BITS F464: Machine Learning
Tutorial Book

Kaivalya Rawal

April 18, 2018

Contents

Chapter 1

Probability

1.1 Brief Review

1.1.1 Probability Distributions of Random Variables

This chapter is meant to be used as a quick review of some of the basic probability concepts that will be used later in this course. For a more thorough review, please complete the course **MATH F113: Probability and Statistics**. A very brief (and incomplete) review is provided in this section, with a few sample questions to refresh your memory.

Probability distributions refer to the probability than an individual observation will take a particular value (or lie in a range of values). These can be represented diagrammatically, or mathematically through precise functions. These functions must satisfy the following properties:

1. They have "parameters", that allow us to change the shape of the distribution.

2. Irrespective of the value of its parameters, the total area under a probability distribution sums to 1.

3. The probability of obtaining values between say x_1 and x_2 is the area under the probability distribution between x_1 and x_2. This probability (and clearly, the area) must lie between 0 and 1.

Probability distributions are associated with random variables, whose characteristics they are supposed to describe. Informally, a random variable is one that takes one of several pre-defined values. Exactly which value will be taken is not known because of the element of chance. A probability distribution describes how likely each value is. That is, one view of a probability distribution is as the relative frequency - in the long run - with which a random variable will take its different values when sampled repeatedly.

1.1.2 Definitions and Notation

Before proceeding with the sample questions please review the following terms and notations:

P(X) The *probability distribution* of the *random variable* X (random variables are conventionally always capitalised).

P(X=x) The exact probability of the random variable X taking the value x. Often shortened to $P(x)$.

P(X|Y) The probability of distribution of X given Y.

Expectation The expected value of a random variable or function of a random variable. Often denoted $E(X)$ for a random variable X.

$$E(h(X)) = \sum h(X)P(X)$$
$$provided \sum |h(X)|P(X) < \infty$$

Independence We say that the random variables $X$and Y are independent if $P(X|Y) = P(X)$, or equivalently, $P(Y|X) = P(Y)$.

Conditional Independence Two random variables X and Y are considered conditionally independent given a third variable Z if $P(X|Y,Z) = P(X|Z)$ or $P(Y|X,Z) = P(Y|Z)$. Note that conditional independence does not imply independence. This can also be denoted $X \perp\!\!\!\perp Y \mid W$.

1.1.3 Required Math

1.1.3.1 Integration By Parts

Recall the formula for integration by parts:

$$\int u\,dv \;=\; uv - \int v\,du$$

1.2 Sample Questions

1.2.1 Shapes, Letters, and Colours

You are given a set of 13 squares and circles, 9 of which are coloured black and the rest are coloured white. Each object also has either the letter A or B on it. There are: 2 black squares with an A, 4 black squares with a B and 1 black circle with an A. Of the remaining, there is 1 white square and 1 white circle each with an A. Here is a diagrammatic representation:

Let *Black* denote the set of black objects, *White* denote the set of white objects, *Square* denote the set of square objects, A the set of objects with an

A and so on. We are drawing shapes at random from this space at random. Assuming all objects are equally likely (the so-called Principle of Indifference) answer the following questions:

1.2.1.1 What is $P(A)$?

Simple counting over the entire set gives us $P(A) = 5/13$.

1.2.1.2 What is $P(A|Square)$?

Counting the all the A's onle from amongst the square shapes yields, $P(A|Square) = 3/8$.

1.2.1.3 Are A and $Square$ independent?

No, since $P(A) \neq P(A|Square)$.

1.2.1.4 Are A and $Black$ independent?

$P(A|Black) = 1/3 \neq P(A) = 5/13$. So, no A and $Black$ are not independent.

1.2.1.5 Are A and $Square$ conditionally independent given $Black$?

$P(A|Square, Black) = 1/3 = P(A|Black)$. So, A and $Square$ are conditionally independent, given $Black$.

1.2.1.6 Are A and $Square$ conditionally independent given $White$?

$P(A|Square, White) = 1/2 = P(A|White)$. So, A and $Square$ are conditionally independent, given $White$.

1.2.1.7 The Law of Total Probability gives us: $P(A) = P(A, White) + P(A, Black)$. Verify that the law holds in this case.

$P(A) = P(A|White)P(White) + P(A|Black)P(Black) = 1/2 \cdot 4/13 + 1/3 \cdot 9/13 = 5/13$. Hence verified.

1.2.1.8 Using Bayes' Rule, calculate $P(Black|A)$?

$$\begin{aligned} P(Black|A) &= \frac{P(A|Black)P(Black)}{P(A)} \\ &= \frac{P(A|Black)P(Black)}{P(A|Black)P(Black) + P(A|White)P(White)} \\ &= \frac{(1/3)(9/13)}{(1/3)(9/13) + (1/2)(4/13)} \\ &= \frac{3}{5} \end{aligned}$$

1.2.2 Urns and Marbles

There are two urns (Urn1 and Urn2). Urn1 has 2 red marbles and 2 blue marbles. Urn2 has 1 red and 3 blue marbles. The urn labels are now covered and a coin is flipped to select an urn. Having selected an urn, we draw a marble from the urn.

1.2.2.1 The marble is red. What is the probability that the urn selected was Urn1?

This can be solved by simply drawing the probability tree. The first branch has a binary choice with probability 0.5 of selecting Urn1 or Urn2. For Urn1 there is a probability of $0.5 : 0.5$ of selecting $red : blue$ marbles. The corresponding choices are $0.25 : 0.75$ for Urn2. Conditioning on a red marble being drawn will lead to $P(Urn1) = \frac{0.25}{0.25+0.125} = \frac{2}{3}$.

This question can also be solved by Bayes Rule (as was probably done in the probability section of your math class in 12th grade). To find $P(Urn1|Red)$, we use:

$$\begin{aligned} P(Urn1|Red) &= \alpha P(Red|Urn1)P(Urn1) \\ P(Red|Urn1) &= 0.5 \\ P(Urn1) &= 0.5 \end{aligned}$$

Therefore, $P(Urn1|Red) = 0.25\alpha$.
Similarly

$$\begin{aligned} P(Urn2|Red) &= \alpha P(Red|Urn2)P(Urn2) \\ P(Red|Urn1) &= 0.25 \\ P(Urn1) &= 0.5 \end{aligned}$$

Therefore, $P(Urn2|Red) = 0.125\alpha$.
Since $0.25\alpha + 0.125\alpha = 1$, $\alpha = 1/0.375$ and $P(Urn1|Red) = \frac{0.25}{0.375} = \frac{2}{3}$

1.2.3 Cricket and Temperature

In a typical English summer, the probability that the temperature falls below 10 degrees Celsius is 0.4. In that case, the English cricket team wins with probability 0.75. The probability that the temperature is between 10 and 30 degrees Celsius is 0.4, in which case the English team wins with probability 0.65. The probability that the temperature is greater than 30 degrees is 0.2 and in that case, the English team wins with probability 0.55.

1.2.3.1 The English team has won. What is the probability that the temperature was below 10 degrees?

To find $P(T < 10|Won)$:

$$\begin{aligned} P(Won|T < 10) &= 0.75 \\ P(T < 10) &= 0.4 \\ P(T < 10|Won) &= \alpha P(Won|T < 10)P(T < 10) \\ &= \alpha \cdot 0.75 \cdot 0.4 \\ &= 0.3\alpha \end{aligned}$$

Similarly, we get

$$\begin{aligned} P(Won|30 > T > 10) &= 0.65 \\ P(30 > T > 10) &= 0.4 \\ P(30 > T > 10|Won) &= \alpha P(Won|30 > T > 10)P(30 > T > 10) \\ &= \alpha \cdot 0.65 \cdot 0.4 \\ &= 0.26\alpha \end{aligned}$$

and

$$\begin{aligned} P(Won|T > 30) &= 0.55 \\ P(T > 30) &= 0.2 \\ P(T > 30|Won) &= \alpha P(Won|T > 30)P(T > 30) \\ &= \alpha \cdot 0.55 \cdot 0.2 \\ &= 0.11\alpha \end{aligned}$$

Since $0.3\alpha + 0.26\alpha + 0.11\alpha = 1$, $\alpha = 1/0.67$ and $P(T < 10|Won) = 0.3\alpha = 0.48$.

