Question	Points	Score
Descriptive Statistics	6	
Probability	10	
Expectation	11	
Distributions	16	
Inference	17	
Simulation	9	
Regression	10	
Modelling	21	
Total:	100	

Instructions

- The formulae sheet for the unit is at the end of the exam. This supports answers for many of the questions.
- For all questions, part marks are given for working. So it is best to show working. Then you can still receive part marks if your final answer is wrong.

Numeric/calculation errors: for each question, they should loose at most 1/2 a mark if they make numeric errors in the question.
e.g., Q1.1(a) if the get Q_1 wrong, then their calc of IQR should be marked as if it was correct.
e.g., if they give a final formula but don't calculate, deduct 1/2 max.
i.e., ignore numeric errors during marking a question, but if they had them, then subtract 1/2 at the end

Keep packets together and keep them in order. Only mark one packet at a time. Packets must be kept in order to find booklets for individual students.

Make a note of **common errors** so we can evaluate the sorts of mistakes commonly made. Also, be on the lookout for unusual but correct answers.

1. Descriptive Statistics: ..... *Total: 6 marks*

(1) Consider the following set of numbers: 12, -25, 2, 3, 21, 8, 10, 14, 3, 18, 21. For each of the questions below, state your answer, showing working if necessary.

(2 marks) (a). What is the range and the inter-quartile range?

> **Solution:** Ordering data yields: -25 2 3 3 8 10 12 14 18 21 21. Range = 46. According to formula, for Q_1, $p = 3$ and $q = 0$ and for Q_3, $p = 9$ and $q = 0$. Thus $Q_1 = 3$ and $Q_3 = 18$ thus IQR=15.
> 1/2 for range, 1/2 for Q_1 and Q_3 each, 1/2 for IQR.

(2 marks) (b). Given the sum ($\sum x$) is 87 and the sum of squares ($\sum x^2$) is 2357, what is the sample standard deviation?

> **Solution:** Mean is 7.909. Mean of squares is 214.27. Sample variance is $11/10 * (214.27 - 7.909 * 7.909) = 166.89$. Solution is 12.92.
> 1/2 for mean, 1.5 if sample variance correct (however done), lose half if forgot 11/10.

(1 mark) (c). Identify a candidate outlier and compute a new sample standard deviation with the outlier removed.

> **Solution:** The outlier is -25. Mean becomes 11.2 and mean of squares now 173.2. Solution 7.28.
> 1/2 for picking the outlier, 1/2 for recomputing.

(1 mark) (2) Suppose two quantitive variables, x and y, have a negative correlation. What can we say about their relationship?

Solution: Generally as x increases, y decreases. There maybe a more complex relationship but correlation only measures the linear relationship.
1/2 for incr/decr, 1/2 for linearity.

2. Probability: ..... *Total: 10 marks*

(1 mark) (1) You are studying the side-effects of a new weight loss drug. You enroll 500 obese patients and their GPs to do a long term trial of the drug. You have GPs fill out a table for their patients with entries as follows:

- weight at start of trial
- weight at 12 months
- complications at 12 months

You then plan to do an analysis of the complications and weight loss. But 178 patients drop out of the study before the end of the 12 months. What potential problems are their with your design? Is there any sampling bias in the recording of complications, and if so what is it called?

Solution: The design should have asked for follow-up questions on any participants who dropped out. Thus it has participation bias.
1/2 for description. 1/2 for name of it.

(2) A group of 3 boys and 7 girls are lined up in a random order. Note that there are $10! = 3628800$ such orders.

(2 marks) (a). What is the probability that a particular boy is in the 3rd position and a particular girl in the 4th position?

Solution: Its 1/90. Probability of a particular child in any position is 1/10. Conditional probability of another child in another position is 1/9.
1 mark for 1/10 part. 1 mark for 1/9 part. 1/2 mark for non-crazy attempt at either. At least 1 mark for any non-crazy attempt all told.

(2 marks) (b). What is the probability that the person in the 3rd position is a boy?

Solution: Its 3/10.
At least 1 mark for any non-crazy attempt all told.

(3) A strange die has the following probabilities:

x	1	2	3	4	5	6
$p(X = x)$	0.10	0.25	0.15	0.05	0.25	0.2

(2 marks) (a). Are the following two events independent?

