W16 Problem Sheet Explanation: Classification

Question 1

Find the best model on the following test set, using the log-likelihood as the objective.

x_0	x_1	y
-2.1	-3.2	0
-3.4	-1.2	0
1.2	3.6	1
-0.1	0.8	1

The models are, where l is the logits:

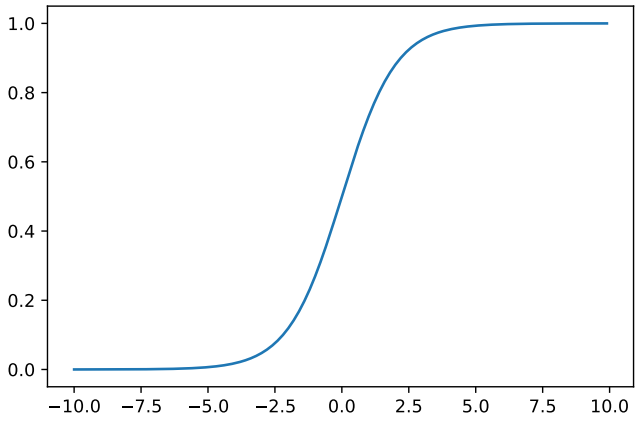
1. $l = x_0 + x_1$
2. $l = 4x_0 + 0.01x_1$
3. $l = 10x_0 + 10x_1$

In linear regression, we would try fit weights to an equation similar to the logits. We would sum up each of the x_i inputs, each multiplied by a coefficient. Our output in linear regression is y, which can be any value between $-\infty$ and ∞.

This is how we will view the logits equation, however we need to transform it so that, instead of getting any number out, we get either a 0 or 1. This is where we will use the **sigmoid** function (also known as logistic function).

The sigmoid function looks like this:

$$f(x) = \frac{1}{1 + e^{-x}}$$



Now, what this function essentially does it allows us to input any real number and output a number in the range $0-1$. So, if we input our l values from the logits functions (which will be any real number), the sigmoid function transforms it into a value between 0 and 1.

As you can see from the shape of the graph, most input values will give an input of either 0 or 1 - this is why we can utilise this within Bernoulli distributions.

However, this function cannot represent the two distinct classes 0 and 1, what it really **represents is the probability of getting a success** (a 1), so $P(Y=1|X)$.

If we put our values into each of the logit equations we get the following:

$$
\begin{aligned}
l_1 &= x_0 + x_1 && (1)\\
&= [-2.1-3.2, -3.4-1.2, 1.2+3.6, -0.1+0.8] && (2)\\
&= [-5.3, -4.6, 4.8, 0.7] && (3)\\
l_2 &= [-0.872, -1.372, 0.516, -0.032] && (4)\\
l_3 &= [-53.0, -46.0, 48.0, 7.0] && (5)
\end{aligned}
$$

Now, looking at the sigmoid graph you can roughly estimate what the output values are going to be for each value.

For l_1 we first have two negative values, -5.3 and -4.6, which will give values close to 0. This intuitively hints that perhaps l_1 could be the model as this means that the probability that these points belong to class 1 is close to 0, which is correct as we know from the original data that they come from class 0. The next point is 4.8 which will imply that that data point is in class 1, however the final point is 0.7, which gives a medium value so we are unsure what class this belongs to.

Doing the same strategy for the other logit equations we can make a guess that perhaps l_3 is the best model as it has the lowest negative numbers for the class 0 data points and the highest positive numbers for the class 1 data points.

Now, let's do this a bit more formally. If we want to find the best model out of the 3 logits, we want to the model that most has most likely created our data. Hence, we will look at the probability that each data point was classified into its correct class for each model.