1.2.4 Mean and Variance

You are given a random variable X with a discrete probability distribution categorised by:

$$P(X = x) = \begin{cases} 1/3 & if\ x = -1, 0, 1 \\ 0 & if\ otherwise \end{cases}$$

1.2.4.1 What is $\mu_X = E(X)$?

$$\mu_X = \frac{1}{3} \cdot (-1 + 0 + 1) = 0$$

1.2.4.2 What is $Var(X) = E[(X - \mu_X)^2]$?

$$\begin{aligned} Var(X) &= E[(X-\mu_X)^2] \\ &= E(X^2) - (\mu_X)^2 \\ &= E(X^2) - (0)^2 \\ &= E(X^2) \\ &= \frac{1}{3} \cdot ((-1)^2 + 0^2 + 1^2) \\ &= \frac{2}{3} \end{aligned}$$

1.2.4.3 Repeat the calculations for the following random variable.

$$P(X = x) = \begin{cases} 1/3 & if\ x = -2, 0, 2 \\ 0 & if\ otherwise \end{cases}$$

Notice that the points are spread out more around the same central location. Hence the mean will be unchanged, and the variance will increase.

$$\mu_X = \frac{1}{3} \cdot (-2 + 0 + 2) = 0$$

$$\begin{aligned} Var(X) &= E[(X-\mu_X)^2] \\ &= E(X^2) - (\mu_X)^2 \\ &= E(X^2) \\ &= \frac{1}{3} \cdot ((-2)^2 + 0^2 + 2^2) \\ &= \frac{8}{3} \end{aligned}$$

1.2.5 Means of some popular distributions.

1.2.5.1 Let X be the random variable denoting the number of dots that come up on the throw of a six-sided die. What is $E(X)$?

Notice that this is a discrete uniform distribution.

$$E(X) = \frac{1}{6} \cdot (1 + 2 + 3 + 4 + 5 + 6) = 3.5$$

1.2.5.2 Let X be a random variable denoting the number of successes in n IID Bernoulli trials, each with probability p of success. What is $E(X)$?

This is the binomial distribution.

The expected value for the number of successes is given by:

$$\begin{aligned} E(X) &= \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} \\ &= \sum_{k=0}^{n} k\frac{n!}{(n-k)!k!}p^k(1-p)^{n-k} \\ &= \sum_{k=1}^{n} np\frac{(n-1)!}{(n-k)!(k-1)!}p^(k-1)(1-p)^{n-k} \end{aligned}$$

Let $i = k - 1$. Then:

$$\begin{aligned} E(X) &= \sum_{k=1}^{n} np\frac{(n-1)!}{(n-k)!(k-1)!}p^(k-1)(1-p)^{n-k} \\ &= np\sum_{i=0}^{n-1}\frac{(n-1)!}{(n-1-i)!i!}p^i(1-p)^{n-1-i} \\ &= np \end{aligned}$$

1.2.5.3 Let X be an exponential random variable with PDF $f(X = x) = \lambda e^{-\lambda x}$ $(x > 0)$. What is $E(X)$?

$$\begin{aligned} E(X) &= \int_0^\infty x\lambda e^{-\lambda x}dx \\ &= -xe^{-\lambda x}\big|_0^\infty + \int_0^\infty e^{-\lambda x}dx \\ &= \frac{1}{\lambda} \end{aligned}$$

1.2.6 Power Law P.D.Fs

1.2.6.1 A continuous real-valued variable has a power-law P.D.F. if $P(x) = Cx^{-\alpha}$ $(\alpha > 0)$. In fact, this function diverges as $x \to 0$: so how can it be a P.D.F.?

It cannot. In most real-world problems, the power-law is only a good fit after some minimum value of $x = x_{min}$. So, as a first approximation, we can take the probabilities of such variables as being modelled by this (modified) P.D.F. instead:

$$P(X = x) = \begin{cases} Cx^{-\alpha} & if\ x \geq x_{min} \\ 0 & if\ x < x_{min} \end{cases}$$

1.2.6.2 Find an expression for C in the (modified) P.D.F. in the previous question.

We know that every probability distribution must sum up to 1. Over its entire domain. Therefore, we have $\int_{-\infty}^{\infty} f(x)dx = 1$

$$\begin{aligned}
\int_{-\infty}^{\infty} f(x)dx &= 1 \\
\int_{-\infty}^{x_{min}} (0)dx + \int_{x_{min}}^{\infty} Cx^{-\alpha}dx &= 1 \\
\int_{x_{min}}^{\infty} Cx^{-\alpha}dx &= 1 \\
C[\frac{x^{-\alpha+1}}{-\alpha+1}]_{x_{min}}^{\infty} &= 1 \\
0 - C \cdot \frac{x_{min}^{-\alpha+1}}{-\alpha+1} &= 1 \\
C &= -(\frac{1-\alpha}{x_{min}^{1-\alpha}}) \\
C &= (\alpha-1)x_{min}^{\alpha-1}
\end{aligned}$$

1.2.6.3 Find the expected value for the random variable having the (modified) power-law P.D.F.

$$\begin{aligned}
E(X) &= \int_{x_{min}}^{\infty} xCx^{-\alpha}dx \\
&= C\int_{x_{min}}^{\infty} x^{-\alpha+1}dx \\
&= \frac{C}{2-\alpha}x^{2-\alpha}\Big|_{x_{min}}^{\infty}
\end{aligned}$$

1.2.6.4 Power-laws with $\alpha \leq 2$ have no finite mean. This means that as we start taking more and more samples from such populations, we will start to see the mean diverge. How can this happen?

What must start to happen is every so often samples with a very large value of the mean must come up. That is, the fluctuation in the means must be very large. There are many natural phenomena that exhibit power-law behaviours with divergent means or divergent variances (or both) like this.

1.3 Summary and Further Reading

- Random variables are described by probability distributions, which are mathematical functions with some special constraints.
- X and Y are considered conditionally independent given Z if $P(X|Y,Z) = P(X|Z)$ or $P(Y|X,Z) = P(Y|Z)$.
- The Law of Large Numbers states that the average of the results obtained from a large number of trials will be close to the expected value.
- The Central Limit Theorem states that the properly normalized sum of independent random variables tends towards a normal distribution even if the original variables are not normally distributed.
- Power law distributions are defined only beyond a certain threshold x_{min} and have no mean for $\alpha \leq 2$ and no variance for $\alpha \leq 3$.

1.4 References

The following sources were used in the creation of this chapter.

- https://en.wikipedia.org/wiki/Random_variable
- https://en.wikipedia.org/wiki/Conditional_independence
- https://en.wikipedia.org/wiki/Law_of_large_numbers
- https://en.wikipedia.org/wiki/Central_limit_theorem
- M. Cargal - Discrete Mathematics for Neophytes - Birkhauser 1988

Chapter 2

Maximum Likelihood Estimation and Expectation Maximization

2.1 Computing the Maximum Likelihood Estimate

2.1.1 Intuition behind the Maximum Likelihood Estimate (MLE)

To get an intuitive feel for the maximum likelihood estimate (MLE), consider an experiment involving say $n = 100$ tosses on an unbiased coin. Most people would expect *around* 50 heads and 50 tails. Now lets modify this experiment. If the coin was *biased* with probability of heads $p(heads) = 0.8$, most people would automatically and obviously correct their estimates to 80 heads and 20 tails respectively.