- A = (X < 4) (i.e., X = 1 or X = 2 or X = 3)
- B = (X is odd) i.e., X = 1 or X = 3 or X = 5)

Solution: $p(A) = 0.5, p(B) = 0.5, p(A \cap B) = 0.25$, so $p(A) * p(B) = p(A \cap B)$ entails the two events are independent.

1 mark for a statement to the effect of $p(A) * p(B) = p(A \cap B)$. 1/2 mark for computing relevant probabilities.

(1 mark) (b). What about for a regular dice?

Solution: But it is not the case for a regular dice since $p(A \cap B) = 0.33$ while $p(A) = 0.5, p(B) = 0.5$.
Needed to compute $p(A \cap B) = 0.33$, $p(A) = 0.5, p(B) = 0.5$. If done so but wrong answer, then 1/2 mark.

(2 marks) (4) In your small town of 1000 people, 600 classify themselves Republican, the remainder are Democrats. But during the election voters may switch sides. In the recent election, 60 Republicans switch and vote Democrat, and 50 Democrats switch and vote Republican. What is the probability that if someone voted Republican they are in fact Democrat?

Solution: Let Dem and Rep be party. Let $VDem$ and $VRep$ be how they vote. $p(Rep) = 0.6$ and $p(Dem) = 0.4$, $p(VDem, Rep) = 0.06$ and $p(VRep, Dem) = 0.05$. Alternatively, you could give the counts, 600, 400, 60, 50, as long as the event is being described properly.

$$p(Dem|VRep) = \frac{p(VRep, Dem)}{p(VRep)} = \frac{50}{600 - 60 + 50} = 0.0847$$

1 mark for stating the probabilities or the numbers for main events. 1 mark for final calc and 1/2 for moderate attempt.

3. Expectation: ..... *Total: 11 marks*

(1) The lifetime in hours of electronic tubes is a random variable have a probability density function

$$f(x) = \frac{1}{C} x e^{-ax}, x \geq 0$$

Furthermore, you are given the following integral:

$$\int_0^\infty x^n e^{-ax} = \frac{1}{a^{n+1} n!}$$

(1 mark) (a). What is the value of C to make the distribution normalise to 1?

Solution: C is found when $n = 1$, so it is $1/a^2$.
1/2 mark for any moderate statement but wrong.

(2 marks) (b). What is the mean of x?

Solution: Use $n = 2$ to get $1/2a^3$ as the integral for x, and divide by $C = 1/a^2$, which is $1/2a$.
1 mark for using $n = 2$ right. 1 mark for any reasonable attempt.

(2) Let $E[X] = 1$ and $E[X^2] = 4$, $E[Y] = -1$ and $E[Y^2] = 3$, and X and Y are independent, then

(2 marks) (a). What is $E[(2 + 4X)^2]$?

Solution:

$$\begin{aligned} E[(2+4X)^2] &= E[4 + 16X + 16X^2] \\ &= 4 + 16\,E[X] + 16\,E[X^2] \\ &= 4 + 16 + 16 * 4 = 84 \end{aligned}$$

1/2 mark for expanding the square, 1/2 mark for linearity, 1/2 mark for substituting.

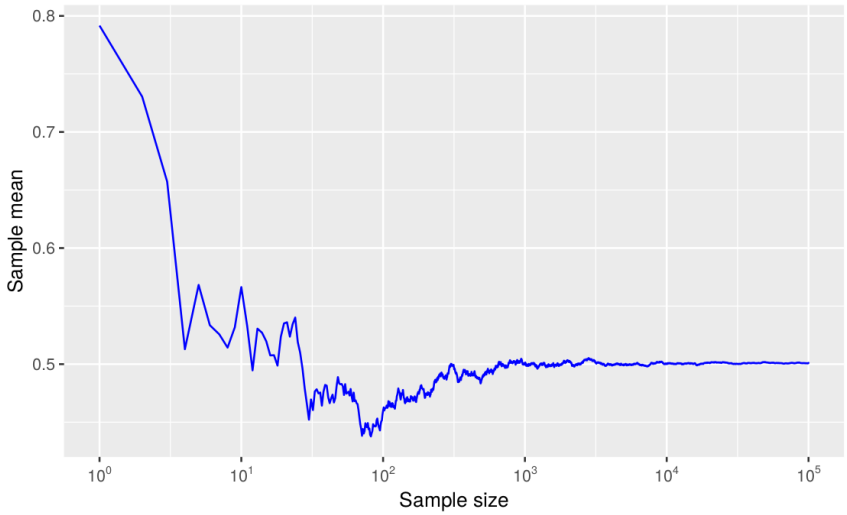
(2 marks) (b). What is $V[(X + 2Y)]$?