Now, as explained above we already have the probability that a data point has come belongs to data class 1, by putting our logits numbers into the sigmoid function:

$$P(Y = 1|l) = \frac{1}{1 + e^{-l}}$$

Now, as we only have two possibilities for the classes, the probability of a data point being in class 0 and the probability of it being in class 1 should add up to equal 1. So, all we need to do to get the other probability is:

$$\begin{aligned} P(Y = 0|l) &= 1 - \frac{1}{1 + e^{-l}} && (6)\\ &= \frac{1 + e^{-l}}{1 + e^{-l}} - \frac{1}{1 + e^{-l}} && (7)\\ &= \frac{e^{-l}}{1 + e^{-l}} && (8)\\ &= \frac{e^{l}}{e^{l}} \cdot \frac{e^{-l}}{1 + e^{-l}} && (9)\\ &= \frac{1}{1 + e^{l}} && (10) \end{aligned}$$

Often we will work with the log probabilities instead, which is what the question has stated us to do, so let's work these out next.

$$\begin{aligned} \log P(Y = 1|l) &= \log\left(\frac{1}{1 + e^{-l}}\right) && (11)\\ &= \log 1 - \log(1 + e^{-l}) && (12)\\ &= -\log(1 + e^{-l}) && (13)\\ \log P(Y = 0|l) &= \log\left(\frac{1}{1 + e^{l}}\right) && (14)\\ &= \log 1 - \log(1 + e^{l}) && (15)\\ &= -\log(1 + e^{l}) && (16) \end{aligned}$$

Now, we just need to put our values into these expressions.

Before when working with the maximum likelihood estimate, we wanted to find the model that gave us the highest probability, and we took the product for each data point. When we worked with the log-likelihood, we took the sum of each data point - so we will do the same here.

So, we want to find the l that the data points most realistically came from, so the summation of the probability that each data point belongs to its labelled class.

First, let's look at $l = x_0 + x_1$ where we had the values $l = [5.3, 4.6, 4.8, 0.7]$:

$$\log P(Y|l) = P(y_1 = 0|l_1) + P(y_2 = 0|l_2) + P(y_3 = 1|l_3) + P(y_4 = 1|l_4) \tag{17}$$
$$= -\log(1 + e^{l_1}) - \log(1 + e^{l_2}) - \log(1 + e^{-l_3}) - \log(1 + e^{-l_4}) \tag{18}$$
$$= -\log(1 + e^{-5.3}) - \log(1 + e^{-4.6}) - \log(1 + e^{-4.8}) - \log(1 + e^{-0.7}) \tag{19}$$
$$= -0.42636 \tag{20}$$

Let's do the same for $l = 4x_0 + 0.01x_1$ where we had the values $l = [0.872, 1.372, 0.516, 0.032]$:

$$\log P(Y|l) = -\log(1 + e^{-0.872}) - \log(1 + e^{-1.372}) - \log(1 + e^{-0.516}) - \log(1 + e^{0.032}) \tag{21}$$
$$= -1.7527 \tag{22}$$

And for $l = 10x_0 + 10x_1$ where we had the values $l = [53.0, 46.0, 48.0, 7.0]$:

$$\log P(Y|l) = -\log(1 + e^{-53}) - \log(1 + e^{-46}) - \log(1 + e^{-0.48}) - \log(1 + e^{-7}) \tag{23}$$
$$= -0.00091147 \tag{24}$$

Before with MLE, the next step would be to find the l which maximises this. As we already have data and we already have numbers there is no need to work with the equation and find the derivative and set equal to 0 - we can simply observe to see which has the highest probability. **From this we can see that the third model has the highest, which is what we predicted earlier anyway**.

Question 2

Using the models and test-set given in Q1, use regularised maximum likelihood, with the objective:

$$\mathcal{L} = \log P(y|\boldsymbol{w}, \boldsymbol{x}) - \frac{1}{2}\boldsymbol{w}^2$$

Compute the regulariser for each model:

$$reg = -\frac{1}{2}\boldsymbol{w}^2$$

And then compute the regularised ML objective as the sum of the regulariser and log-likelihood.

So, as mentioned in the first question, we view the logits vector as the weights. So for each model we have:

$$l = x_0 + x_1 \implies \boldsymbol{w} = \begin{pmatrix} 1 & 1 \end{pmatrix} \tag{25}$$
$$l = 4x_0 + 0.01x_1 \implies \boldsymbol{w} = \begin{pmatrix} 4 & 0.01 \end{pmatrix} \tag{26}$$
$$l = 10x_0 + 10x_1 \implies \boldsymbol{w} = \begin{pmatrix} 10 & 10 \end{pmatrix} \tag{27}$$