Modifying this experiment further, lets flip it to estimate the parameter of the distribution (in this case $\theta = p(heads)$) from the data observed, instead of the other way round. So now we want to guess the bias of the coin after having observed 80 heads and 20 tails from a total of 100 tosses. Again, most people would immediately be able to guess that $p(heads) = \theta = 0.8$. This is actually the MLE estimate for θ. Now if I was to repeat the experiment, but this time perform $10,000$ tosses, and obtain exactly $8,000$ heads, most people would still guess $\theta = 0.8$, but this time, their confidence in their estimate would be far greater. This intuitive process of guessing parameters actually leads us to the maximum likelihood estimate. This is what makes MLE so special, easy, and intuitive.

From this example, the following features of MLE become clear:

- It is a statistical estimate, and not a certainty (the actual value of θ may

not be exactly 0.8)

- It is a point estimate (as opposed to a range)
- It is the most obvious estimate - and the de-facto estimate people tend to use
- It is easy and natural to compute, though not necessarily unbiased
- The estimate improves with an increase in sample size (mathematically speaking, it achieves the *Cramer Rao bound* as the sample tends to infinity)

What we have done here, is essentially take a likelihood function representing a probability $L = P(D|\theta)$, and found which value of D, given a fixed value of θ would maximise it. Then we flipped this around, and estimated which value of θ would maximise $P(d|\theta)$ if d was given (ie. if data was observed). This forms the basis of Maximum Likelihood esdtimation. We often use maximise the log-likelihood function $l = log(L)$ instead of the likelihood function for simplicity.

2.1.2 Required Math

2.1.2.1 Lagrange Multipliers

This technique is used to maximise a general function $f(x, y)$ under some constraint of the form $g(x, y) = 0$.

This is done by introducing a *Lagrange multiplier* λ. The function f is maximised at the point where the *Lagrange function* L is stationary. Thus, to maximise f, we just need to find the stationary points of L.

$$L(x, y, \lambda) = f(x, y) - \lambda \cdot g(x, y)$$

We can find the stationary points using the usual method of setting the first derivative to zero.

You can learn more about this in the Appendix of this book.

2.1.2.2 The identity function

The identity function is a mathematical trick used to define a function that is one for certain inputs, and zero for all others. This is written so:

$$I(x = j) = \begin{cases} 1 & if\ x = j \\ 0 & if\ x \neq j \end{cases}$$

2.2 Sample Questions on MLE

2.2.1 Maximise $f(x_i|p) = p^{x_i}(1-p^{x_i})$ given the $x_i = \{1,0,0,1,0,1,1,1,1,1\}$.

When solving MLE questions, we start by making the usual *IID* assumption - each observation in the dataset is independent of the others, and follows the same distribution. Therefore in this case, we get:

$$\begin{aligned} L = P(d|p) &= \prod f(x_i) \\ L &= p^7(1-p)^3 \\ l = log(L) &= 7log(p) + 3log(1-p) \\ \frac{\partial l}{\partial p} &= \frac{7}{p} - \frac{3}{1-p} = 0 \\ p &= \frac{7}{10} \end{aligned}$$

2.2.2 Observing closely, this question can be thought of as a binomial distribution. Find the MLE estimate for p using the binomial distribution formula given 7 successes and 3 trials, as before.

We will rewrite the likelihood function, and proceed in an identical manner.

$$\begin{aligned} L = P(d|p) &=_{10} C_7 p^7(1-p)^3 \\ l = log(L) &= log(_{10}C_7) + 7log(p) + 3log(1-p) \\ \frac{\partial l}{\partial p} &= \frac{7}{p} - \frac{3}{1-p} = 0 \\ p &= \frac{7}{10} \end{aligned}$$

2.2.3 Let $x_1, x_2, \ldots, x_n$ be a sample of observations from a Poisson distribution with parameter λ. Find the maximum likelihood estimate of λ in terms of the x_i and n.

The likelihood function is the probability of observing $x_1, x_2, \ldots, x_n$ in n independent draws from a Poisson distribution.

Again, we shall utilise the *IID* assumption, write the likelihood function as a product, take its logarithm, and differentiate it and set the result to zero.

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$
$$L(\lambda) = \frac{e^{-n\lambda}\lambda^{\sum_i x_i}}{x_1! \cdots x_n}$$
$$\log L(\lambda) = -n\lambda + \sum x_i \log\lambda - \log c$$
$$\lambda_{MLE} = \frac{\sum_i x_i}{n}$$

2.2.4 Let $x_1, x_2, \ldots, x_n$ be a sample from an exponential distribution, which has a density function $f(X = x) = \lambda e^{-\lambda x}$ $(x > 0)$. Derive the maximum likelihood estimate of λ in terms of the x_i and n.

The likelihood function for continuous random variables is the joint p.d.f.

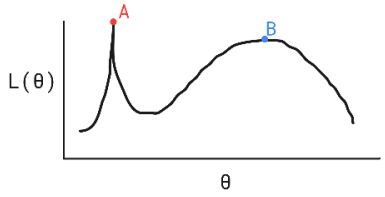
Following a procedure identical to the previous question gives us:

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$
$$L(\lambda) = \lambda^n e^{-\lambda \sum x_i}$$
$$\lambda_{MLE} = \frac{n}{\sum x_i}$$

2.2.5 Let $x_1, x_2, \ldots, x_n$ be a sample from a normal distribution with parameters μ and σ^2. Derive maximum likelihood estimates of μ and σ^2.

$$
\begin{aligned}
L = P(X = x_i) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \\
L = P(D|\mu, \sigma) &= \prod \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \\
&= (\frac{1}{\sqrt{2\pi}\sigma})^n e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2} \\
l = log(L) &= -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \\
\mu_{MLE} &= \frac{1}{n}\sum x_i \\
\sigma^2_{MLE} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{MLE})^2
\end{aligned}
$$

2.2.6 For the following plot of the Likelihood function, describe the location of the maximum likelihood estimate.



MLE is a point estimate. It yields the single point where the value of θ maximises $L(\theta)$. Thus, the estimate will be located at point A. Notably, in most scenarios B would actually represent a better estimate of θ, and probably be what most people would be looking for.

2.2.7 Extend the coin toss MLE estimation problem to a dice roll.

In the coin toss problem, there is a single parameter, p, the probability of heads on a single coin toss. For the equivalent dice problem, we have 6 parameters $\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6$, each representing the probability of the dice landing with corresponding face. Note that $\sum \theta = 1$.

Note that in an experiment conducted we may observe for example $X = \{3, 4, 5, 2, 2, 1, 6, 5, 2, 4\}$, and we could represent the number of times a face i appeared by n_i, giving us $\{n_1 = 1, n_2 = 3, n_3 = 1, n_4 = 2, n_5 = 2, n_6 = 1\}$. In this case, it is an $\theta_i = \frac{n_i}{\sum n_i}$ is an intutitive guess. As it happens more often thatn not, this intuitive guess is actually the maximum likelihood estimate! Lets prove this mathematically.