Solution: Compute as $E[(X + 2Y)^2] - E[(X + 2Y)]^2$.

$$\begin{aligned} E[(X+2Y)^2] &= E[X^2 + 4XY + 4Y^2] \\ &= E[X^2] + 4E[X]E[Y] + 4E[Y^2] \\ &= 4 + 4 \cdot 1 \cdot -1 + 4 \cdot 3 = 12 \\ E[X+2Y] &= E[X] + 2E[Y] = 1 + 2 \cdot -1 = -1 \\ V[(X+2Y)] &= 12 - 1 = 11 \end{aligned}$$

1/2 mark for expanding variance into expected values, 1/2 mark for expanding the square, 1/2 mark for linearity, 1/2 mark for substituting.

(4 marks) (3) We have X distributed as uniform between $(0, 1)$. We sample $100,000$ values of X and record the mean of the first n values, plotting the sample mean against the sample size (n), producing the curve below.



So, for instance, at $n = 10$ the mean is approximately 0.565, and when $n = 100$ the mean is approximately 0.454.

Describe how the Weak Law of Large Numbers and the Central Limit Theorem affect the look of this curve.

Solution: The WLLN says the average approaches the mean. The CLT says the mean is approximately normally distributed with variance equal sample variance divided by n, so sample standard deviation divided by $\sqrt{n}$.

Thus WLNN means as n gets bigger the plotted values should get closer to 0.5. The CLT has the same effect, but also means the difference from the mean should shrink as per $1/\sqrt{n}$, which means the plotted values should get closer and closer to the mean. Since the cure has n in log scale, the effect of shrinking is accentuated.
2 marks for roughly interpreting WLLN and CLT without considering curve. 1 mark for describing effect of WLNN on curve, and 1 mark for CLT.

4. Distributions: ..... *Total: 16 marks*

(2 marks) (1) You toss 10 fair coins. What is the probability of getting one or less heads?

> **Solution:** The sample space is 2^{10} toss outcomes. 1 outcomes has zero heads. 10 outcomes have exactly 1 head. So probability is $(1+10)\frac{1}{2^{10}} = 0.0107$.
> 1 mark for expressing as zero or exactly 1 head. 1/2 mark for giving full size of sample space.

(2) A fair die is rolled and the value of the top face found.

(2 marks) (a). What is the mean and variance of the value?

> **Solution:** Mean is 3.5. Variance is $\frac{1}{6}(1+4+9+16+25+36) - 3.5^2 = 2.917$.
> 1 mark each.

(2 marks) (b). Hence or otherwise determine an approximate 95% most probable interval for the sum of 100 independent rolls of the dice. State any assumptions you make.

> **Solution:** Use the sum version of the CLT. The sum has mean 350 (100×3.5) and variance 291.7 (100×2.917), which is a std.dev. of 17.08. Need to find $Z_{0.975} = 1.96$. So the interval is $[350 - 1.96 * 17.08, 350 + 1.96 * 17.08] = [316.52, 383.48]$.
> 1 mark for computing variance or std.dev. 1 mark for doing 2-sided and getting 1.96.

(3) Two stores are side by side and have opening hours of 8 hours a day.

(1 mark) (a). The first store sells dresses at \$50 each. Customers arrive at a rate of 4 an hour and half the customers buy a single dress, the rest buy none. What is the distribution of dollar sales made by the first store in a day?

Solution: The number of sales per day is Poisson with rate 16. The dollar sales is 50 times this.
1/2 mark each

(1 mark) (b). The second store sells cigarette packets at \$6 each. Customers arrive at a rate of 20 an hour and all the customers buy a single pack, the rest buy none. What is the distribution of dollar sales made by the second store?

