Tutorial 5

1. (MLR) Consider the horseshoe female crab data given in the csv file `crab.csv`. We would want to form a model for the weight of the female crabs (kg), which depends on its width (cm) and its spine condition (1 = both good, 2 = one worn or broken, 3 = both worn or broken).

   (a) Produce a scatter plot of variable weight against width for different condition of spine.

   (b) Fit a linear regression model for weight which has two explanatories, width and spine.

   (c) Is the fitted model significant?

   (d) Derive R^2 and adjusted R^2 of the fitted model.

   (e) Write down the fitted model.

   (f) Two female crabs of the same width, find the difference of their weight if one has spines are of good condition and another one with broken spines.

   (g) Predict the weight of a female crab that has width of 27 cm and has both spines worn or broken.

2. The K-nearest neighbor classifier

   The table below provides a training data set containing six observations, three predictors, and one qualitative response variable, Y.

Obs	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

   Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

   (a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

   (b) What is our prediction with $K = 1$? Why?

   (c) What is our prediction with $K = 3$? Why?

   (d) If the Bayes decision boundary (the gold standard decision boundary) in this problem is highly non-linear, then would we expect the best value for K to be large or small? Why?

3. Measures of classifier performance

   Suppose we have developed a K-nearest neighbors classifier for predicting diabetes status. The following table shows the actual response Y (1 =yes, 0 =no) and fitted value $\widehat{Y}$ using the classifier for 10 test data points. A test data point is predicted to be $\widehat{G} = 1$ if $\hat{Y} > \delta$, for a specified threshold value δ. (Recall that we use $\delta = 0.5$ in class, also known as the majority rule).

   (a) We define

   $$TPR = \frac{TP}{TP + FN}; \quad FPR = \frac{FP}{FP + TN}.$$

   For each of the thresholds $\delta = 0.3,\ 0.6$ and 0.8, derive TPR and FPR in making predictions with the K-nearest neighbors classifier for the 10 test data points. Plot TPR against FPR for the three thresholds.

i	Y_i	$\hat{Y}_i$
1	1	0.9
2	1	0.5
3	0	0.7
4	1	0.4
5	1	0.5
6	0	0.2
7	0	0.7
8	1	0.9
9	0	0.1
10	0	0.1

		$\widehat{G}$: 1	$\widehat{G}$: 0	Total
Y	1	TP	FN	$TP+FN$
	0	FP	TN	$FP+TN$
	Total	$TP+FP$	$FN+TN$	$n=10$

(b) Can we add the two points $(0,0)$ and $(1,1)$ to the plot of TPR against FPR in part (a). Explain why or why not.

4. The CSV file `Caravan.csv` contains data on 5822 real customer records on caravan insurance purchase. This data set is owned and supplied by the Dutch data mining company, Sentient Machine Research, and is based on real world business data. Each record consists of 86 variables, containing socio-demographic data (variables 1-43) and product ownership (variables 44-86). Variable 86 (**Purchase**) indicates whether the customer purchased a caravan insurance policy.

   For this business, assume that the overall error rate (equivalently, the *accuracy*) is not of interest. Instead, the company wants to use the classifier to predict who are the potential customers likely to purchase insurance. Then the metric *precision* will be important, since it relates the proportion of individuals who will actually purchase the insurance, among the group of individuals who are predicted to purchase insurance.

   (a) Without any classifier, if the company tries to sell insurance to a random selection of customers, what is the success rate?

   (b) Standardize the input features. *Hint*: Use `scale()` command in R.

   (c) Randomly select 1000 observations to form the test data, and the remaining observations will be the training data.

   (d) Use 1-nearest neighbor classifier for the training data to predict if a customer will purchase insurance. Compute the precision of the classifier.

   (e) Repeat question 4d, for k-nearest neighbor classifier where $k = 3, 5, 10$. Which value of k gives the best precision?