1. A good way to understand penalised logistic regression is by thinking about a case where we have two very highly correlated feature variables. Consider the task of collecting data from 100 people where we are trying to predict whether someone is over 167cm (about the average height in the U.S.) based on two variables: the measured length of their left leg in centimetres (X) and the measured length of their right leg in centimetres (Y). The two quantities X and Y will be very similar, but due to measurement error and some natural variation, they will not be identical for all people.

Fitting a logistic regression model to data of this form, here is what the results might look like (I completely made this up but believe the numbers are realistic):

$$\text{Prob (over 167cm)} = F(0.05 + 3.20 \times X - 1.44 \times Y)$$

Answer the following questions based on this model.

a. Look at the signs of the coefficients. What does the model seem to imply about the relationship between the length of someone's right leg (Y) and their height?

The negative sign, simplistically, implies that people with longer right legs are less likely to be over 167cm.

b. In most cases we would expect the values of X and Y for a specific observation to be very similar. Simplify the logistic regression model by assuming that X and Y are identical and writing the Prob (over 167cm) as a function of X alone (i.e., not Y).

$$F(0.05 + 3.20X - 1.44Y) = F(0.05 + 3.20X - 1.44X) = F(0.05 + 1.76X)$$

c. Another way to simplify the model would be to force the coefficients for X and Y to be the same (i.e., we would have b = c). Assuming the X and Y are so similar that their differences are not important, how could we re-write the logistic regression with coefficients for X and Y that are equal?

Want both parameters to sum to 1.76 ⇒ each are $1.76/2 = 0.88$

$$\Rightarrow F(0.05 + 0.88X + 0.88Y)$$

d. Compute the elastic net complexity scores for the original logistic regression model, the model from (b) and the model from (c). Likely all the models perform nearly equally as well. Which model(s) would seem to be preferred by the elastic net?

(a) $|3.2| + |-1.44| = 4.64$ (c) $|.88| + |.88| = 1.76$

(b) $|1.76| = 1.76$ So (b) & (c) tie for lowest complexity

e. There is another complexity score used to fit what is called **ridge regression** equal to the sum of the squared values of the coefficients. Compute the ridge penalty for the three models. Which model(s) would seem to be preferred by penalised regression using the ridge penalty?

(a) $|3.2|^2 + |-1.44|^2 = 12.31$ (c) $|\cancel{0}.88|^2 + |.88|^2 = 1.55$

(b) $|1.76|^2 = 3.0976$ So (c) lowest ridge penalty

2. In this question, you will derive a formal form of the logistic elastic net model. To begin, let's consider a prediction task with a single feature variable X and only three training data points. We will assume that the classes are the numbers 0 and 1 and the reference class we are trying to predict is 1. The training data is given and denoted by the following (L1, L2, and L3 are the labels):

X1 = 2	L1 = 1
X2 = 1	L2 = 0
X3 = 5	L1 = 1

We will be describing a logistic regression of the form:

Prob (L = 1) = F (a + b × X)

In the questions below, we will use the following notation to denote the probability of the logistic regression predictions:

P1 = F (a + b × X1)
P2 = F (a + b × X2)
P3 = F (a + b × X3)

Based on this data, answer the following questions.

a. Logistic regression tries to find parameters (here, **a** and **b**) that maximizes the probability of observing the training data. Write down the probability of observing the label L1 in terms of the predicted probability P1. [Hint: The solution is very simple]

$$P1$$

b. Repeat the previous question for the probability of observing the label L2 in terms of the predicted probability P2. [Hint: If you not used to think about probabilities, try think about a concrete example. For instance, if P2 (the probability that L2 is 1) is equal to 0.8, what is the probability that X2 is equal to 0?]

$$1-P2$$

c. The probability of independent events occurring at the same time is equal to the product of the individual probabilities. A fair coin toss has a 0.5 probability; the probability that it comes up head three times in a row is 0.5 × 0.5 × 0.5 = 0.125. Write down (in terms of P1, P2, and P3) the probability of observing the actual labels for the three data points given the model.

$$[P1] \times [1-P2] \times [P3]$$

d. A number raised to the power of 1 is equal to itself; any non-zero number raised to the power of 0 is equal to one. Using these rules, write an equation for the probability of observing L1 given the probability P1 that is true regardless of the value of L1. No conditional statements allowed!

$$[P1]^{L1}\,[1-P1]^{(1-L1)}$$

e. Combine the previous questions to write a single quantity for the probability of observing the training data given the model in terms of P1, P2, P3, L1, L2, and L3 that is true regardless of the values of the labels.

$$[P1]^{L1}[1-P1]^{(1-L1)} \times [P2]^{L2}[1-P2]^{(1-L2)} \times [P3]^{L3}[1-P3]^{(1-L3)}$$

f. The goal of logistic regression is to find parameters that maximize the probability you wrote in the previous question. A trick to solving it is to notice that the probability must be non-negative and therefore we can take the logarithm of the equation. The maximizing value of the logarithm is the same value that maximizes the original probability. Take logarithm of the value you have in (e) below. Simplify by applying the following rules: log(a × b) = log(a) + log(b) and log(a^b) = b × log(a)]

$$L1 \cdot \log(P1) + (1-L1) \cdot \log(1-P1) + L2 \log(P1) +$$
$$(1-L2) \log(1-P2) + L3 \log(P3) + (1-L3) \log(1-P3)$$

g. Rewrite the solution to (f) using summation notation. This should make it easier to read and much more compact.

$$\sum_i \left[L_i \log(P_i) + (1-L_i) \log(1-P_i) \right]$$

h. Now, write (g) in terms of the values Xi along with the parameters **a** and **b**.

$$\sum_i \left[L_i \log(a + bX_i) + (1-L_i) \log(1-a-bX_i) \right]$$

i. Since we have gotten this far, finish by extrapolating the previous question and write down the optimisation task that describes logistic regression in terms of features Xij and parameters bj, where i is an index of the observations and j is an index of the features.

$$\sum_i \left[L_i \log\left(a + \sum_j b_j X_{ij}\right) + (1-L_i) \log\left(1 - a - \sum_j b_j X_{ij}\right) \right]$$

3. The logarithm of the joint probabilities you computed as the last step in the previous question is what I was called the **FIT** of the model in the slides from today.

a. Write down the full optimisation task that describes the elastic net.

$$\text{maximise}\left[\sum_i L_i \log\left(a+\sum_j b_j x_{ij}\right) + (1-L_i)\log\left(1-a-\sum_j b_j x_{ij}\right)\right] - \lambda \cdot \sum_{j=1}^{M} |b_j|$$

b. Notice that there are a lot of negative signs above. Usually, we multiple the result in (a) by negative -1 and convert the maximization task into a minimization task. Modify your answer to (a) to write the elastic net as a minimisation task.

$$\text{minimise}\left[\sum (L_i-1)\log\left(1-a-\sum_j b_j x_{ij}\right) - L_i \log\left(a+\sum_j b_j x_{ij}\right) + \lambda \sum_j |b_j|\right]$$