BDA/IDAR Coursework 2

- Please submit TWO files to Dropbox 2 on moodle: a .rmd file and knit it to one of the following (.doc/.pdf/.html) files. Please include any R code, plots or results.
- Your files should be named as follows:
  MSc/BSc_CW2_xxxxxxxx_initial_lastname.rmd (.pdf/.html/.doc)
  where xxxxxxxx is your student ID. For instance, MSc_CW2_12345678_T_Han.rmd.
- Don't forget to write down your programme (MSc or BSc), name and student ID on the first page of your answer sheets as well.
- Each question below has two weightings. The first weighting is for MSc students and the second weighting is for BSc students. For instance, (10% | 0%) means that the question is worth 10% for MSc students and 0% for BSc students (optional).

1. Bayesian Networks and Naïve Bayes Classifiers (20% | 25%)

Marking scheme:
- MSc: 5.0% each.
- BSc: (a), (c) 7.5% each; (b), (d) 5.0% each.

(a) Given a training dataset including 30 instances 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.

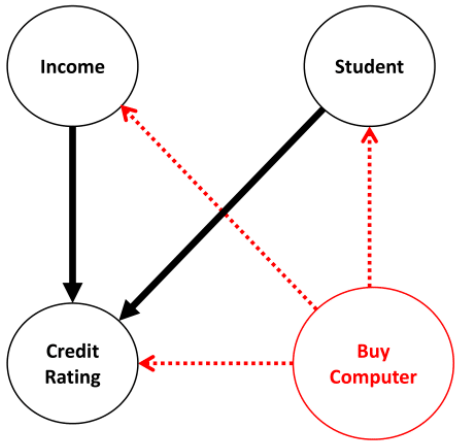
(b) Make predictions for 2 testing instances by using the Bayesian network classifier.

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

(d) Make predictions for 2 testing instances by using the naïve Bayes classifier.

Training Instances	Income	Student	Credit Rating	Buy Computer
Instance_1	High	True	Fair	Yes
Instance_2	Low	False	Excellent	No
Instance_3	Low	True	Fair	No
Instance_4	High	False	Fair	No
Instance_5	Low	True	Excellent	Yes
Instance_6	High	False	Fair	Yes
Instance_7	High	True	Excellent	Yes
Instance_8	Low	True	Fair	No
Instance_9	Low	False	Excellent	Yes
Instance_10	Low	True	Excellent	No
Instance_11	High	True	Fair	No
Instance_12	Low	False	Fair	Yes
Instance_13	Low	True	Fair	No
Instance_14	High	False	Excellent	No
Instance_15	Low	True	Fair	Yes
Instance_16	High	False	Excellent	Yes
Instance_17	High	True	Excellent	No
Instance_18	Low	True	Fair	No
Instance_19	Low	False	Excellent	Yes
Instance_20	Low	True	Excellent	No
Instance_21	High	False	Excellent	Yes
Instance_21	Low	True	Excellent	Yes
Instance_23	High	False	Excellent	No
Instance_24	High	True	Fair	No
Instance_25	Low	False	Fair	Yes
Instance_26	Low	True	Fair	No
Instance_27	Low	True	Fair	Yes
Instance_28	Low	True	Fair	Yes
Instance_29	Low	False	Fair	No
Instance_30	High	True	Excellent	No

Testing Instances	Income	Student	Credit Rating	Buy Computer
Instance_31	Low	False	Excellent	?
Instance_32	High	False	Fair	?



2. Decision Trees and Random Forests (20% | 25%)

Marking scheme: - MSc: 5.0% each. - BSc: (a), (c) 5.0% each; (b), (d) 7.5% each.

To predict room occupancy using the decision tree classification algorithm.

(a) Load the room occupancy data and train a decision tree classifier. Evaluate the predictive performance by reporting the accuracy obtained on the testing dataset.

(b) Output and analyse the tree learned by the decision tree algorithm, i.e. plot the tree structure and make a discussion about it.

(c) Train a random forests classifier, and evaluate the predictive performance by reporting the accuracy obtained on the testing dataset.

(d) Output and analyse the feature importance obtained by the random forests classifier.

3. SVM (20% | 25%)

Marking scheme: - MSc: 4.0% each. - BSc: 5.0% each.

To predict the wine quality using the support vector machine classification algorithm.

(a) Download the wine quality data and use the training dataset to conduct the grid-search to find the optimal hyperparameters of svm by using the linear kernal.

(b) Train a svm classifier by using the linear kernal and the corresponding optimal hyperparameters, then make predictions on the testing dataset, report the predictive performance.

(c) Conduct the grid-search to find the optimal hyperparameters of svm by using the RBF kernal.

(d) Train a svm classifier by using the RBF kernal and the corresponding optimal hyperparameters, then make predictions on the testing dataset, report the predictive performance.

(e) Conduct the ROC curve analysis to compare the predictive performance of svm classifiers trained by using the linear and RBF kernels respectively.

Hint: Given a pre-defined hyperparameter space - C: [0.01, 0.1, 1, 5, 10], and γ: [0.01, 0.03, 0.1, 0.5, 1].

4. Hierarchical Clustering (20% | 25%)

Marking scheme: - MSc: 5.0% each. - BSc: (a-c) 7.0% each; (d) 4.0%.

Consider the `USArrests` data. We will now perform hierarchical clustering on the states.

(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.

(b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

(c) Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.

(d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer.

5. PCA and K-Means Clustering **(20% | 0%)**

Marking scheme: - MSc: (a) 2.0%, (b-c) 4.0% each, (d-g) 2.5% each.

In this problem, you will generate simulated data, and then perform PCA and K-means clustering on the data.

(a) Generate a simulated data set with 20 observations in each of three classes (i.e. 60 observations total), and 50 variables.

Hint: There are a number of functions in `R` that you can use to generate data. One example is the `rnorm()` function; `runif()` is another option. Be sure to add a mean shift to the observations in each class so that there are three distinct classes.

(b) Perform PCA on the 60 observations and plot the first two principal components' eigenvector. Use a different color to indicate the observations in each of the three classes. If the three classes appear separated in this plot, then continue on to part (c). If not, then return to part (a) and modify the simulation so that there is greater separation between the three classes. Do not continue to part (c) until the three classes show at least some separation in the first two principal component eigenvectors.

(c) Perform K-means clustering of the observations with K = 3. How well do the clusters that you obtained in K-means clustering compare to the true class labels?

Hint: You can use the `table()` function in `R` to compare the true class labels to the class labels obtained by clustering. Be careful how you interpret the results: K-means clustering will arbitrarily number the clusters, so you cannot simply check whether the true class labels and clustering labels are the same.

(d) Perform K-means clustering with K = 2. Describe your results.

(e) Now perform K-means clustering with K = 4, and describe your results.

(f) Now perform K-means clustering with K = 3 on the first two principal components, rather than on the raw data. That is, perform K-means clustering on the 60×2 matrix of which the first column is the first principal component's corresponding eigenvector, and the second column is the second principal component's corresponding eigenvector. Comment on the results.

(g) Using the `scale()` function, perform K-means clustering with K = 3 on the data after scaling each variable to have standard deviation one. How do these results compare to the true class labels? Will the scaling affect the clustering?