Solution: The number of sales per day is Poisson with rate 160. The dollar sales is 6 times this.
1/2 mark each

(3 marks) (c). Now the Poisson distribution is approximately Gaussian when the rate is larger, and this applies moderately well when the rate is 15 or more. Give Gaussian approximations to the distributions of dollar sales for each of the two stores.

Solution: The sales counts respectively are $N(16, 16)$, and $N(160, 160)$. Now we scale. The first is $N(800, 40000)$, and the second is $N(960, 5760)$.
1 mark each for first two $N()$s. 1 mark for final scaling.

(4) One store has daily sales in dollars of approximately $N(1000, 40000)$ (that is, a variance of 40000), and a second has sales in dollars of approximately $N(800, 4000)$.

(2 marks) (a). What is the probability the first store makes more money than the second?

Solution: Means and variances both add, so the distribution on the difference is $N(1000 - 800, 40000 + 4000)$ which is $N(200, 44000)$. The Z value for \$0 is $\frac{-200}{\sqrt{44000}} = -0.953$. So the probability of greater is approximately 0.83 (Z tables only good 2 decimal places).
1 mark for getting $N(,)$ right. 1/2 mark for getting Z value.

(2 marks) (b). What is the probability that the first store fails to make \$600 in sales? What is it for the second store?

Solution: The Z value for \$600 for the first store is $\frac{-400}{\sqrt{40000}} = -2.0$, and for the second is $\frac{-200}{\sqrt{4000}} = -3.16$. So for the first store this is 2.3% and for the second 0.1%.
1/2 mark for getting each Z right. 1/2 mark for each probability from Z.

(1 mark) (c). What is the probability that one or other store fails to make \$600 in sales?

Solution: It is $1 - (1 - 0.023) * (1 - 0.001) \approx 2.4\%$.
1/2 mark for any reasonable attempt.

5. Inference: ..... *Total: 17 marks*

(3 marks) (1) The Hitchhiker's Estimator for a population was given in the tutorials to be $\hat{\mu} = 42$. Explain in words when this works well, and give its bias and variance.

Solution: The bias is $42 - \mu$. The variance is 0 (since it is a constant). If the true μ happens to be close to 42, then by lucky chance the HE will work well. Otherwise, it works poorly.
1 mark for bias, 1 for variance, 1 for interpretation. 1/2 marks for moderate attempts.

(2) Suppose that a hairdresser with a single store has 5 customers arrive on average every hour while they are open. Assume that customers' arrivals don't depend on time of day or on whether other customers have arrived. We wish to model the arrival of customers in the morning (9am to 12 midday).

(2 marks) (a). Describe a suitable distribution for X the number of customers arriving in the morning, and give its mean and variance.

Solution: Poisson with rate $\lambda = 15$. This is also the mean and variance.
1 mark for Poisson (and 1/2 mark off if wrong lambda), 1 for mean and variance.

(4 marks) (b). The hairdresser opens a new store. In their first 30 days, the total number of customers who arrive in the morning is 420. Using just this information what is the maximum likelihood estimate for the average number of customers arriving in the morning in the new store? Give an approximate 99% confidence interval for the rate of the new store.

Solution: Estimate $\lambda_{ML} = \frac{30}{420} = 14$. Using the CLT, the mean is $N(\lambda_{ML}, \lambda_{ML}/30)$. For 99% two sided confidence interval $Z = 2.58$. Using the CLT, the CI is $14 \pm 2.58\sqrt{\frac{14}{30}}$, which is 14 ± 1.76.
1 mark for λ_{ML}. 1 mark for giving $N(,)$ via CLT with right parameters. 1 mark for giving Z. 1 marking for getting CI.

(4 marks) (c). Test the hypothesis that the new store has a higher rate of customers than the original store. Get a p-value and give your recommendation.

Solution: Now, the new store has a lower rate, so either it is rejected, or perhaps we don't get enough evidence. Have H_0 given by $\lambda > 15$. Using the CLT, then the Z-score for 14 is $\frac{14-15}{\sqrt{15/30}} = -1.414$. This gives a p-value for a one-sided test of approximately 0.08. Thus there is mild evidence that H_0 is rejected. But you could say there is mild evidence against the new store having a higher rate of customers.
1 mark for stating H_0. 1 mark for computing Z. 1 marking for giving p-value or possibly its inverse (1-p). 1 mark for intepretation.

