x} f_X(u)du \int_{-\infty}^{y} f_Y(v)dv \\ &= P(X \leq x)P(Y \leq y) \end{aligned}$$

Part 4:
Week 10: Linear Regression 1 & 2
Week 11: Logistic Regression

Building Models!

We can use data to make predictions.

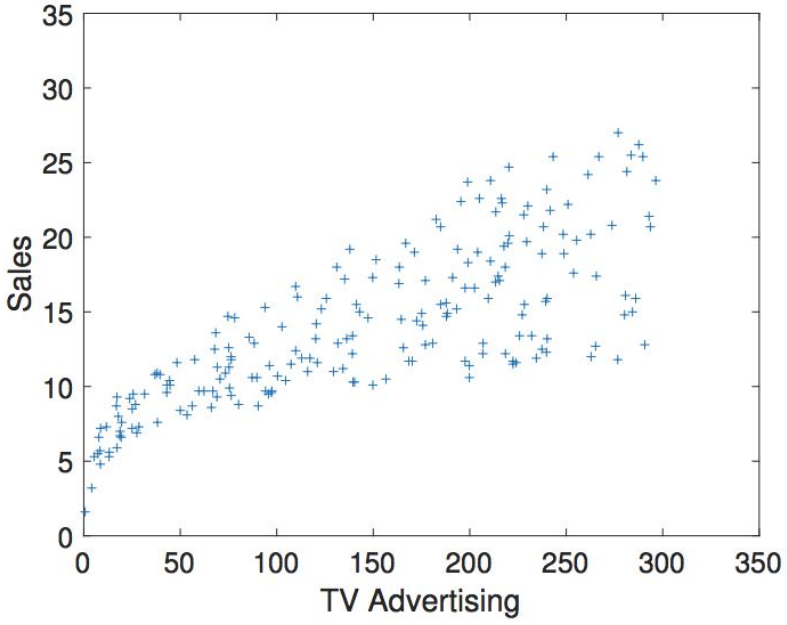
We do this by building a model.

A model is an equation that summarises how variables relate to each other. For example, if an ounce of gold is $400 then a model for gold could be:

$$\$400 * Weight = Price$$

In computer science we get our using data models by “training” functions.

Say we’re given a graph of data:



We want an equation (or model) that has the least average distance from each dot. What we do is make guess. This is done by taking random values for our equation. This equation will be in the general form of $f(x) = ax + b$. Since this equation was a guess it is not likely to fit the data very well. To help us calculate a better equation we use a cost function. A cost function is a function that calculates the average distance from each data point to the equation that we guessed.

For an model with a general equation of the form $f(x) = ax + b$, an example of a cost function is:

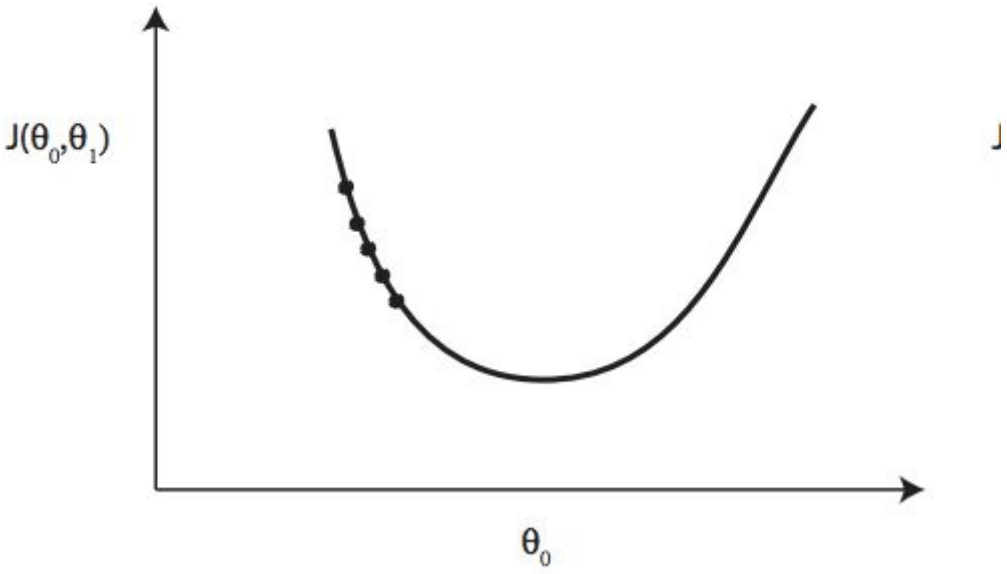
$$J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Here, J represents the cost function. θ_0 represents the constant in the equation -> b. θ_1 represents the weight given to the variable -> a. h_0 represents the current best guess of our equation. $x^{(i)}$ and $y^{(i)}$ represent a data point from our data. m is the number of data points we have.

