1. You are tasked with designing a movie recommendation model for Netflix. You build a logistic regression model to predict the probability that a viewer likes a movie. In the test data set of 10 movies, the model delivers the following probabilities (with the class label = 1 if the viewer likes the movie, and 0 if she does not.)

Movie id	Pred. Probability	True class	Pred. class
1	0.91	1	1
2	0.89	1	1
3	0.84	0	1
4	0.82	1	1
5	0.79	1	0
6	0.73	0	0
7	0.62	1	0
8	0.54	0	0
9	0.49	0	0
10	0.48	0	0

		Pred. class	
		0	1
True class	0	4	1
	1	2	3

a. Viewers get upset if recommended a movie that they don’t like and so you set a high threshold of probability 0.8 or higher to recommend a movie. Write down the confusion matrix for the above data using this threshold.

The pred. class column contains the prediction for threshold p>=.8. See the resulting confusion matrix.

b. What is the accuracy of the model in the test data?

Accuracy = (TP+TN)/(N+P) = 7/10=70%

c. What is the precision in the test data? The recall?

Precision = TP/P*=3/4 = 75%. Recall = TP/P=3/5=60%.

d. Compute the true positive rate (TPR) and the false positive rate (FPR) in the test data for the following threshold probabilities: 1, 0.75, .5, 0.25, and 0?

For p=1, TP=0 & FP=0, so TPR=0, and FPR=0.
For p=.75, TP=4 & FP=1, so TPR=4/5=0.8, and FPR=1/5=0.2 (Note that 5 movies have predicted probabilities greater than ,75, so all will be classed as positives but only 4 are true positives and 1 is a false positive).
Similarly:
For p=.5, TP=5 & FP=3, so TPR=5/5=1, and FPR=3/5=0.6.
For p=.5, TP=5 & FP=3, so TPR=5/5=1, and FPR=3/5=0.6.
For p=.25, TP=5 & FP=5, so TPR=5/5=1, and FPR=5/5=1.
For p=0, TP=5 & FP=5, so TPR=5/5=1, and FPR=5/5=1.

2. You are given a dataset with 100 observations that has a continuous outcome y, and a single feature, x. You fit two OLS models. The first is a simple linear model i.e. regression of y on x. The second is a more flexible model in which you regress y on a cubic polynomial of x (i.e. x, x^2, x^3).

   a. Suppose the true relationship between y and x is linear ($y = \alpha + \beta x$). Would you expect the training MSE of your linear model to be lower than that of the cubic model? Explain your answer briefly.

      The cubic model would have a lower training MSE since it would not only fit the pattern but also some of the noise i.e. it would overfit the training data.

   b. What would be your answer to part (a) if you were comparing the test MSE? Explain briefly.

      The cubic model would have a higher test MSE. The reason is the same as in part a. The cubic model is an overfit and will perform poorly in the test data – it will have high variance.

   c. Suppose now that the true relationship between y and x is not linear, but we do not know the degree of non-linearity. Would you expect the training MSE of the linear model to be higher or lower than that of the cubic model? Explain briefly.

      The training MSE of the cubic model will be lower. As before, it will fit some of the noise in the data but it will also now fit some of the true nonlinear pattern.

   d. What would be your answer to part (c) if you were comparing the test MSE? Explain briefly.

      We can't say for sure. The answer will depend on the degree of nonlinearity in the true relationship. If it is highly nonlinear, the cubic model will do a better job fitting than the linear model i.e. much less bias. On the other hand, if the true relationship is only mildly nonlinear, then the linear model might still do better in the test data.

3. <u>True and False</u>

State whether the following statements are true or false, and explain your reason briefly (no more than 5 lines).

   a. LASSO and Ridge regression coefficients can sometimes be identical to OLS regression coefficients.
      TRUE. If the regularization parameter is zero i.e. no penalty on the size of the coefficients, then the objective function is the same as that of the OLS regression, and coefficients will be identical.

b. Sometimes, one might pick a classification model with a Receiver Operating Characteristic (ROC) curve that lies below the 45-degree line.

   FALSE. Random guessing will yield a ROC curve that lies on the 45-degree line. Any worthwhile modeling exercised should improve upon random guessing i.e. yield a ROC that is above the 45-degree line.

c. Suppose you run the following constrained least squares regression (LASSO) with *p* predictors and *n* observations where we seek to minimize, for a particular value of *s*:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_p x_{ij})^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s$$

   As we relax the constraint i.e. increase *s* from zero, the test MSE must decrease monotonically.

   FALSE. At *s=0*, the model is very underfitted (the prediction is simply the mean value), and so the bias is very high, as is the test MSE. As *s* increases, the fit improves and bias falls sharply with only a modest increase in variance. So, test MSE decreases. However, when *s* becomes large, there is effectively no regularization, and the model starts to overfit. The test MSE will increase.

d. For the model in part c above, as we relax the constraint i.e. increase *s* from zero, the training MSE will decrease and then stay constant.

   TRUE. At *s=0*, the model is very underfitted (the prediction is simply the mean value), and the training MSE is high. Increasing *s* allows an improved fit and training MSE falls. At large enough values of *s*, the model returns OLS coefficients which minimize the training MSE. At that point training MSE becomes constant.