We need a way to define our likelihood function in terms of the observed x_i. To do this, we start with the simple observation for an $m = 6$ faced dice:

$$P(x|\theta) = \prod_{j=1}^{m} \theta_j^{I(x=j)}$$

so, for example for $x_0 = 3$, we have $P(3) = \theta_3$, as described before.

$$\begin{aligned} P(3|\theta) &= \prod_{j=1}^{m} \theta_j^{I(x_i=j)} \\ &= \theta_1^{I(3=1)} \cdot \theta_2^{I(3=2)} \cdot \theta_3^{I(3=3)} \cdot \theta_4^{I(3=4)} \cdot \theta_5^{I(3=5)} \cdot \theta_6^{I(3=6)} \\ &= \theta_1^0 \cdot \theta_2^0 \cdot \theta_3^1 \cdot \theta_4^0 \cdot \theta_5^0 \cdot \theta_6^0 \\ &= \theta_3 \end{aligned}$$

We can now define our likelihood function for given that the observed x_i are of the form $X = \{x_0 \dots x_n\}$:

$$\begin{aligned} L = P(D|\theta) &= \prod_{i=1}^{n} \prod_{j=1}^{m} \theta_j^{I(x_i=j)} \\ &= \prod_{j=1}^{m} \prod_{i=1}^{n} \theta_j^{I(x_i=j)} \\ &= \prod_{j=1}^{m} \theta_j^{\sum_{i=1}^{n} I(x_i=j)} \\ &= \prod_{j=1}^{m} \theta_j^{n_j} \\ l = log(L) &= \sum_{j=1}^{m} n_j log(\theta_j) \end{aligned}$$

Recall that n_j is the number of times the face marked j appears in the dataset.

Thus we need to maximise $l = \sum_{j=1}^{m} n_j log(\theta_j)$, subject to the constraint $\sum \theta = 1$.

Using lagrange multipliers, this can be shortened to a system of equations.

$$f - \lambda g = \sum n_j log(\theta_j) - \lambda(\sum \theta_j - i)$$

Now differentiating with each θ_j, and setting to zero, we shall get:

$$\begin{aligned}\frac{\partial(f-\lambda g)}{\partial\theta_j} &= 0\\ \frac{n_1}{\theta_1}-\lambda &= 0\\ &\vdots\\ \frac{n_6}{\theta_6}-\lambda &= 0\end{aligned}$$

This yields

$$\begin{aligned}\theta_1 &= \frac{n_1}{\lambda}\\ \theta_2 &= \frac{n_2}{\lambda}\\ &\vdots\\ \theta_6 &= \frac{n_6}{\lambda}\end{aligned}$$

Which, when combined with $\sum\theta = 1$, gives us

$$\begin{aligned}\sum\theta_j &= 1\\ \sum\frac{n_j}{\lambda} &= 1\\ \lambda &= \frac{1}{\sum n_j}\\ \theta_j &= \frac{n_j}{\sum n_j}\end{aligned}$$

Thus proving the familiar $\theta_i = \frac{n_i}{\sum n_i}$. Notice that this is simply the multidimensional equivalent of the coin toss.

2.3 The Expectation Maximization (EM) Algorithm

MLE requires extensive analytical examination to obtain an estimate for a parameter. This can often be difficult, or even impossible. The EM algorithm provides a computational means of calculating a point estimate. Moreover, it also works when labels are missing from the dataset.

Lets consider the first MLE problem of coin tosses. Suppose we have data from a sequence of Bernoulli trials, each with a probability of success p. We have

seen how the maximum likelihood estimate of p is s/n where s is the number of successes observed in the data, and n is the total number of trials. This can now be extended to a model with more than 1 coin.

Each Bernoulli trial is like tossing a biased coin with probability p of landing heads. Maximum likelihood estimation can now be trivially applied on both coins. If we have coins A and B, then to obtain the maximum likelihood estimates of the parameters p_A and p_B, we repeatedly do the following:

1. Randomly pick a coin (for the moment with equal probability)
2. Toss the coin some number (say 10) times
3. Record the experiment number, the number of heads observed; and which coin was chosen
4. Repeat the experiment (that is, return to Step 1)

For example, if we repeated this experiment 5 times, giving coin A or B 10 tosses each time, we might get the following data:

R1(B): *HTTTHHTHTH*

R2(A): *HHHHTHHHHH*

R3(A): *HTHHHHHTHH*

R4(B): *HTHTTTHHTT*

R5(A): *THHHTHHHTH*

The maximum likelihood estimates of p_A and p_B will simply be the proportion of heads on tosses for which A and B were used, ie $P_A = s_A/n_A = 24/30$, and $P_B = s_B/n_B = 9/20$. This is trivial. Now lets generalise this experiment further by considering a scenario without labels.

Consider this harder problem. You still have 2 coins A and B, and you have data from say 5 repetitions:

R1: *HTTTHHTHTH*

R2: *HHHHTHHHHH*

R3: *HTHHHHHTHH*

R4: *HTHTTTHHTT*

R5: *THHHTHHHTH*

Since there is no record of which coin was used in each of the five trials, it is non-trivial to estimate P_A and P_B. A possible way to do this is:

1. Start with some guess about p_A and p_B . Call these $p_A^{(0)}$ and $p_B^{(0)}$.

2. For each of the repeats *R1–R5*, calculate $P_A = P(D|p_A^{(0)})$ and $P_B = P(D|p_B^{(0)})$ (where D is the sequence of Head's and Tails's on that repetition). If $P_A > P_B$, then assume A was used, otherwise assume B was used. That is, manually label each of the five trials as having used A or B.

3. Now we have a complete table like we did before, and can calculate maximum likelihood estimates as before.

We can iterate through this entire procedure, using the maximum likelihood estimate obtained at the end of step (3) as the guesses in step (1) in the subsequent iteration.

The "expectation-maximisation" algorithmm or EM algorithm is a refinement of this basic idea, except we don't label any trial as having been observed exclusively from coin A or coin B. Instead, the current guesses for p_A and p_B are used to obtain new "weighted" instances of each trial.

The weights are proportional to $P(coin = A|D)$ and $P(coin = B|D)$. These are proportional to $P(D|coin = A)P(coin = A)$ and $P(D|coin = B)P(coin = B)$. Since the coins were known to be chosen with equal probability($P(coin = A) = P(coin = B)$), then $P(coin = A|D) = \alpha P(D|coin = A)$ and $P(coin = B|D) = \alpha P(D|coin = B)$. We can calculate this using the current guess of p_A and p_B.

For example, with an initial guess of 0.60 for p_A and 0.50 for p_B, we find the following:

R1. D1 = HTTTHHTHTH. Then, $P(Coin = A|D1) = 0.45$ and $P(Coin = B|D1) = 0.55$. So, we create a new set of weighted instances generated by both coins:

$$CoinA : (5 \cdot 0.45) = 2.2H \ \ and \ \ (5 \cdot 0.45) = 2.2T$$
$$CoinB : (5 \cdot 0.55) = 2.8H \ \ and \ \ (5 \cdot 0.55) = 2.8T$$

$$(Both) : 5H \ \ and \ \ 5T \ \ \ (as\ before)$$

R2. D2 = HHHHTHHHHH. Then, $P(Coin = A|D2) = 0.80$ and $P(Coin = B|D2) = 0.2$, and weighted instances:

$$CoinA : (9 \cdot 0.8) = 7.2H \ \ and \ \ (1 \cdot 0.8) = 0.2T$$
$$CoinB : (9 \cdot 0.2) = 1.8H \ \ and \ \ (1 \cdot 0.2) = 0.8T$$

$$(Both) : 9H \ \ and \ \ 1T \ \ \ (as\ before)$$

and so on, for repetitions *R3–R5*. In each case, the number of heads in the weighted instances is the expected value of the number of heads, given the

current estimate of the parameters. The totals obtained after all calculations are:

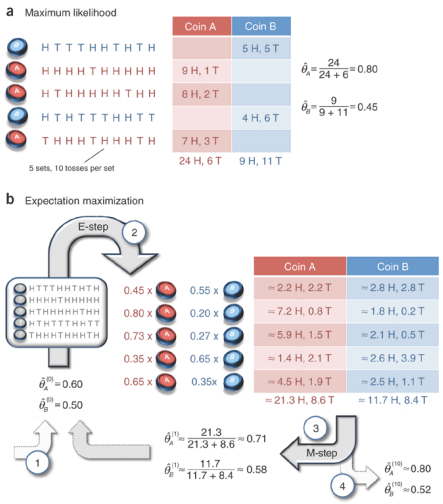
$$CoinA: 21.3H \ \ and \ \ 8.6T$$
$$CoinB: 11.7H \ \ and \ \ 8.4T$$

This gives a new estimate of $p_A = \frac{21.3}{21.3+8.6} = 0.71$ and $p_B = \frac{11.7}{11.7+8.4} = 0.58$.