This function calculates the average distance of data from the equation. It cycles through the data from 1 to M. It then feeds the x values into the equation the we made up, which is represented by h_0. It compares the y value produced from the function (the value of $h_0(x^{(i)})$) with the correct Y value from the data -> $y^{(i)}$. The error in the guess is $h_0(x^{(i)}) - y^{(i)}$, this is the difference between the guess produced by our equation and the real value of y for that data point. To avoid negative values that may cancel out positive values we square this difference. We then divide the total amount of this difference by m to get an average. In Summary ->

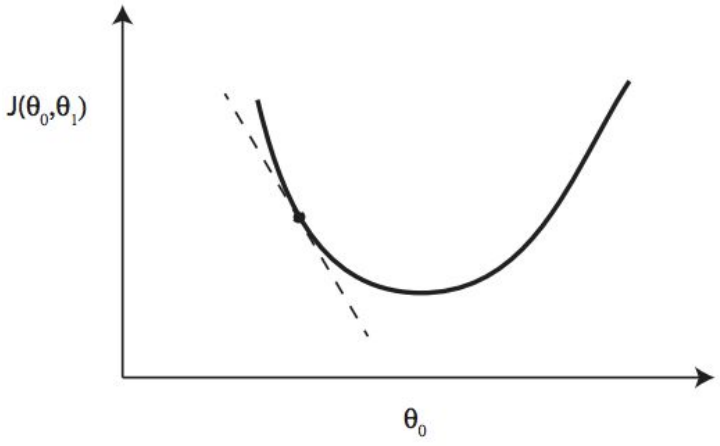
- Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$
- Parameters: θ_0, θ_1
- Cost Function: $J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$
- Goal: Select θ_0 and θ_1 that minimise $J(\theta_0, \theta_1)$

To get a good model for our datal we want an equation that will produce the lowest possible cost. Do find our lowest possible cost model we can use something called “Gradient Descent”. To understand gradient descent it’s good to think in graphs. Let’s imagine our cost model looks like this:



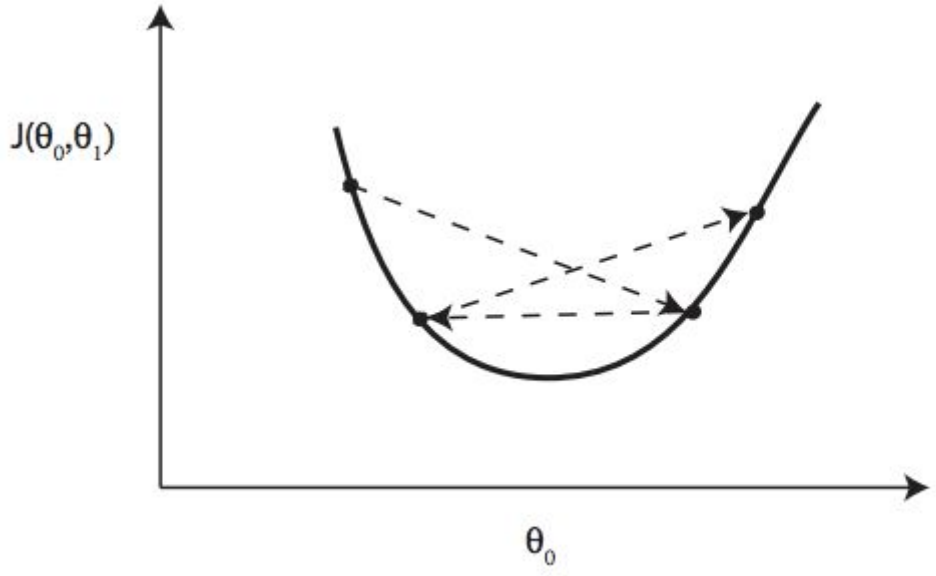
Here the equation (or model) we are trying to calculate has the general form $f(x) = \theta^1 X$. On this graph $J(\theta_0, \theta_1)$ is the cost we have calculated using a cost model (just an equation for calculating the distance data points are from our equation) and θ^1 is the weight we put on our x variable. From the graph we can see that if we change θ^1 the "cost" or the average distance our data has from our model changes. We can also see that there is a value of θ_0 that would have a lowest possible cost.