So, each of the regularisers are:

$$reg_1 = -\frac{1}{2}\begin{pmatrix}1 & 1\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix} \tag{28}$$
$$= -\frac{1}{2}\cdot(1+1) \tag{29}$$
$$= -1 \tag{30}$$
$$reg_2 = -\frac{1}{2}\begin{pmatrix}4 & 0.01\end{pmatrix}\begin{pmatrix}4\\0.01\end{pmatrix} \tag{31}$$
$$= -\frac{1}{2}\cdot(16+0.0001) \tag{32}$$
$$= -8.00005 \tag{33}$$
$$reg_3 = -\frac{1}{2}\begin{pmatrix}10 & 10\end{pmatrix}\begin{pmatrix}10\\10\end{pmatrix} \tag{34}$$
$$= -\frac{1}{2}\cdot(100+100) \tag{35}$$
$$= -100 \tag{36}$$

Now, to compute the whole expression, the regularised log-likelihood, all we need to do is add our values for our regularisers to the values we computer in question 1, so:

$$\mathcal{L}_1 = 0.42636 - 1 \tag{37}$$
$$= -1.42636 \tag{38}$$
$$\mathcal{L}_2 = 1.7527 - 8.00005 \tag{39}$$
$$= -9.75275 \tag{40}$$
$$\mathcal{L}_3 = 0.00091147 - 100 \tag{41}$$
$$= -100.00091147 \tag{42}$$

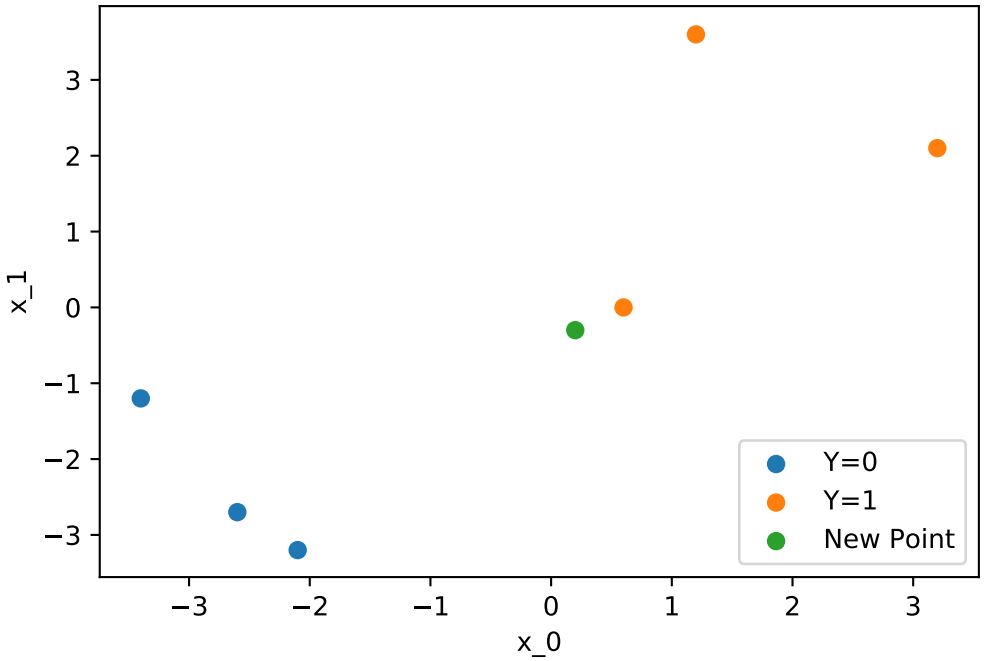
Question 4

Find the predicted class-label for the test point $(0.2, -0.3)$, with $K = 3$ using the standard Euclidean distance.

x_0	x_1	y
-2.1	-3.2	0
-3.4	-1.2	0
-2.6	-2.7	0
3.2	2.1	1
1.2	3.6	1
0.6	0	1

K-nearest neighbours is a pretty self-explanatory classification method. We look

at the point we want to predict, look at its closest neighbours and make our class prediction based on its neighbours.



The plot above shows all the data points given, with class 0 in blue, class 1 in orange and the new data point (the one which we need to predict) has been placed in green. From first observations, we might guess that the point belongs to class 1 as it has an orange point sitting very close to it, however we do not know.