(3) A company self-insures its large fleet of cars against collision. To determine the nature of its repair costs per collision, it has randomly chosen a sample of 16 accidents. Suppose the average repair cost for these is \$4,200 with a sample standard deviation of \$1,200.

(2 marks) (a). Give a 95% two-sided confidence interval for the mean. State your assumptions clearly.

Solution: Assuming the repair costs are Gaussian, the 95% CI should use a Students t with $df = 15$, and the two-sided t value is 2.131. Not quite correct is instead using a Z value of 1.96. Yields $4200 \pm 2.131 * 1200/\sqrt{16} = 4200 \pm 639$ or $4200 \pm 1.96 * 1200/\sqrt{16} = 4200 \pm 588$. Clearly, a Gaussian isn't ideal since the cost cannot be symmetric, and surely it will be skewed left.
1 mark for correct form of CI. 1/2 mark for correct t/Z. 1/2 for assumption discussion.

(2 marks) (b). Give a 95% two-sided confidence interval for the standard deviation. State your assumptions clearly.

Solution: Assumptions as above. Confidence interval for σ^2 is given by $\frac{(n-1)s^2}{\chi^2_{n-1}}$. Upper and lower χ^2_{15} for 95% are 6.262 and 27.488. Thus σ^2 is in $\left(\frac{15*1200^2}{27.488}, \frac{15*1200^2}{6.262}\right)$. Thus σ is in $(886.5, 1857.2)$.
1 mark for giving $\frac{(n-1)s^2}{\chi^2_{n-1}}$. 1/2 mark for computing bounds of χ^2_{15}.

6. Simulation: ..... *Total: 9 marks*

(1) Consider the simple Bayesian network below. Assume this network correctly models the relationships between A, B, C, D and E.

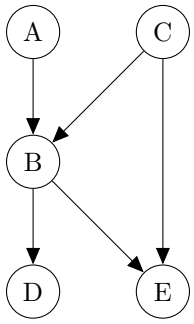


Figure 1: Simple Bayesian network

(2 marks) (a). Write a factorised probability expression for the full joint distribution $p(A, B, C, D, E)$.

Solution:

$$p(A, B, C, D, E) = p(A)p(C)p(B|A, C)p(D|B)p(E|B, C)$$

1 mark for any reasonable attempt. 1/2 mark off for alomst correct.

(1 mark) (b). Which variables if any are independent of E when controlling for B. i.e., assuming B is known.

Solution: A and C.
1/2 mark for getting one right.

(3 marks)

(c). When using a Gibbs sampler to sample from the joint distribution of A, C, D, E when conditioned on $B = b$. What are the four conditional probability distribution functions one needs to be able to calculate? Write out how you would compute them.

Solution:

$$
\begin{aligned}
p(A|B=b,C,D,E) &\propto p(A)p(B|A,C) \\
p(C|B=b,A,D,E) &\propto p(C)p(B|A,C)p(E|B,C) \\
p(D|B=b,A,C,E) &\propto p(D|B) \\
p(E|B=b,A,C,D) &\propto p(E|B,C)
\end{aligned}
$$

1.5 marks if the LHSs are correct. Proportionally allocate remaining 1.5. 2.5 marks total if 2/4 correct.

(2) Suppose we wish to use rejection sampling to sample from the distribution with pdf given by the function:

$$p(x) = |x| \quad \text{for } -1 \leq x \leq 1$$

using a random number generator which generates samples distributed uniformly between -1 and 1.

(2 marks) (a). Write the rejection probability that would be used in the algorithm. What proportion of samples would we reject?

Solution: The short answer is to draw on figure and see the difference. Stating this with picture is full marks. The long answer is to use the formulas. $q(X) = |x|$ and $C = \frac{1/2}{1} = 1/2$, and plug into formulas in Lecture 10. So rejection probability is $1 - \frac{1/2 \cdot |x|}{1/2} = 1 - |x|$. Total rejected samples is $1 - 1/2 = 1/2$. Total rejection is 0.5 of samples.
1 mark for each part. Lose 1/2 if they give acceptance instead of rejection.

