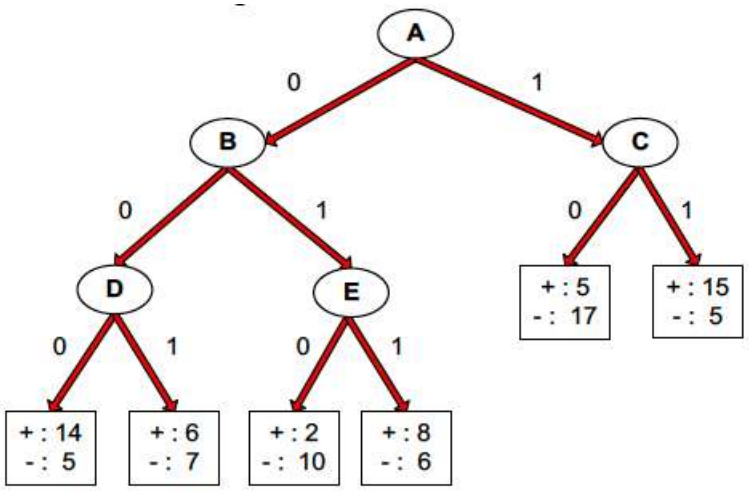
CSE 572 Data Mining HW2: Data and Data Preprocessing Problem

Zhandaulet Yespossynov | ASU ID 1233282975 | zyesposs@asu.edu

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 include a website address (e.g., Github, Dropbox, OneDrive, GoogleDrive) that can allow the TA to read your code.
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?

https://github.com/zyesposs/cse-572-hw1.git << link to github.

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.
Q1:



Q1: (10 points) What is the training error rate for the tree? Explain how you get the answer?

Training Error is computed as number of misclassified samples divided by total samples. So, for given tree misclassified one are (5+6+2+6+5+5) = 29 and total samples (14+5+6+7+2+10+8+6+5+17+15+5)=100. There Training Error rate = 29/100.
That means about 29% of training samples were classified wrong. Let me explain one misclassification. Let's take ABD so when A=0, B=0, D=0, its shows positive is 14 and 5 is negative and as positive bigger than negative it predicts it as positive. But wait, it has 5 negatives, and those will be counted as misclassified as positive. Like this we need to count for all others and get training error rate.

Q2 : (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?

Following the tree given we need to go from A=0 go left and B=1 go right and E=0 got left, we will reach 2 positive and 10 negative. So, model will predict negative as 2 < 10. In case of Titanic, it means passenger is not survived. We will skip C and D because there are not in this path and tree only checks features that appear in its current branch.

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?

We have 10 records and 4 positive and 6 negative.

So, we need to find formula GiniTotal = 1 – (P(+)^2 + P(-)^2).
P(+) = 4/10 and P(-) = 6/10

GiniTotal = 1-((4/10)^2+(6/10)^2)= 1-((16+36)/100) = 1-52/100 = 48/100=0.48

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

A	*Count*	*+*	*-*	*Gini for group*
T	*7*	*4*	*3*	*1-(4/7)^2 – (3/7)^2 = 0.48*
F	*3*	*0*	*3*	*1-(0/3)^2 – (3/3) ^2 = 0*

After split Gini = 7/10 x (0.48) + 3/10 x 0 = 0.34

Gini Gain = 0.48-0.34 = 0.14

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

B	*Count*	*+*	*-*	*Gini for group*
T	*4*	*3*	*1*	*1-(3/4)^2 – (1/4)^2 = 0.37*
F	*6*	*1*	*5*	*1-(1/6)^2 – (5/6) ^2 = 0.27*

After split Gini = 4/10 x (0.37) + 6/10 x 0.27 = 0.31 Gini Gain = 0.48-0.31 = 0.17

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

B because it yields the lowest weighted Gini - 0.31 and the highest gain 0.17.

Task 4: (10 points) Please answer and explain.
Q1: (5 points) Are decision trees a linear classifier? Why?
No, because decistion trees are not acting as linear classifier where its separating classes using straight single line or hyperplane in features space and also splits in DT is simple that forms a non-linear boundary overall. If we plot data split by a decision tree, the boundary looks like steps or rectangles, not a single diagonal line.

Q2: (5 points) Is Misclassification error better than Gini index as the splitting criteria for decision trees? Why?

No, Gini index is generally better than misclassification error for splitting in decision trees.

Gini index is preferred because it is more sensitive to class probabilities and produces purer child nodes, leading to better decision trees than misclassification error.

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?

Bagging reduces variance but its trees often stay too similar since all use same features so correlation stays high. Random Forest fixes this by picking random subset of features for each split, making trees more different and less correlated. This randomness helps overcome baggings weakness and improves accuracy and generalization.

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. 】
After mapping the points, SVM separates them by the line x1x2=0 where support vectors lie at x1x2=±1, so distance from middle line to either side is 1and margin is 1.

Task 7: (10 points) Recall that the equation of the circle in the 2-dimensional plane is $(x_1 - a)^2 + (x_2 - b)^2 - r^2 = 0$. Please expand out the formula and show that every circular region is linearly separable from the rest of the plane in the feature space (x_1, x_2, x_1^2, x_2^2).

If we expand (x1−a)2+(x2−b)2−r2=0 we get x12+x22−2ax1−2bx2+(a2+b2−r2)=0.
If we take features (x1,x2,x12,x22), this becomes linear equation in that space, so any circular region can be linearly separated from rest of the plane.

Task 8: (10 points) Recall that the equation of an ellipse in the 2-dimensional plane is $c(x_1 - a)^2 + d(x_2 - b)^2 - 1 = 0$. Please show that an SVM using the polynomial kernel of degree 2, $K(u, v) = (1 + u \cdot v)^2$, is equivalent to a linear SVM in the feature space $(1, x_1, x_2, x_1^2, x_2^2, x_1x_2)$ and hence that SVMs with this kernel can separate any elliptic region from the rest of the plane.

When we expand c(x1−a)2+d(x2−b)2−1=0 we get cx12+dx22−2acx1−2bdx2+(ca2+db2−1)=0.
This equation looks linear if we see it in feature space (1,x1,x2,x12,x22,x1x2) SVM with polynomial kernel of degree 2, K(u,v)=(1+u ·v)2, map data into same kind of space, so it act like linear SVM there and can separate any ellipse region from rest of plane.