This is one iteration of the EM algorithm. It can be used repeatedly to converge to the final estimates of parameters p_A and p_B.

2.4 Summary and Further Reading

- The general MLE problem can be solved by the following steps:
  - Given $d = \{\ldots x_i \ldots\}$
  - Assume P_θ, a probability function that represents each (independent, and identically distributed) x_i.
  - Construct the likelihood function: $L = P(D|\theta)$
  - maximize by differentiating the log likelihood: $\theta_{MLE} = argmax_\theta(L(\theta))$
- This can be compared with EM in the following graphic (from *Nature*):



- Both MLE and EM provide point estimates, and generally interpretable results.

- The existence, and uniqueness of MLE estimates is not guaranteed.
- MLE estimates may not be unbiased.
- MAP estimates improve upon MLE by incorporating prior knowledge.
- The EM algorithm greedily maximizes an objective function, iteratively updating parameter values.
- EM estimates are susceptible to starting estimates. Gibbs sampling offers an alternative to EM that improves upon this.
- They are both particularly susceptible to overfitting.

2.5 References

The following sources were used in the creation of this chapter.

1. https://onlinecourses.science.psu.edu/stat504/node/28
2. https://en.wikipedia.org/wiki/Maximum_likelihood_estimation
3. https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm
4. http://www.cmi.ac.in/~madhavan/courses/dmml2017/literature/EM_algorithm_2coin_example.pdf
5. http://www.nature.com/nbt/journal/v26/n8/full/nbt1406.html
6. Chuong B Do, Serafim Batzoglou - What is the expectation maximization algorithm? - Springer Nature, 2008

Chapter 3

Linear Models

3.1 Prerequisites

This section explains some terms and reviews some mathematical concepts. These can also be found in the appendices at the end of this book.

3.1.1 Definitions

3.1.1.1 Bias

Given an estimator $\hat{\theta}$ for a variable θ, we define

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

An estimator is said to be unbiased if this value is equal to zero, and biased otherwise.

3.1.1.2 Variance

Given an estimator $\hat{\theta}$ for a variable θ, we define

$$Variance(\hat{\theta}) = E((\hat{\theta} - E(\theta))^2)$$

Variance can never be negative.

3.1.1.3 MSE

The *Mean Squared Error* or MSE for an estimator $\hat{\theta}$ is defined as

$$MSE(\hat{\theta}) = E((\hat{\theta} - \theta)^2)$$

It is just one measure of estimator accuracy out of many possible metrics.

3.1.2 Required Math

3.1.2.1 Lagrange Multipliers

This technique is used to maximise a general function $f(x, y)$ under some constraint of the form $g(x, y) = 0$.

This is done by introducing a *Lagrange multiplier* λ. The function f is maximised at the point where the *Lagrange function* L is stationary. Thus, to maximise f, we just need to find the stationary points of L.

$$L(x, y, \lambda) = f(x, y) - \lambda \cdot g(x, y)$$

We can find the stationary points using the usual method of setting the first derivative to zero.

You can learn more about this in the Appendix of this book.

3.2 The Linear Model

This is by no means a thorough treatment of linear models. For a complete review please refer to your lecture notes. We will *not* cover the Gauss Markov Assumptions that lead to the linear model, but simply describe the final model.

3.2.1 Best Fit Lines

The objective of building linear models is to capture relationships between independent and dependent variables. The independent variables are known as *predictors* and denoted x,and the dependent variables are known as response variables denoted y. We shall *assume* that these variables are linearly related, and try to determine this relationship using a sample drawn from a population. This is known as the line of best fit.

A common misconception people often have is that there can never be points with identical x values having different y values in the population. This is not true. A linear model simply captures the expected value of y, $E(y|x)$ for every value of x. The population is described by

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\beta_0$and $\beta_1$are constants, but the residual ϵ_i is not. Note that for each x, there can be multiple y values, but there is a single $E(Y)$ for each x. That is

$$E(Y) = \mu_y = \beta_0 + \beta_1 x$$

The linear model we construct is thus

$$\hat{y_i} = a + bx_i$$

Here $\hat{y}$ is an estimator for y. Linear models are thus used to predict (estimate) y values for points where only the x value is known. a and b are point estimates of $\beta_0$and $\beta_1$respectively.

Linear models are often written as $y_i = E(Y_i|X_i) + \epsilon_i$ instead. This is merely a notational difference. Note that $E(Y|X)$ is just the expected value of y for a given x, which is known to be $\beta_0 + \beta_1 x$. So, with either notation, the study of linear models thus involves one thing at its core: the estimation of β_0 and β_1.

We will investigate this in the sample questions by calculating a and b. This is usually done by defining a function upon the dataset that varies with different values of a and b. This "cost" function is minimised to obtain an estimate for β_0 and β_1. Intuitively, the cost function represents how far the model estimates a and b are from the true population parameters $\beta_0$and β_1. Minimising this provides us with the best possible estimates of a and b.

3.2.2 Gradient Descent

Gradient descent is a numeric technique to obtain values for estimators that iteratively improves the estimator value. If we want to estimate θ that minimises some cost function $f(\theta)$, we *repeat till satisfaction*:

$$\theta_{i+1} = \theta_i - \eta \nabla_{\hat{\theta}}$$

Here each θ_i is an estimate for θ, which we hope to gradually improve with increasing i. $\nabla_{\hat{\theta}}$ represents the gradient of the cost function, and η is an arbitrary constant.

That is, we keep updating our value of θ using this equation, until we are convinced that we have found a reasonable value for our estimator. Note that η is a constant here that needs to be tuned manually for each different cost function $f(\theta)$.

3.2.3 Regularisation

The traditional cost function that is minimised when doing linear regression is a measure of the "inaccuracy" of the model. Sometimes, along with this we add a second term to the cost representing the "complexity". Thus we minimise a total cost that measures both inaccuracy and complexity. We do this to ensure that the models we produce are not only accurate (low inaccuracy), but also simple (low complexity).