(1 mark) (b). If uniform random number generator gave us the value 0.75, with what probability would we accept the sample?

Solution: According to above, reject is $1 - |x| = 1 - 0.75 = 0.25$. So accept is 0.75.
1/2 mark is wrong computation.

7. Regression: ..... *Total: 10 marks*

Below is a table of the wealth (measured in GDP per capita), and inequality (measured with the Gini coefficient for wealth) of a small number of countries:

Country	GDP (in $US)	Wealth Gini
Armenia	3,068	0.684
Australia	27,193	0.622
Benin	1,225	0.713
Brazil	16,397	0.620
Canada	28,731	0.688
Finland	24,416	0.615
Egypt	4,406	0.689
Ireland	27,197	0.581
Japan	25,924	0.547
Venezuela	3,168	0.712

(2 marks) (1) We have a new country Malaysia which has `GDP` of $US 9,422. Suppose we use a 1-nearest neighbour method to estimate the `Gini` for it. Define your distance measure used to compute nearest neighbours and do the computation. What is the `Gini` estimate with this?

Solution: Distance is absolute value of difference in `GDP`. Closest neighbour is Egypt. Estimate is 0.689.
1 mark for giving reasonable distance. 1 for giving Egypt.

(4 marks) (2) Build a simple linear regression model to predict a country's `Gini` given their `GDP`, and complete your solution by giving the linear prediction formula. The key statistics for data in this table are:

$$\begin{aligned} n &= 10 \\ \overline{GDP} &= 16172.5 \\ \overline{GDP^2} &= 388202517 \\ \overline{Gini} &= 0.6471 \\ \overline{Gini^2} &= 0.42175 \\ \overline{GDP * Gini} &= 10010.81 \end{aligned}$$

Solution: The solution for β_1 of

$$\hat{\beta}_1 = \frac{\mathrm{SS}_{XY}}{\mathrm{SS}_{XX}} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \overline{X}^2} = \frac{10010.81 - 0.6471 \cdot 16172.5}{388202517 - 16172.5^2} = -0.00000359$$

and the solution for β_0 of

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 0.6471 - -0.00000359 \cdot 16172.5 = 0.705$$

2 marks for giving correct formula. 1 mark each for correct values.

(4 marks) (3) Theory tells us that the estimate $\hat{\beta}_1$, the coefficient given in the formula sheet, has a sampling distribution of

$$\frac{1}{\sqrt{\frac{RSS}{n(n-2)}\frac{1}{\overline{GDP^2}-\overline{GDP}^2}}}(\hat{\beta}_1-\beta_1) \sim \texttt{Student-t}(n-2)$$

$$\text{for } RSS = n\left(\overline{Gini^2}-\overline{Gini}^2-\left(\overline{GDP^2}-\overline{GDP}^2\right)\hat{\beta}_1^2\right)$$

Using a null hypothesis that $\beta_1 = 0$, compute a p-value using the LHS of the sampling distribution.

Solution: First,

$$RSS = 10 \cdot (0.42175 - 0.6471^2 - (388202517 - 16172.5^2)0.00000359^2) = 0.0138$$

Then the standard error (part of factor on the LHS) is

$$\sqrt{0.0138/80/(388202517 - 16172.5^2)} = \sqrt{1.362 \cdot 10^{-12}} = 1.167 \cdot 10^{-6}$$

β_1 is divided by this to give -3.076, and it is Student t with 8 degrees of freedom. We have a 2-sided case, so we need double the probability on the t-tables. The t-tables has $t_{0.01} = 2.896$ and $t_{0.005} = 3.250$. A rough guess to 2 decimal places is we are at $t_{0.007}$ so doubling we get a p-value of 0.014 ± 0.004.

Computing RSS is 1 mark. Computing standard error $1.167 \cdot 10^6$ is 1 mark. Computing t statistic 1 mark. Arguing from full RHS statistic to a p-value is 1 mark. Ignore 1 or 2 sided.

8. Modelling: ..... *Total: 21 marks*

(1) The internet company MemeBuzz wishes to try to predict whether a meme will go viral (a binary variable) based on some easily measured characteristics. They use variables:

Viral - the Boolean target variable.