Looking at its neighbours, it only has one close one so it is hard to tell whether its closest 3 neighbours are going to be class 0 or 1. So, what we will do first is **calculate the distance between the new point and all other points** to see which are the closest. We will just use the Euclidean distance which is defined as:

$$D(x, y) = \|x - y\| = \sqrt{\sum_i (x_i - y_i)^2}$$

We will just observe the squared distances (as we are just trying to find the minimum value not specific value), so do not need to worry about the square root. So, for each point we get the following distances:

x_0	x_1	y	Distance
-2.1	-3.2	0	13.7
-3.4	-1.2	0	13.77
-2.6	-2.7	0	13.6
3.2	2.1	1	14.76
1.2	3.6	1	16.21
0.6	0	1	0.73

The closest three points are the first, third and sixth. The first and third have class 0 and the third has class 1, so therefore **we predict that our new point is of class 0**.

Question 5

This question wants us to perform **weighted nearest neighbour** on the data provided in question 4.

The method we used above (just normal k-nearest neighbours) has two problems:

1. The value of k majorly affects the outcome even if there is just a slight adjustment
2. How to decide the outcome - normally you just take the majority vote, however this does not seem very robust

Weighted nearest neighbour solves both these problems by:

1. Gives a larger weight to those points closest and a smaller weight to points further away. If k is increased, then this shouldn't matter too much as the points will be further away and have a small weight so won't affect the overall outcome as much. The weights are allocated using a **kernel function**, known in the notes as $k(\boldsymbol{x}_\lambda, \boldsymbol{x}_{\lambda'})$.
2. You take the weighted average to determine the outcome. We will get one number for each class - the class with the **highest average** is the outcome.

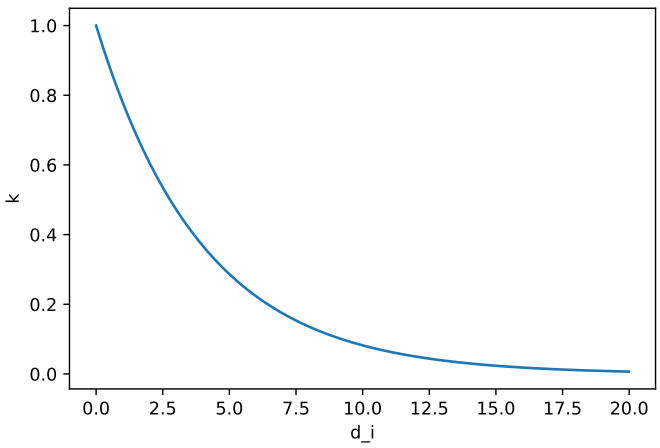
So, first, let's use the kernel function to get the weights. If we denote our previous distances by d_i, then our kernel function is:

$$k(\boldsymbol{x}_\lambda, \boldsymbol{x}_{\lambda'}) = e^{-\frac{d_i}{2b}}$$

In this function, b is just parameter, and the question uses $b = 2$, so we have:

$$k(\boldsymbol{x}_\lambda, \boldsymbol{x}_{\lambda'}) = e^{-\frac{d_i}{4}}$$

What is important about this function is the negative in the power - this means that our exponential function is decreasing. The function can be seen below, notice how the smaller the distance, the higher the weight.



Putting all our previous distances in, we get our weights:

x_0	x_1	y	Distance	Weights
-2.1	-3.2	0	13.7	0.0325
-3.4	-1.2	0	13.77	0.0320
-2.6	-2.7	0	13.6	0.0334
3.2	2.1	1	14.76	0.0250
1.2	3.6	1	16.21	0.0174
0.6	0	1	0.73	0.8332

Now, the next step is to get our weighted average using the following equation (given in notes):

$$p_y(\boldsymbol{x}) = \frac{\sum_\lambda k(\boldsymbol{x}, \boldsymbol{x}_\lambda)\delta_{y,y_\lambda}}{\sum_\lambda k(\boldsymbol{x}, \boldsymbol{x}_\lambda)}$$

Now, $\boldsymbol{x}$ is the point in which we are trying to get and $\boldsymbol{x}_\lambda$ represents all the other points in our dataset. We have already worked out the distances between this new point and the data points and we have also worked out the weights.

The other term in the function, δ_{y,y_λ}, is the delta kronecker function. This outputs a 1 if $y = y_\lambda$ and a 0 otherwise. In other words, it outputs a 1 when the class of the λ^{th} data point is the same as the class weighted average that we are trying to calculate.

