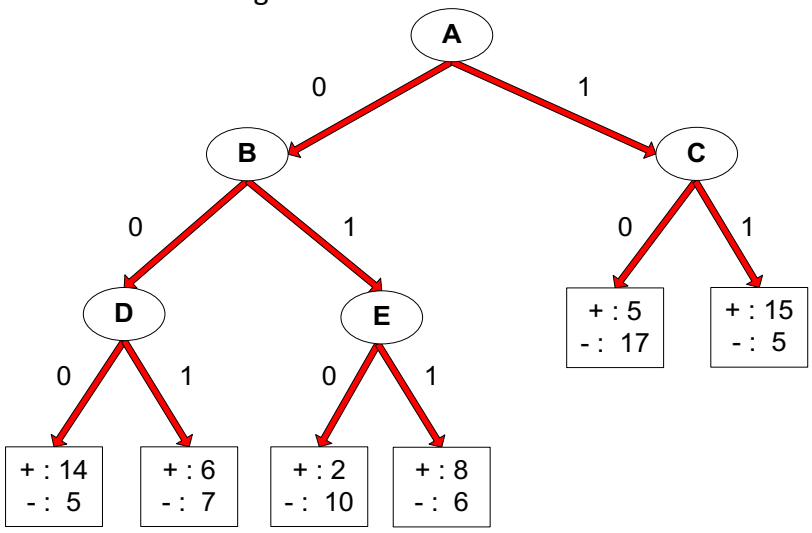
DM Models 1

Task 1 (20 points) For the Titanic challenge (https://www.kaggle.com/c/titanic), we need to guess whether the individuals from the test dataset had survived or not. Please:

1) Preprocess your Titanic training data;
2) (5 points) Learn and fine-tune a decision tree model with the Titanic training data, plot your decision tree;
3) (5 points) Apply the five-fold cross validation of your fine-tuned decision tree learning model to the Titanic training data to extract average classification accuracy;
4) (5 points) Apply the five-fold cross validation of your fine-tuned random forest learning model to the Titanic training data to extract average classification accuracy;
5) (5 points) Which algorithm is better, Decision Tree or Random Forest? What are your observations and conclusions from the algorithm comparison and analysis?

Task 2 (15 points) Understanding Training Error and Testing

Consider the decision tree shown in the diagram below. The counts shown in the leaf nodes correspond to the number of training records associated with the nodes.



(a) (10 points) What is the training error rate for the tree? Explain how you get the answer?

(b) (5 points) Given a test instance T={A=0, B=1, C=1, D=1, E=0}, what class would the decision tree above assign to T? Explain how you get the answer?

Task 3 (20 points) Understand Splitting Process
Consider the following data set for a binary class problem.

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

Q1: (5 points) What is the overall gini before splitting?

Q2: (5 points) What is the gain in gini after splitting on A?

Q3: (5 points) What is the gain in gini after splitting on B:

Q4: (5 points) Which attribute would the decision tree choose?

Task 4: (10 points) Please answer and explain.
Q1: (5 points) Are decision trees a linear classifier? Why?
Q2: (5 points) Is Misclassification error better than Gini index as the splitting criteria for decision trees? Why?

Task 5: (10 points) What are the weaknesses of bagging? What is the difference between bagging and random forests, and why such difference can overcome the weaknesses of bagging?

Task 6: (20 points) Construct a support vector machine that computes the kernel function. Use four values of +1 and -1 for both inputs and outputs:
[−1, −1] (negative)
[−1, +1] (positive)
[+1, −1] (positive)
[+1, +1] (negative).
Map the input [x1, x2] into a space consisting of x1 and x1x2. Draw the four input points in this space, and the maximal margin separator. What is the margin? 【To be consistent with our lecture notes, margin is defined as the distance from the middle way/hyperplane to either support vectors. 】

Task 7: (10 points) Recall that the equation of the circle in the 2-dimensional plane is (x1 − a)^2 + (x2 − b)^2 − r2 = 0. Please expand out the formula and show that every circular region is linearly separable from the rest of the plane in the feature space (x1, x2, x1^2, x2^2).

Task 8: (10 points) Recall that the equation of an ellipse in the 2-dimensional plane is c(x1 − a)^2 + d(x2 − b)^2 − 1 = 0. Please show that an SVM using the polynomial kernel of degree 2, K(u, v) = (1 + u · v)^2, is equivalent to a linear SVM in the feature space (1, x1, x2, x1^2, x2^2, x1x2) and hence that SVMs with this kernel can separate any elliptic region from the rest of the plane.

Please submit a PDF report. In your report, please answer each question with your explanations, plots, results in brief. DO NOT paste your code or snapshot into the PDF. At the end of your PDF, please include a website address (e.g., Github, Dropbox, OneDrive, GoogleDrive) that can allow the TA to read your code.