$$Total\ Cost = Accuracy\ Cost + Complexity\ Cost$$

3.3 Sample Questions

3.3.1 Show that $MSE = (variance) + (bias)^2$

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= \text{E}\left[(\hat{\theta} - \theta)^2\right] \\
&= \text{E}\left[\left(\hat{\theta} - \text{E}[\hat{\theta}] + \text{E}[\hat{\theta}] - \theta\right)^2\right] \\
&= \text{E}\left[\left(\hat{\theta} - \text{E}[\hat{\theta}]\right)^2 + 2\left(\hat{\theta} - \text{E}[\hat{\theta}]\right)\left(\text{E}[\hat{\theta}] - \theta\right) + \left(\text{E}[\hat{\theta}] - \theta\right)^2\right] \\
&= \text{E}\left[\left(\hat{\theta} - \text{E}[\hat{\theta}]\right)^2\right] + \text{E}\left[2\left(\hat{\theta} - \text{E}[\hat{\theta}]\right)\left(\text{E}[\hat{\theta}] - \theta\right)\right] + \text{E}\left[\left(\text{E}[\hat{\theta}] - \theta\right)^2\right] \\
&= \text{E}\left[\left(\hat{\theta} - \text{E}[\hat{\theta}]\right)^2\right] + 2\left(\text{E}[\hat{\theta}] - \theta\right)\text{E}\left[\hat{\theta} - \text{E}[\hat{\theta}]\right] + \left(\text{E}[\hat{\theta}] - \theta\right)^2 \\
&= \text{E}\left[\left(\hat{\theta} - \text{E}[\hat{\theta}]\right)^2\right] + 2\left(\text{E}[\hat{\theta}] - \theta\right)\left(\text{E}[\hat{\theta}] - \text{E}[\hat{\theta}]\right) + \left(\text{E}[\hat{\theta}] - \theta\right)^2 \\
&= \text{E}\left[\left(\hat{\theta} - \text{E}[\hat{\theta}]\right)^2\right] + \left(\text{E}[\hat{\theta}] - \theta\right)^2 \\
&= \text{Variance}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2
\end{aligned}$$

3.3.2 Find the gradients $\frac{\partial MSE}{\partial b}$ and $\frac{\partial MSE}{\partial a}$ for the linear model $Y = a + bX$.

For this model ie ($Y = a + bX$), all the points (x_i, y_i) will be of the form $y_i = a + bx_i + \epsilon_i$. Thus the *MSE* can be written as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

Now differentiating wrt b we get

$$\begin{aligned}
\frac{\partial MSE}{\partial b} &= \frac{1}{n}\sum_{i=1}^{n} 2(y_i - (a + bx_i))\frac{\partial(y_i - (a + bx_i))}{\partial b} \\
&= \frac{2}{n}\sum_{i=1}^{n}(y_i - (a + bx_i))(-x_i) \\
&= -\frac{2}{n}\sum_{i=1}^{n} x_i(y_i - (a + bx_i))
\end{aligned}$$

Similarly for a we get

$$\frac{\partial MSE}{\partial a} = \frac{1}{n}\sum_{i=1}^{n} 2(y_i - (a + bx_i))\frac{\partial (y_i - (a + bx_i))}{\partial a}$$
$$= -\frac{2}{n}\sum_{i=1}^{n}(y_i - (a + bx_i))$$

Thus we finally get

$$\nabla_b MSE = \frac{\partial MSE}{\partial b} = -\frac{2}{n}\sum_{i}^{n} x_i(y_i - (a + bx_i))$$

$$\nabla_a MSE = \frac{\partial MSE}{\partial a} = -\frac{2}{n}\sum_{i}^{n}(y_i - (a + bx_i))$$

3.3.3 Derive expressions for a and b that fit the linear model $Y = a + bX$.

We will use the partial derivatives computed above, and set them to zero.

$$\frac{\partial MSE}{\partial a} = -\frac{2}{n}\sum_{i=1}^{n}(y_i - (a + bx_i)) = 0$$

$$\frac{\partial MSE}{\partial b} = -\frac{2}{n}\sum_{i=1}^{n} x_i(y_i - (a + bx_i)) = 0$$

Solving for 'a' and 'b' yields

$$b = \frac{\sum_{i=1}^{n} x'_i y'_i}{\sum_{i=1}^{n} x'^2_i} \qquad a = \bar{y} - b\bar{x}$$

where $x'_i = x_i - \bar{x}$ and $y' = y_i - \bar{y}$.

3.3.4 Show that the point $(\bar{x}, \bar{y})$ lies on the linear model with least MSE.

We already know that the MSE is minimised when $\nabla_a MSE = \frac{\partial MSE}{\partial a} = -\frac{2}{n}\sum_i^n (y_i - (a + bx_i)) = 0$

$$\frac{\partial MSE}{\partial a} 0$$

$$-\frac{2}{n}\sum_{i}^{n}(y_i - (a + bx_i)) 0$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i \frac{1}{n}\sum_{i=1}^{n}(a + bx_i)$$

$$a + b\frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{y}a + b\bar{x}$$

Thus, for the linear model $Y = a + bX$, if we minimise the MSE to obtain estimates for a and b, the resultant model necessarily passes through the mean ($\bar{x}$,$\bar{y}$) of the data.

3.3.5 Find the gradient update equations for a and b.

The estimatiors a and b in the linear model $Y = a + bX$ can be computed using gradient descent like any other estimators. Therefore:

$$a_{i+1} = a_i - \eta\nabla_a \qquad \textit{and} \qquad b_{i+1} = b_i - \eta\nabla_b$$

3.3.6 Describe the changes required to fit the model $Y = a + bX + cX^2$ instead of the linear model $Y = a + bX$ to the data.

There are no changes required to the algorithm. We just have an extra coeefficient in the model, and thus need to perform the approprite gradient descent update to find its value along with the values of the other coefficients.

3.3.7 Explore the performance of gradient descent for various cost functions.

There are several challenges that arise in using gradient descent for various kinds of cost functions.

- **Convex functions:** For a convex cost function, gradient descent can start with any arbitrary estimate and eventually reach the global minimum, thus finding the optimal value for the estimator. Linear regression with MSE as the cost is always convex.
- **Flat functions:** Any cost function which has areas of zero slope (or "flat" areas) can get stuck in a situation where gradient descent doesn't update the estimator value since the gradient is zero.

- **Multiple minima:** Gradient descent can get stuck in a local minima if a function has more than one minima. In this scenario, the performance of gradient descent will depend hugely on the initial guess for the estimator.

There are many modifications to gradient descent that have been suggested, to overcome some of these problems.

- **Decaying η:** In this version of gradient descent, the value of η is initially fairly large, and is gradually reduced. Most implementations allow for an exponential, linear, or stepwise decay. The intuition here is that while η is large, the estimator can be prevented from getting stuck at a local minimum. Finally, when the η value is low, the estimator can iterativel optimise to the exact point that minimises the cost function without the danger of moving too far away.
- **Random restarts:** In this technique, the usual gradient descent algorithm is applied multiple times, with different starting guesses for the estimator. Each run yields a (potentially) different local minimum. The lowest of these is assumed to be the global minimum.
- **Momentum:** To provide gradient descent with the capability to climb out of local minima when moving in a general direction, a momentum term can be added to the gradient descent update. This requires keeping track of the change in the estimator value ($\Delta\theta_i$) in each step. One version of this update formula is $\theta_{i+1} = \theta_i - \eta\nabla_\theta + \alpha\Delta\theta_i$.

3.3.8 Given a dataset with features $(x_1, x_2 \dots x_n, y)$, construct a linear model for Y.

If the dataset consisted of two numeric variables, we could use the familiar $Y = a + bX$ model. This can be extended to three features $Y = a + bX_1 + cX_2$.

This can thus be naturally extended to all n independent variables (*predictors*). The dependent variable (*response*) can thus be modelled using the formula:

$$\begin{aligned} Y =& \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots \beta_n X_n \\ =& \beta_0 + \sum_{i=1}^{n} \beta_i X_i \end{aligned}$$

However, this fails if any of the predictors is non-numeric.

3.3.9 Construct a linear model provided one of the input predictors is binary.

Consider a linear model for data of the form (x, y, z). Here, x is a numeric predictor, y is a predictor taking either value 0 or 1, and z is the response variable.

We can now split the data into two portions, based on the value of y. Thus, we solve two separate regression tasks, separately fitting the model $Z = a + bX$ at first to all the data points that have $y = 0$, and then to those that have $y = 1$. This yields two regression lines: $Z = a_1 + b_1X$ for use whenever $y = 0$ and $Z = a_2 + b_2X$ for use when $y = 1$.

