Question 1. Suppose we have been given training data consisting of the number (L) of hours of lectures attended by a student, the number of hours (S) spent on self-study and whether the student passed/failed the final exam. Using this data we train a logistic regression model, which has parameters $\theta_0 = -8.75$, $\theta_1 = 0.25$ and $\theta_2 = 0.1$.

1. Estimate the probability for a student with $L=35$ and $S=20$ to pass the exam

2. Estimate how many hours $S$ of self-study a student who could attend only $L=25$ hours of classes needs to study to have that same probability to pass the exam.

Model Solution 1. 1. The probabilistic interpretation of Logistic regression gives $P(Y = y | \theta, x) = \frac{1}{1 + e^{-y \theta^T x}}$ for $y = 1$ and $y = -1$. In this example the output $Y$ is whether the student passes (say we label this +1) or fails (say we label this −1), the feature vector $x = [1, L, S]^T$ (we include a dummy first feature which is constant) and parameter vector $\theta = [-8.75, 0.25, 0.1]^T$. When $L = 35$, $S = 20$ then $x = [1, 35, 20]^T$, $\theta^T x = -8.75 \times 1 + 0.25 \times 35 + 0.1 \times 20 = 2$ and so $\text{Prob passed} = P(Y = +1) = \frac{1}{1+e^{-2}} = 0.88$.

2. To get the same value for $P(Y = +1)$ we need to keep $\theta^T x = 2$. But now $L = 25$ so we need to find $S$ such that $-8.75 \times 1 + 0.25 \times 25 + 0.1 \times S = 2$. Subtracting $-8.75 \times 1 + 0.25 \times 25$ from both sides, we need $0.1 \times S = 2 - (-8.75 \times 1 + 0.25 \times 25) = 4.5$. Dividing both sides by 0.1 we then have that $S = 45/0.1 = 45$.

Question 2. Consider the following two-dimensional classification task, where 'x' indicates class $y = 1$ points and 'o' class $y = -1$ points.

1. Describe a logistic regression model and explain how it can be fitted to this training data

2. Will a logistic regression model fitted to this training data correctly predict the labels for all these training points, or will it make mistakes for some points? Explain your answer.

3. Now suppose that we modify the cost function used to train the model so that includes a penalty/regularisation term $-\lambda \theta_j^2$ i.e. only one parameter $\theta_j$ is penalised. We select a very large value for $\lambda$ (think $\lambda = +\infty$). Given the training data above, explain how the training error will change when $j = 1, 2$?
Model Solution 2. 1. In a logistic regression model the output is predicted using hypothesis/model \( y = \text{sign}(\theta^T x) \) where \( \theta \) are unknown parameters and \( x \) is the vector of inputs/features. In this example there are two features \( x_1 \) and \( x_2 \). Usually we would also add an extra constant feature so that for output \( y \) the feature vector used would be \([1, x_1, x_2]^T\) and our prediction is \( \text{sign}(\theta_0 + \theta_1 x_1 + \theta_2 x_2) \). Given training data \((x^{(i)}, y^{(i)}), i = 1, \ldots, m\) we select the model parameters \( \theta \) by minimising the cost function \( J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \log(1 + e^{-y^{(i)} \theta^T x^{(i)}}) \).

2. Yes, since the data is linearly separable i.e. we can draw a straight line that cleanly separates the two classes of data. The following sktech an example separating line added to above figure:

3. The cost function is now \( J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \log(1 + e^{-y^{(i)} \theta^T x^{(i)}}) - \lambda \theta_j^2 \). When \( j = 1 \) and \( \lambda \) is very large the parameter vector which minimises the cost function will have \( \theta_1 = 0 \) since the penalty term strongly penalises non-zero values of \( \theta_1 \). This means that the separating plane will be forced to be a vertical line (since it can only vary with \( x_2 \)). Since a vertical line can still cleanly separate the data, the classification accuracy on the training data will still be good. When \( j = 2 \) the minimising parameter vector will have \( \theta_2 = 0 \) and so the separating plane will be forced to be a horizontal line (since it can only vary with \( x_1 \)). The data is not separable by a horizontal line, and so the classification accuracy on the training data will be worse, e.g.

**Question 3.** Consider the following two-dimensional classification task, where ‘x’ indicates class \( y = 1 \) points and ‘o’ class \( y = -1 \) points.
1. Is the data shown above linearly separable? Explain your answer.

2. Suppose we want to use a logistic regression model with this training data. Discuss how we might select the features/inputs used so as to correctly label the training data.

