DAR Coursework

1. Statistical learning methods **(10%)**

For each of parts (a) through (d), indicate whether we would generally expect the performance of a non-parametric statistical learning method to be better or worse than a parametric method. Justify your answer.

(a) The number of predictors p is large, and the number of observations n is small. (2%)

(b) The sample size n is large, and the number of predictors p is also large. (2%)

(c) The sample size n is small, and the relationship between the predictors and response is highly linear. (3%)

(d) The standard deviation of the error terms, i.e. $\sigma = \text{sd}(\varepsilon)$, is extremely high. (3%)

2. Linear regression **(20%)**

This question involves the `Auto` dataset included in the "ISLR" package.

(a) Use the `lm()` function to perform a simple linear regression with `acceleration` as the response and `cylinders` as the predictor. Use the `summary()` function to print the results. Comment on the output. For example:

i. Is there a relationship between the predictor and the response? (3%)

ii. How strong is the relationship between the predictor and the response? (3%)

iii. Is the relationship between the predictor and the response positive or negative? (3%)

iv. What is the predicted `acceleration` associated with an `cylinders` of 3.0? What are the associated 99% confidence and prediction intervals? (3%)

(b) Plot the response and the predictor. Use the `abline()` function to display the least squares regression line. (3%)

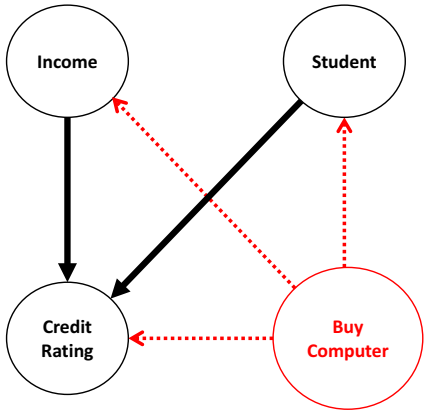
(c) Plot the 99% confidence interval and prediction interval in the same plot as (b) using different colours and legends. (5%)

3. Bayesian networks and naïve Bayes classifiers. **(30%)**

a) Given a training dataset including 30 observations and a Bayesian network indicating the relationships between 3 features (i.e. Income, Student and Credit Rate) and the class attribute (i.e. Buy Computer), please create the conditional probability tables by hand. (10%)

Training Observations	Income	Student	Credit Rating	Buy Computer
Observation_1	High	True	Fair	No
Observation_2	Low	False	Excellent	No
Observation_3	Low	True	Fair	No
Observation_4	High	False	Fair	No
Observation_5	Low	True	Excellent	Yes
Observation_6	High	False	Fair	Yes
Observation_7	High	True	Excellent	Yes
Observation_8	Low	True	Fair	No
Observation_9	Low	False	Excellent	Yes
Observation_10	Low	True	Excellent	No
Observation_11	High	True	Fair	No
Observation_12	Low	False	Fair	No
Observation_13	Low	True	Fair	No
Observation_14	High	False	Excellent	No
Observation_15	Low	True	Fair	Yes
Observation_16	High	False	Excellent	Yes
Observation_17	High	True	Excellent	No
Observation_18	Low	True	Fair	No
Observation_19	Low	False	Excellent	Yes
Observation_20	Low	True	Excellent	No
Observation_21	High	False	Excellent	Yes
Observation_21	Low	True	Excellent	Yes
Observation_23	High	False	Excellent	No
Observation_24	High	True	Fair	No
Observation_25	Low	False	Fair	Yes
Observation_26	Low	True	Fair	No
Observation_27	Low	True	Fair	No
Observation_28	Low	True	Fair	Yes
Observation_29	Low	False	Fair	No
Observation_30	High	True	Fair	Yes

Testing Observations	Income	Student	Credit Rating	Buy Computer
Observation_31	Low	True	Excellent	?
Observation_32	High	True	Fair	?



b) Make predictions for 2 testing observations by using a Bayesian network classifier. (5%)

c) Based on the conditional independence assumption between features, please create the conditional probability tables by hand. (10%)

d) Make predictions for 2 testing observations by using a naïve Bayes classifier. (5%)

4. Predicting wine quality by using support vector machine classification algorithm. (40%)

a) Download the full wine quality training and testing datasets from Moodle, and use the training dataset to find out the optimal value of hyperparameter C for a linear kernel-based svm. Define the value of the random seed equals 1 and cost = c(0.01, 1, 100). (5%)

b) Train a svm classifier by using the linear kernel and the corresponding optimal value of hyperparameter C, then make predictions on the testing dataset, report the classification accuracy. (10%)

c) Use the training dataset to find out the optimal values of hyperparameters C and for an RBF kernel-based svm. Define the value of the random seed equals 1, cost = c(0.01, 1, 100) and gamma=c(0.01, 1, 100). (5%)

d) Train a svm classifier by using the RBF kernel and the corresponding optimal values of hyperparameters C and gamma, then make predictions on the testing dataset, report the classification accuracy. (10%)

e) Train a logistic regression model. Then use the testing dataset to conduct an ROC curve analysis to compare the predictive performance of the trained logistic regression model and those two svm classifiers trained by using linear and RBF kernels respectively. (10%)