Tone - a 4-valued categorical input variable indicating whether the meme is happy, sad, angry, or neutral.

Format - a 3-valued categorical input variable indicating whether the meme is a gif, an image or plain text.

Slang - a Boolean input variable indicating whether any slang is used in the meme.

We are considering developing a classification algorithm to predict the target `Viral` given the inputs `Tone, Format, Slang`.

(3 marks) (a). Write out the formula for a Naïve Bayes classifier for case where `Viral=F`.

Solution:

$$p(Viral{=}F, Tone, Format, Slang) =$$
$$p(Viral{=}F)p(Tone|Viral{=}F)p(Format|Viral{=}F)p(Slang|Viral{=}F)$$
$$p(Viral{=}T, Tone, Format, Slang) =$$
$$p(Viral{=}T)p(Tone|Viral{=}T)p(Format|Viral{=}T)p(Slang|Viral{=}T)$$
$$p(Viral{=}F|Tone, Format, Slang) =$$
$$\frac{p(Viral{=}F, Tone, Format, Slang)}{p(Viral{=}F, Tone, Format, Slang) + p(Viral{=}T, Tone, Format, Slang)}$$

Partial formula, partial marks. If unnormalised, 2.5/3.

(3 marks) (b). Write out the formula for a logistic regression classifier for the case where `Viral=F`.

Solution: We have linear parameters $\beta_0, \beta_{Tone}, \beta_{Format}$ and β_{Slang}, and we have to convert the discrete features into indicators which take values 0 or 1. Call the Tone outcomes H,S,A or N for short. β_{Tone} has values for different outcomes of Tone, let us call them $\beta_{Tone,H}$ for Happy, $\beta_{Tone,S}$ for Sad, and $\beta_{Tone,A}$ for Angry. $\beta_{Tone,N}$ for neutral is optional. Likewise for Format, with outcomes G, I or P, and β_{Format} has different outcomes of Format. Slang has outcomes T or F.

$$\mu = \beta_0 + \beta_{Tone,H}1_{Tone=H} + \beta_{Tone,S}1_{Tone=S} + \beta_{Tone,A}1_{Tone=A} + \beta_{Format,G}1_{Format=G} + \beta_{Format,I}1_{Format=I} + \beta_{Slang}1_{Slang=T}$$

$$p(Viral = F|Tone, Format, Slang) = \frac{1}{1+e^{-\mu}}$$

Whether logistic is e^{μ} or $e^{-\mu}$ doesn't matter. 2 marks for expressing μ. Doesn't matter if they had 2 or 3 outcomes for Format in μ or 3 or 4 outcomes of Tone in μ. 1 mark for logistic.

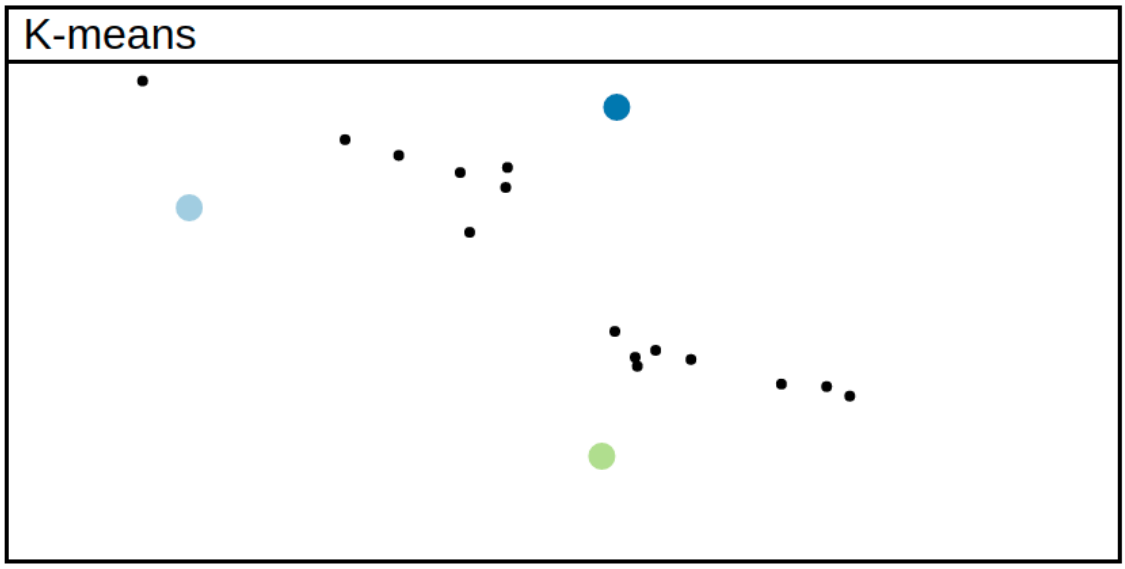
(3 marks) (c). Discuss the difference between the estimation/learning algorithms for the Naïve Bayes classifier versus the logistic regression classifier.