Gradient descent works by using integration. Remember that we start by guessing a value for θ_0 . Gradient descent works by checking the slope of curve at the point of our guess. Like this:



Remember that a negative slope means that the graph is moving downwards, so if we increase θ^1 we will move closer to the minimum possible value for our cost function. This will give us a new guess for θ^1 . We can continue to use this until we reach the actual

minimum value. The amount we move by is called the learning rate. We can pick the learning rate ourselves. If the learning rate is too low it will take us forever to step to the local minimum. If it is too high we risk jumping over the minimum value repeatedly like this:



So we update out value for θ_0 like this

$$-\alpha\frac{\partial}{\partial\theta_0}J(\theta_0,\theta_1)$$

Alpha or α is the learning rate.
Then we multiply this by the rate of change, ie the change in the cost function or $\delta J(\theta_0,\theta_1)$ for small changes in θ_0 or $\delta\theta_0$. Which looks like:

$$\theta_0-\alpha\frac{\partial}{\partial\theta_0}J(\theta_0,\theta_1)$$

When we build these models it is useful to know what is the likelihood that it is the correct model. To do this we must first make some assumptions. When we collect data we get “noise”. Noise is random variance in our measurements. For example, if we measured the weight of a sample of people the “noise‘ would the elements that make our measurements incorrect. It could be the weight of the people's shoes or clothes. The effect of noise is that our measurements are not 100 percent accurate, which means our model will never be 100% accurate. Since we can never have a perfect model we must do regressions to find a model which best suits our data. We can use “probabilistic interpretation” to interpret the probability of our model being correct.

To help us account for noise we can assume two things. We first assume that the noise has a gaussian distribution. A gaussian distribution just describes a bell curve. A general equation for this distribution is:

$$f_Z(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Z-\mu)^2}{2\sigma^2}}.$$

Remember this models the average effect of the noise on each data point. It accounts for what makes our model not 100% accurate. The second thing we assume is that the noise has a mean of 0. This means that noise in either direction will cancel out. The effect of this is that noise should not majorly influence our model.

So to calculate the likelihood of our model being correct, for each data point we check the difference between what our model produces. Which is:

$$(h_0(x^{(i)}) - y^{(i)})^2$$

Then we put this through our model for how our noise should have affected our output.

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{(y^{(i)} - h_\theta(x^{(i)}))^2}{2}}$$

This function is the same as the function for our bell curve. Remember that the mean of our noise is assumed to be zero. So μ is no longer in the equation. And our Z is replaced with the difference between the output from our model and the value we observed from our data.

To calculate the overall probability that our model is correct we check the probability that it is correct for each data point and then multiply these probabilities to get an overall probability for the model. The probability for the model is the product of the probability for the model being correct at each data point. In maths speak this is:

$$f_{D|\Theta}(d|\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y^{(i)} - h_\theta(x^{(i)}))^2}{2}}$$

Here, the θ is our guess for the model. The d is the data. m is the number of datapoints that we have and the messy equation is modeling the difference between our guesses for output and expected output accounting for noise. The bell curve is to account for noise. So

what we have now is a model for calculating the likelihood of our data given the guess for our model. To get the best guess we want to find the model that produces the maximum value from this equation. This is the model that would maximise the likelihood of producing our data. But first! With some sneaky maths we can make this model simpler.

$$f_{D|\Theta}(d|\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y^{(i)} - h_\theta(x^{(i)}))^2}{2}}$$
$$= \frac{1}{\sqrt{2\pi}^m} e^{-\sum_{i=1}^{m} \frac{(y^{(i)} - h_\theta(x^{(i)}))^2}{2}}$$

- Taking logs: $\log f_{D|\Theta}(d|\theta) = \log \frac{1}{\sqrt{2\pi}^m} - \sum_{i=1}^{m} \frac{(y^{(i)} - h_\theta(x^{(i)}))^2}{2}$

We can then ignore the $\log \frac{1}{\sqrt{2\pi}^m}$ since it will be the same for every iteration. The result is that we want to maximise the second part of the equation. Ie:

$$-\sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)}))^2$$

Maximising a negative is the same as minimizing a positive. So we are trying to minimise:

$$\sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)}))^2$$

Which is our cost function from before. This maths was simply to highlight a few assumptions that we made when using our cost function. Namely:

- Noise is additive
- Noise on each observation is independently and identically distributed
- Noise is Gaussian

Changing how we account for our noisel would lead to a different loss function. We can use gradient descent or some more mathsing to get find the model that will minimise our function.