Now, whenever we need to make a prediction, we first check the value of y, and use the appropriate regression line based on its value.

$$Z = a + bX \quad \begin{cases} Z = a_1 + b_1X & if\ Y = 0 \\ Z = a_2 + b_2X & if\ Y = 1 \end{cases}$$

3.3.10 Show that the combined model constructed above is identical to a single model $Z = a+bX+cY+dXY$.

The approach described above of splitting the data based on the value of y is actually identical to fitting a single regression model to the entire dataset with some auxiliary variables. To do so, we first augment the data tuples by adding an xy term $(x, y, z) \longrightarrow (x, y, xy, z)$. We then fit the model $Z = a+bX+cY+dXY$.

To see how this is identical to the previous case, consider the equation $Z = a + bX + cY + dXY$ when $Y = 0$ and when $Y = 1$.

$$Z = a + bX + cY + dXY = \begin{cases} Z = a + bX & if\ Y = 0 \\ Z = (a + c) + (b + d)X & if\ Y = 1 \end{cases}$$

Comparing this equation with those obtained above, we can calculate a, b, c, and d in terms of a_1, a_2, b_1, and b_2.

$$\begin{aligned} a &= a_1 \\ b &= b_1 \\ c &= a_2 - a_1 \\ d &= b_2 - b_1 \end{aligned}$$

Thus the two models are identical.

3.3.11 Show that the least square estimate for b is the same as the MLE estimate if $e_i \sim_{i.i.d} N(0, \sigma^2)$

Recall that the least squares estimate of b required $\frac{\partial MSE}{\partial b} = 0$.

$$\begin{aligned}\frac{\partial MSE}{\partial b} &= 0 \\ -\frac{2}{n}\sum_{i=1}^{n} x_i(y_i - (a + bx_i)) &= 0 \\ \sum_{i=1}^{n} x_i(y_i - (a + bx_i)) &= 0\end{aligned}$$

Now lets calculate the MLE estimate of b. Recall the definition of MLE.

$$b_{MLE} = arg\, max_b P(D|b)$$

We now maximise by setting the derivative of the log likelihood to zero.

$$\begin{aligned}Likelihood(b) &= P(D|b) \\ L(b) &= \prod_{i=1}^{n} P(y_i|x_i, b) \\ log(L(b)) &= \sum_{i=1}^{n} log(P(y_i|x_i, b)) \\ l(b) &= \sum_{i=1}^{n} log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{(y_i-(a+bx_i))^2}{2\sigma^2}\right)}\right) \\ l(b) &= n\, log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (a + bx_i))^2 \\ \frac{\partial l(b)}{\partial b} &= -\frac{1}{2\sigma^2}(-\frac{2}{n})\sum_{i=1}^{n} x_i(y_i - (a + bx_i)) \\ \frac{\partial l(b)}{\partial b} &= \frac{1}{n\,\sigma^2}\sum_{i=1}^{n} x_i(y_i - (a + bx_i))\end{aligned}$$

$$\begin{aligned}\frac{\partial l(b)}{\partial b} &= 0 \\ \frac{1}{n\,\sigma^2}\sum_{i=1}^{n} x_i(y_i - (a + bx_i)) &= 0 \\ \sum_{i=1}^{n} x_i(y_i - (a + bx_i)) &= 0\end{aligned}$$

This shows that either ways, the same constraint is finally being applied to compute b. Therefore a linear model minimising MSE is the same as an MLE estimate for the linear model.

3.3.12 How would the gradient descent update change if we introduced regularisation?

So far we have used MSE as our metric to compare models. The best model is the one with least MSE, and thus we try to minimise it. This can tend to overfit the data at times. As we have seen, this estimate is the same as the MLE estimate, and MLE can overfit frequently. When we introduce a regularisation term to our cost function, we decide to penalise our models not just for high MSE, but also for high complexity, thus directly reducing overfitting.

The cost function becomes

$$J(\theta) = MSE(\theta) + \alpha f(\theta)$$

Where $f(\theta)$ is a regularization term (to prevent over-fitting and to get a smoother curve) and α is a hyperparameter that determines the extent of influence the regularization term has on the total cost $J(\theta)$.

There are two kinds of regularisation used commonly when building linear models:

1. L1 Regulariaztion: $|\theta| = \sum_{i=1}^{n} |\theta_i|$
2. L2 Regulariaztion: $||\theta|| = \sum_{i=1}^{n} \theta_i^2$

The gradient descent update, when using a regularised cost function, is:

$$\theta = \theta - \lambda \frac{\partial(J(\theta))}{\partial\theta}$$

3.3.13 While constructing linear models of the form $Y = a + b \cdot \phi(X)$, show that choice of ϕ between sigmoid and tanh doesn't matter.

Even if the predictors and response have a relationship that can be modelled using linear regression, it is possible that the data requires a nonlinear transformation. For example, there may not be a direct relationship between x and y, but y may be varying linearly with the sin of x. In this case, the dataset (x_i, y_i) can be transformed into another dataset $(\sin(x_i), y_i)$. A linear model is now expressive enough to capture this relationship.

In this question, we consider a scenario where a nonlinear transform (sigmoid and tanh respectively) is being applied to the data to build the linear model, and show that if a model can be built for one of the two with a certain accuracy, then it can also be built for the other, with an identical accuracy.

We have $\phi_1(x) = \text{sigmoid}(x) = \frac{1}{1+e^{-x}}$ and $\phi_2(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. We know these are related by

$$\tanh(x) = 2\text{sigmoid}(2x) - 1$$

.

Also, our linear models are of the form:

$$Y = w_0 + \sum_{j=1}^{n} w_j \text{sigmoid}(x)$$
$$Y = u_0 + \sum_{j=1}^{n} u_j \tanh(x)$$

If we let $x' = \frac{x}{2}$

$$\begin{aligned}
Y =& w_0 + \sum_{j=1}^{n} w_j \text{sigmoid}(x) \\
=& w_0 + \sum_{j=1}^{n} w_j \text{sigmoid}(2x') \\
=& w_0 + \sum_{j=1}^{n} \frac{w_j}{2}(2\text{sigmoid}(2x') - 1 + 1) \\
=& w_0 + \sum_{j=1}^{n} \frac{w_j}{2}(2\text{sigmoid}(2x') - 1) + \sum_{j=1}^{n} \frac{w_j}{2} \\
=& (w_0 + \sum_{j=1}^{n} \frac{w_j}{2}) + \sum_{j=1}^{n} \frac{w_j}{2} \tanh(x') \\
=& u_0 + \sum_{j=1}^{n} u_j \tanh(x)
\end{aligned}$$

Where $u_j = \frac{w_j}{2}$ *for* $j = (1 \dots n)$ and $u_0 = w_0 + \sum_{j=1}^{n} \frac{w_j}{2}$.

3.3.14 Show that minimising a regularized error function with *complexity cost* $= a^q + b^q$ and a regularized error function subject to $a^q + b^q = \eta$ is the same.

Lets first rewrite the expression for unregularised and regularised error. Note that the additional fraction $\frac{1}{2}$ has no effect on the cost and is introduced only for convenience.

$$\textit{Unregularised Error} = \frac{1}{2} \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

$$\textit{Regularised Error} = \frac{1}{2} \sum_{i=1}^{n} (y_i - (a + bx_i))^2 + \frac{\lambda}{2}(a^q + b^q)$$

We need to show that when the regularised error is optimised subject to the constraint $a^q + b^q = \eta$, the model obtained is the same as when optimising for the regularised error.