As explained before, we need to get a weight average for each class, so let's start with $y = 0$. I will use the notation k_i to denote $k(\boldsymbol{x}, \boldsymbol{x}_i)$:

$$\begin{aligned} p_0(\boldsymbol{x}) &= \frac{\sum_\lambda k(\boldsymbol{x}, \boldsymbol{x}_\lambda)\delta_{0,y_\lambda}}{\sum_\lambda k(\boldsymbol{x}, \boldsymbol{x}_\lambda)} \\ &= \frac{(k_1 \cdot 1) + (k_2 \cdot 1) + (k_3 \cdot 1) + (k_4 \cdot 0) + (k_5 \cdot 0) + (k_6 \cdot 0)}{k_1 + k_2 + k_3 + k_4 + k_5 + k_6} \\ &= \frac{0.0325 + 0.0320 + 0.0334}{0.0325 + 0.0320 + 0.0334 + 0.0250.0174 + 0.8332} \\ &= \frac{0.0979}{0.9735} \end{aligned}$$

Now, let's do the same for class 1:

$$\begin{aligned} p_1(\boldsymbol{x}) &= \frac{\sum_\lambda k(\boldsymbol{x}, \boldsymbol{x}_\lambda)\delta_{1,y_\lambda}}{\sum_\lambda k(\boldsymbol{x}, \boldsymbol{x}_\lambda)} \\ &= \frac{(k_1 \cdot 0) + (k_2 \cdot 0) + (k_3 \cdot 0) + (k_4 \cdot 1) + (k_5 \cdot 1) + (k_6 \cdot 1)}{k_1 + k_2 + k_3 + k_4 + k_5 + k_6} \\ &= \frac{0.0250 + 0.0174 + 0.8332}{0.0325 + 0.0320 + 0.0334 + 0.0250.0174 + 0.8332} \\ &= \frac{0.8755}{0.9735} \end{aligned}$$

The final step is to compare the values - in weighted nearest neighbour we take the class with the **highest weighted average**, which is class 1.

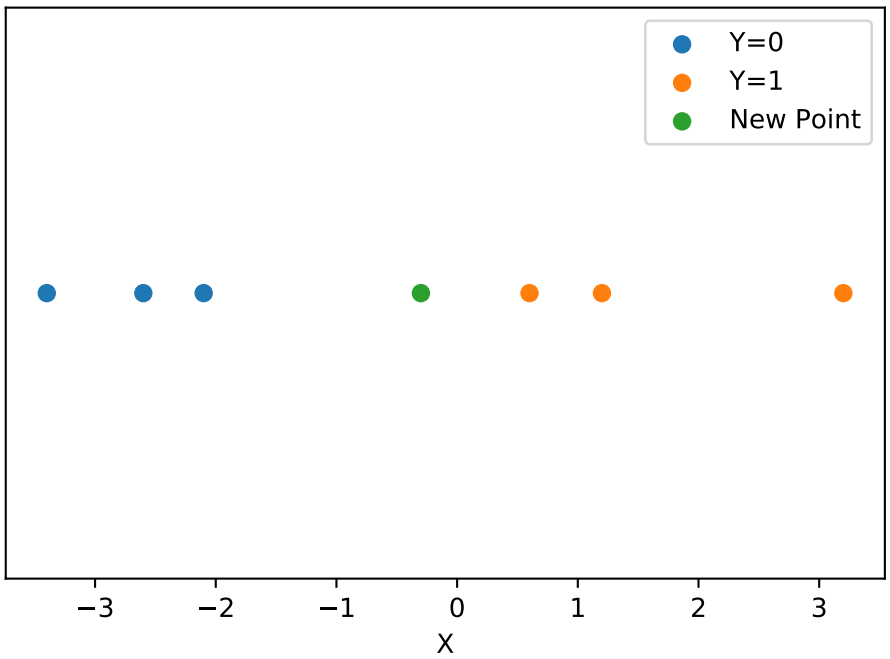
Question 6

Now, let's look at Bayesian inference. In the notes it is explained that Bayesian inference is a step-up from nearest centroid classification.

We have been given the following data:

X	Y
-2.1	0
-3.4	0
-2.6	0
3.2	1
1.2	1
0.6	1

And the data point we are trying to predict the class for is $X = -0.3$, let's call this the 'new point'. We can visualise this data as:



What we will try do now is to fit two separate Gaussian distributions over our two sections of data (one for each class).