**Model Solution 3.**

1. No, it’s not linearly separable since we cannot draw a straight line that cleanly separates the two classes of data.

2. It’s looks as if there is a curve (rather than a straight line) that cleanly separates this data, e.g.

   ![Graph showing data points and a curve]

   Data is not separable by a straight line, but can be separated by a curve

So we might for example consider using $x_1^2$ and $x_2^2$ as features instead of (or as well as) $x_1$ and $x_2$.

**Question 4.** Suppose we are given patient data consisting of $m$ pairs $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots (x^{(m)}, y^{(m)})$ where $x^{(i)}$ is the measured blood sugar level and $y^{(i)}$ equals 1 if the patient has been diagnosed diabetic and -1 otherwise.

1. With reference to Bayes Rule explain what is meant by the likelihood, prior and posterior. Explain how the maximum likelihood estimate of a model parameter differs from the maximum a posteriori (MAP) estimate.

   Suppose we decide to use a logistic regression model for this patient data.

   2. Give the expression for $p(x^{(i)}|y^{(i)})$ used in a logistic regression model

   3. Now give the expression for the likelihood of the observed data $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})$

   4. Discuss why a logistic regression model requires linear separability.

**Model Solution 4.**

1. For random variables $Y$ and $\theta$ Bayes Rule is $P(\theta|Y) = P(Y|\theta)P(\theta)/P(Y)$. Here $P(Y|\theta)$ is the likelihood, $P(\theta)$ the prior and $P(\theta|Y)$ the posterior. In maximum likelihood estimator the parameter values $\theta$ in a model are selected to maximise the likelihood $P(Y|\theta)$. A MAP estimator selects the parameter values $\theta$ that maximise the posterior $P(\theta|Y)$. 

3
2. The probabilistic interpretation of Logistic regression gives
\[ P(y(i)|x(i), \theta) = \frac{1}{1 + e^{-y(i) \theta^T x(i)}} \]
for \( y = 1 \) and \( y = -1 \).

3. The likelihood is
\[ \prod_{i=1}^{m} P(y(i)|x(i), \theta) = \frac{1}{1 + e^{-y(1) \theta^T x(1)}} \times \frac{1}{1 + e^{-y(2) \theta^T x(2)}} \times \cdots \times \frac{1}{1 + e^{-y(m) \theta^T x(m)}} \]

4. In a logistic regression model the prediction is \( \text{sign}(\theta^T x) \). Another way to see this is that the prediction given feature vector \( x \) is the value of output \( y \) that makes the probability \( \frac{1}{1 + e^{-y \theta^T x}} \) highest – when \( \theta^T x > 0 \) then this is \( y = +1 \) and when \( \theta^T x < 0 \) then its \( y = -1 \). Linear separability is when a straight line cleanly separates the features corresponding to each class of output i.e. on one side of the line all outputs belong to one class. <insert a sketch illustrating this>. \( \theta^T x \) defines a line and prediction \( \text{sign}(\theta^T x) \) assigns all outputs on one side of the line to the same class. So this model needs linear separability in order to produce accurate predictions.

**Question 5.** Suppose you are given a data set of \( m \) training pairs \((x(i), y(i)), (x(2), y(2)), \ldots (x(m), y(m))\).

We want to fit the following model to the data:
\[ y = f(x) + n \]
where \( f(x) = \sum_{j=1}^{n} \theta_j \sin(jx) \) and \( n \) is Gaussian noise with mean zero and variance \( \sigma^2 \).

1. Give an expression for \( p(y|x, \theta) \)
2. Now give an expression for the likelihood and log-likelihood of the training data
3. Suppose now that we assume a Gaussian prior with mean zero and variance 1 on the parameters \( \theta \). Give an expression for the posterior probability given observation of the training data.

**Model Solution 5.**

1. \( p(y|x, \theta) \propto e^{-\frac{(y-f(x))^2}{\sigma^2}} \)
2. The likelihood is proportional to
\[ \prod_{i=1}^{m} p(y(i)|x(i), \theta) = e^{-\frac{(y(1)-f(x(1)))^2}{\sigma^2}} \times e^{-\frac{(y(2)-f(x(2)))^2}{\sigma^2}} \times \cdots \times e^{-\frac{(y(m)-f(x(m)))^2}{\sigma^2}} \]
3. The log-likelihood is proportional to \( \log \prod_{i=1}^{m} p(y(i)|x(i), \theta) = -\sum_{i=1}^{m} \frac{(y(i)-f(x(i)))^2}{\sigma^2} \)
4. The log-posterior is proportional to
\[ \log \prod_{i=1}^{m} p(y(i)|x(i), \theta) + \log p(\theta) = -\frac{1}{\sigma^2} \sum_{i=1}^{m} (y(i) - f(x(i)))^2 - \sum_{j=1}^{n} \theta_j^2 \]