To do this, lets use the technique of lagrange multipliers to minimise unregularised error subject to this constraint. We have

$$f(a,b) = \frac{1}{2}\sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

Subject to the condition (again, with the fraction added for convenience).

$$g(a,b) = \frac{1}{2}(a^q + b^q - \eta) = 0$$

This yields the Lagrange function

$$L(a,b,\lambda) = \frac{1}{2}\sum_{i=1}^{n}(y_i - (a + bx_i))^2 + \frac{\lambda}{2}(a^q + b^q - \eta)$$

$$\begin{aligned} L(a,b,\lambda) =& \frac{1}{2}\sum_{i=1}^{n}(y_i - (a + bx_i))^2 + \frac{\lambda}{2}(a^q + b^q - \eta) \\ =& \frac{1}{2}\sum_{i=1}^{n}(y_i - (a + bx_i))^2 + \frac{\lambda}{2}(a^q + b^q) - \frac{\lambda\eta}{2} \\ =& \frac{1}{2}\sum_{i=1}^{n}(y_i - (a + bx_i))^2 + \frac{\lambda}{2}(a^q + b^q) + CONST. \\ =& Regularised\ Error + CONST. \end{aligned}$$

Thus the stationary point of the Lagrange function coincides with that of the regularised error, yielding identical linear models. Note that if $q = 1$, we get L_1 regularisation, and if $q = 2$, we get L_2 regularisation.

3.4 Summary and Further Reading

- We sample from a population (x_i, y_i) belonging to $y_i = \beta_0 + \beta_1 x_i + \epsilon_i = E(Y_i|X_i) + \epsilon_i$ using the linear model $\hat{y_i} = a + bx_i$. Here $\hat{y}$ is an estimator for y used to predict (estimate) y values for points where only the x value is known. a and b are point estimates of $\beta_0$and $\beta_1$respectively.
- The gradient descent update is $\theta_{i+1} = \theta_i - \eta\nabla_{\hat{\theta}}$ and is used to find the values of parameters a and b.
- A regularisation term is often added to the cost to prevent overfitting and fit simpler models.

- $MSE = (variance)+(bias)^2$. The *Cramer Rao* bound implies the variance is greater than zero, and thus MSE is always positive. The lowest MSE may not be achieved with an unbiased estimator, and requires a thorough treatment of the bias variance tradeoff.
- The linear model with least MSE passes through $(\bar{x}, \bar{y})$, the mean of the data.
- The simple linear model described in this chapter can be extended to multiple predictor variables, and even to categoric variables.
- If the residuals are i.i.d Gaussian $e_i \sim_{i.i.d} N(0, \sigma^2)$, then the MLE estimate for b is the same as the least squares estimate.
- We can apply a nonlinear transformation to the predictor variables before fitting our linear model to capture more complicated relationships than simple linear models.

3.5 References

The following sources were used in the creation of this chapter.

1. https://onlinecourses.science.psu.edu/stat501/
2. http://ruder.io/optimizing-gradient-descent/
3. https://en.wikipedia.org/wiki/Lagrange_multiplier
4. https://en.wikipedia.org/wiki/Mean_squared_error#Proof_of_variance_and_bias_relationship
5. http://dept.stat.lsa.umich.edu/~kshedden/Courses/Stat401/Notes/401-multreg.pdf
6. Christopher M. Bishop - Pattern Recognition And Machine Learning - Springer 2006

Chapter 4

Support Vector Machines

4.1 Brief Review

4.2 Sample Questions

4.3 Final Review

Appendix A

Definitions

A.1 Random Variable

A random variable, conventionally capitalised and written as X, is a variable whose possible values are numerical outcomes of a random phenomenon. A random varaible may be discrete or continuous, as it may take on values from a finite or infinite set respectively. Random variables are associated with their probability distributions.

A.2 Probability Distributions

The probability distribution of a random variable describes the probability of the random variable taking each possible value from the set of allowed values. These distributions may be defined graphically or as mathematical functions. In fact, the probability distribution of a discrete random variable can simply be a list of probabilities associated with each of its possible values. This is also sometimes called the probability function or the probability mass function. Correspondingly, Ccntinuous random variables, are defined by their probability density functions.

A.3 Conditional Independence

Two random variables X and Y are considered conditionally independent given a third variable Z if $P(X|Y,Z) = P(X|Z)$ or $P(Y|X,Z) = P(Y|Z)$. This can also be denoted $X \perp\!\!\!\perp Y \mid W$. Note that conditional independence does not imply independence, or vice versa.

A.4 Bias

Given an estimator $\hat{\theta}$ for a variable θ, we define

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

An estimator is said to be unbiased if this value is equal to zero, and biased otherwise.

A.5 MSE

The *Mean Squared Error* or MSE for an estimator $\hat{\theta}$ is defined as

$$MSE(\hat{\theta}) = E((\hat{\theta} - \theta)^2)$$

It is just one measure of estimator accuracy out of many possible metrics.

A.6 Variance

Given an estimator $\hat{\theta}$ for a variable θ, we define

$$Variance(\hat{\theta}) = E((\hat{\theta} - E(\theta))^2)$$

Variance can never be negative.

Appendix B

Required Math

This appendix is far from complete. It very briefly covers those math concepts that you may have forgotten since the *Math F111*, *Math F112*, *Math F113* and *Math F211* courses. It is recommended that you at least review *Math F113 (Probability and Statistics)* thoroughly, beyond the material covered in this appendix.

B.1 Lagrange Multipliers

This technique is used to maximise a general function $f(x, y)$ under some constraint of the form $g(x, y) = 0$.

This is done by introducing a *Lagrange multiplier* λ. The function f is maximised at the point where the *Lagrange function* L is stationary. Thus, to maximise f, we just need to find the stationary points of L.

$$L(x, y, \lambda) = f(x, y) - \lambda \cdot g(x, y)$$

We can find the stationary points using the usual method of setting the first derivative to zero.

As an example, consider finding the discrete probability distribution that maximises infomational entropy (maximisation of Shannon Entropy). This requires maximising the following objective function (Entropy):

$$f(p_1, p_2, \cdots, p_n) = -\sum_{j=1}^{n} p_j \log_2 p_j.$$

This is subject to the condition that $\{p_i \dots p_n\}$ sum up to one.

$$g(p_1, p_2, \cdots, p_n) = \sum_{j=1}^{n} p_j = 1$$

Finding the Lagrange function, we get

$$\frac{\partial}{\partial p_k}(f + \lambda(g-1))\bigg|_{p_k=p_k^*} = 0$$

Which in turn leads to a system of n equations $k = 1, \cdots, n$

$$\frac{\partial}{\partial p_k}\left\{-\left(\sum_{j=1}^{n} p_j \log_2 p_j\right) + \lambda\left(\sum_{j=1}^{n} p_j - 1\right)\right\}\bigg|_{p_k=p_k^*} = 0.$$

Differentiating, we get

$$-\left(\frac{1}{\ln 2} + \log_2 p_k^*\right) + \lambda = 0$$

Which shows that all p_k^* are equal since they depend only on λ. That is $\sum_j p_j = 1$.

$$p_k^* = \frac{1}{n}$$

Giving the final answer as the uniform distribution, as expected.

B.2 The Identity Function

The identity function is a mathematical trick used to define a function that is one for certain inputs, and zero for all others. This can be defined as:

$$I(x = j) = \begin{cases} 1 & if\ x = j \\ 0 & if\ x \neq j \end{cases}$$

Different values of j represent different differnet identity functions. Once an identity function has been defined (fixed j), x, the input to the function variaes, and the function yields either a 1 or a 0 as per this definition.

B.3 Integration By Parts

When integrating a product of two functions (u and dv), the following formula comes in handy:

$$\int u\,dv \;=\; uv - \int v\,du$$

Thus, one of the functions (u) needs to be differentiated, and the other function (dv) needs to be integrated. This can simplify an integral at times.