Now, how do we normally try to fit a normal distribution for 1-dimensional data? We can just find the mean and variation, using these to get our parameters for our

distributions.

Let's calculate the mean, $E_y(X)$ for each class y:

$$\begin{aligned} E_0(X) &= \frac{-2.1 - 3.4 - 2.6}{3} \\ &= -2.7 \\ E_1(X) &= \frac{3.2 + 1.2 + 0.6}{3} \\ &= \frac{5}{3} \\ &\approx 1.67 \end{aligned}$$

Next, let's calculate the variance, which can be defined as:

$$Var_y(X) = E_y(X^2) - [E_y(X)]^2$$

So, to calculate this we need to calculate $E_y(X^2)$ for each class:

$$\begin{aligned} E_0(X) &= \frac{-2.1^2 - 3.4^2 - 2.6^2}{3} \\ &= 7.5767 \\ E_1(X) &= \frac{3.2^3 + 1.2^3 + 0.6^3}{3} \\ &= 4.0133 \end{aligned}$$

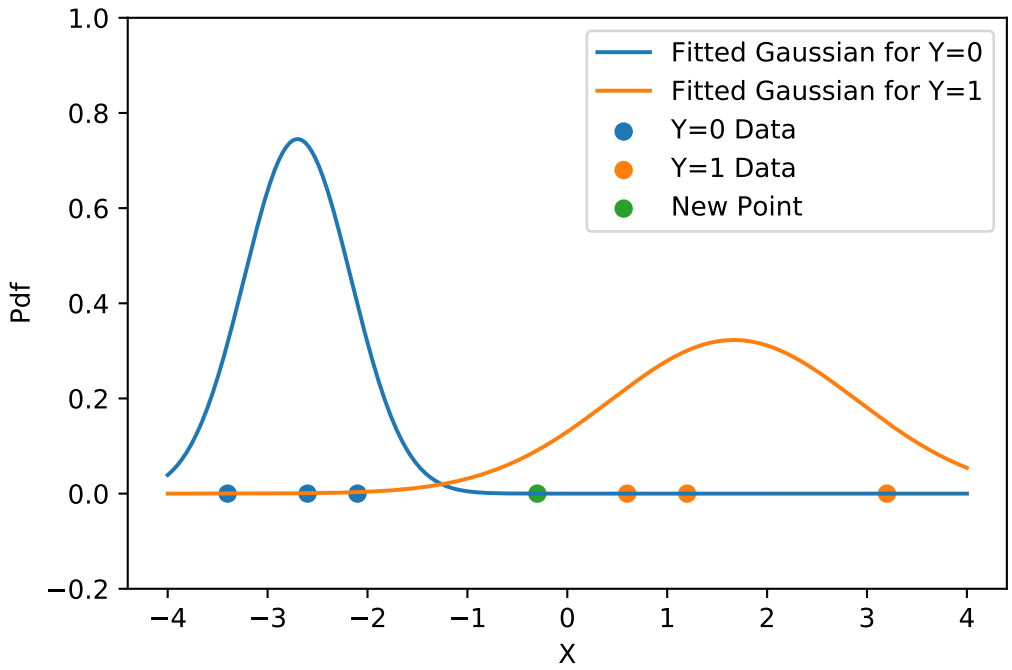
Now, we can get the variance for each class:

$$\begin{aligned} Var_0(X) &= 7.5767 - (-2.7)^2 \\ &= 0.2867 \\ Var_1(X) &= 4.0133 - \left(\frac{5}{3}\right)^2 \\ &= 1.2355 \end{aligned}$$

Now, we can get the standard deviations:

$$\begin{aligned} \sigma_0 &= \sqrt{0.2867} = 0.5354 \\ \sigma_1 &= \sqrt{1.2355} = 1.1116 \end{aligned}$$

We now have our means and standard deviations, let's plot these distributions over our data:



From this figure, we can observe where our new point is and that the probability is a lot higher for the class of $Y = 1$. However, you cannot always visualise the graphs, so to do normally do this you would calculate the actual values (using the PDF of the Gaussian distribution). These values represent **the probability that this data point came from that distribution/class**, so you would take the highest value to get our class outcome.