This mathsing looks like this:

- Select θ to maximise $\log L(\theta) = -\frac{1}{2}\sum_{j=1}^{n}(y_j - \theta x_j)^2$
- Compute derivative with respect to θ:

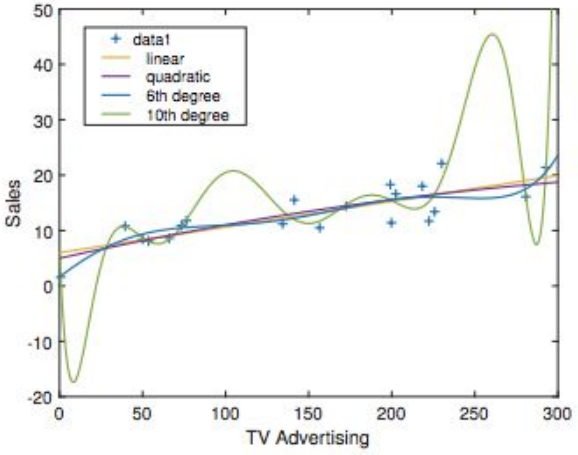
$$\frac{dL}{d\theta} = \sum_{j=1}^{n}(y_j - \theta x_j)x_j = \sum_{j=1}^{n} y_j x_j - \theta \sum_{j=1}^{n} x_j^2$$

- Set derivative equal to 0 and solve for θ:

$$\sum_{j=1}^{n} y_j x_j - \theta \sum_{j=1}^{n} x_j^2 = 0$$

$$\Rightarrow \quad \theta = \frac{\sum_{j=1}^{n} y_j x_j}{\sum_{j=1}^{n} x_j^2}$$

We make models so that we can make predictions. For this reason it is important our model fits the data we currently have and generalises to new data. Trying to evaluate how well a model generalises is called regularisation. It comes down to selecting the correct formula for a line. Eg does the data look like a cubic or quadratic graph. If we use a high polynomial eg: $x^7 = 0$ our model might “overfit”. This is where our model trys to reach every data point thus ends up fitting the noise:



There can conversely be “underfitting” if our polynomial is too low.

What we’ve been doing with our model selection is called Maximum Likelihood Estimation. It means that we our trying to find the model that has the maximum likelihood of being correct. We’ve been using principles from Bayes theorem to help us calculate our likelihood:

$$f_{\Theta|D}(\vec{\theta}|d) = \frac{f_{D|\Theta}(d|\vec{\theta}) f_{\Theta}(\vec{\theta})}{f_D(d)}$$

posterior *likelihood* *prior*

f(d), which is the probability of us having our data, is a constant so that has been ignored. Also $f_{\Theta}(\Theta)$ is our prior belief that our model is correct. Since we assumed all models to be equally likely to be correct we ignored this term. So we were saying that the probability that our model is correct given the data is that same as the probability of getting our data given the model. That allowed us to do our maths which resulted in this equation:

$$\sum_{i=1}^{m}(y^{(i)} - h_\theta(x^{(i)}))^2$$

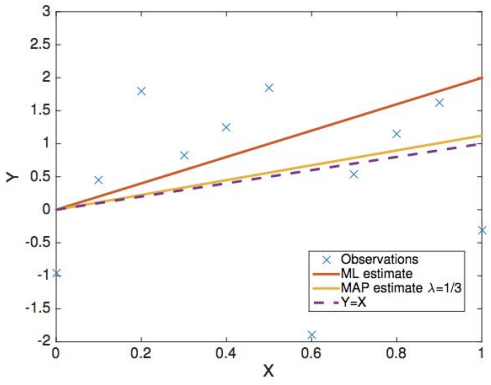
However, there's another form of model estimation called Maximum a Posteriori or "MAP" estimation. In this form of estimation we do not assume that all models have the same probability of being correct. In other words we need to take account of $f_{\Theta}(\Theta)$ from bayes theorem. So we have an extra piece to our equation:

$$\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{1}{\lambda}\sum_{j=1}^{n}\theta_j^2$$

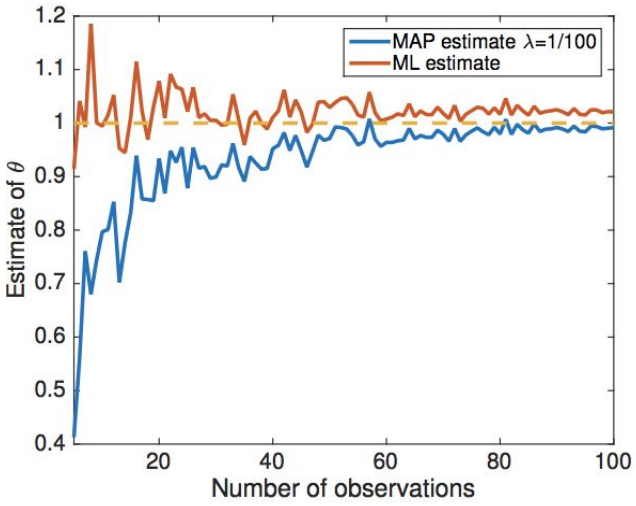
Where θ_j^2 represents our prior beliefs about a model and λ is a constant that we decide. What λ does is account for how certain we are about our prior. If λ is near zero then the effect of the second part of the equation will be greater than if λ was very large.