Solution:

- complexity generally: logistic does function maximisation so requires quite a few cycles with likelihood calculations, so is generally a lot slower than NB
- scaling with p number of features and n number of data: NB is np whereas logistic is Lnp for L cycles of gradient descent or maybe Lnp^2 for more complex fitting routines
- estimation style: logistic regression does point estimation of parameters; NB computes actual probabilities
- resultant models: both are linear

1 mark for each useful/valid point.

(2) The figure below shows some points in 2-D that we wish to model with the K-means algorithm. The 3 larger round dots are the proposed initial cluster centers. The distance measure used in this case is the geometric distance between the two points.



(3 marks) (a). Draw on the figure the results of the first step of the k-means algorithm: assigning points to cluster centers. Connect each point to its appropriate cluster center with a straight line.

Solution: Connect each point with a straight line to its closest.
1 point is unclear which one it is closer too, so can allow either. 2 marks for reasonable attempt.

(3 marks) (b). Draw on the figure the results of the second step of the k-means algorithm, recomputing the cluster centers.

Solution: The new centers are the center point for each cluster.
2 marks for reasonable attempt.

(3 marks) (c). Give an intuitive argument as to why the k-means algorithm should converge in a finite number of steps.

Solution: At each of the two steps (reassign points to clusters, recenter clusters), the total distance of all points to their assigned cluster center is decreasing. Note, completely different explanations could be given, too.
Assess different solutions for reasonableness. 2 marks for reasonable attempt.

(3 marks) (3) A toy dataset is described as follows:

- $n = 10$ data points
- binary target y
- two binary predictors, x_1 and x_2

Laying out the data in columns, where each row is a variable gives:

y	1	1	0	1	0	0	0	1	1	0
x_1	0	0	0	0	0	1	1	1	1	1
x_2	1	1	0	1	1	0	0	0	1	0

We want to build a decision tree to predict y from x_1 and x_2. Should we test x_1 or x_2 at the root of the tree, or maybe no tests at all. Give a metric for choosing which tests to do, evaluate it on the data, and select which test (or no test) is best.

Solution: The only metric we have done is information gain $I(y) - I(y|x_1)$ versus $I(y) - I(y|x_2)$, and we want larger, or conditional information $I(y|x_1)$ versus $I(y|x_2)$, and we want the smaller. Below we use conditional information. Compute this as

$$I(y|x_1) = p(x_1 = T)I(y|x_1 = T) + p(x_1 = F)I(y|x_1 = F)$$

and likewise for x_2. Now $p(x_1 = T) = 0.5$ and $p(x_2 = T) = 0.5$.

$$\begin{aligned} I(y|x_1 = T) &= 2/5 \log_2 5/2 + 3/5 \log_2 5/3 = 0.971 \\ I(y|x_1 = F) &= 3/5 \log_2 5/3 + 2/5 \log_2 5/2 = 0.971 \\ I(y|x_2 = T) &= 4/5 \log_2 5/4 + 1/5 \log_2 5/1 = 0.723 \\ I(y|x_2 = F) &= 1/5 \log_2 5/1 + 4/5 \log_2 5/4 = 0.723 \end{aligned}$$

So $I(y|x_1) = 0.971$ and $I(y|x_2) = 0.971$. Also note $I(y) = 1$ so no test is worse. x_2 is the best test.
1 mark for give use of one of the $I(|)$ forms as test. 1 mark for stating x_2 is best, regardless of reason. 1 mark for evaluating $I(|)$.