MAP vs Maximum Likelihood Estimation

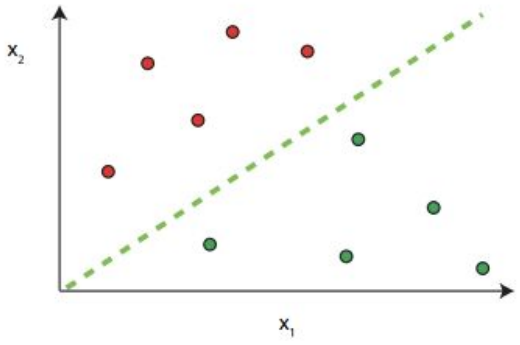
Difference between MAP and ML really kicks in when we only have a small number of observations, yet still need to make a prediction. Our prior beliefs are then especially important.



The more data we have the less impact our priors will have. This is because with more observations we can reason about or data under more certainty.



This methodology doesn't work for every problem! We also have a set of problems called classification problems. This is things like, are emails spam, are transactions fraudulent etc. These questions normally need to be assigned to a box. To solve these problems we first graph the data, we then try to draw a line to classify data into categories (not all data is separable like this either).



Since we're doing things a little differently we need to use a different cost function to calculate our line. The function we use is

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \log(1 + e^{-y^{(i)}\theta^T x^{(i)}})$$

This function calculates how wrong our parameter selection is. Parameters are the constants we multiple our variables by to get our line. For example the general equation

$$\theta X = Y$$

our parameter would be θ.

What we do is y(i) is the classifier. For our email example it could take the value -1 or 1 depending on whether the email was spam or not. The x(i) is the data point vetor we are looking at. Our function multiplies the x(i) vector by the θ^T vector to get the value for that equation. This is the same as subing the value for x(i) into our equation. It then multiplies this by the negative for the actual Y value at that data point. The will minimize the average distance from points that are incorrectly classified to the division boundary. As before to get the correct values for θ we make a guess and then use gradient descent. To use gradient descent we get the derivative of our cost function and then use the slope of that derivate to help us approach a local minimum value for that function.

The derivative looks like this:

For $J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \log(1 + e^{-y^{(i)}\theta^T x^{(i)}})$:

- $\frac{\partial}{\partial\theta_j} J(\theta) = \frac{1}{m}\sum_{i=1}^{m} -y^{(i)}x_j^{(i)} \frac{e^{-y^{(i)}\theta^T x^{(i)}}}{1+e^{-y^{(i)}\theta^T x^{(i)}}}$
- (Remember $\frac{d\log(x)}{dx} = \frac{1}{x}$, $\frac{d\exp(x)}{dx} = exp(x)$ and chain rule $\frac{df(z(x))}{dx} = \frac{df}{dz}\frac{dz}{dx}$)

We use this to update θ as follows:

$$\text{for } j{=}0 \text{ to } n \left\{tempj := \theta_j + \frac{\alpha}{m}\sum_{i=1}^{m} y^{(i)}x_j^{(i)} \frac{e^{-y^{(i)}\theta^T x^{(i)}}}{1+e^{-y^{(i)}\theta^T x^{(i)}}}\right\}$$
$$\text{for } j{=}0 \text{ to } n \left\{\theta_j := tempj\right\}$$

The sneaky α that has been added to our cost function is the learning rate. This is the size of the jumps we want to take towards the local minimum. Doing this over all of our data will help us draw a line that will separate the data into different classes.

We can do some probabilistic interpretation to see how confident we are in our model. probabilistic interpretation asks what is the probability that our model is correct. To find the probability of it being correct we can look at the probability of us getting our data given our guess at the model. This looks like:

$$P(d|\theta) = \prod_{i=1}^{m} \frac{1}{1 + e^{-y\theta^T x}}$$

At the bottom we can see our cost function, which calculates how correct each guess was. Then we use the inverse to get a probability.

The closer our bottom value give us to one the more certain we are about our prediction.