CS 481, Midterm 2 Exam (Spring 2015)

Time: 85 minutes, Total: 75

Name:

Question	Available Marks	Obtained Marks
1	08	
2	15	
3	15	
4	15	
5	10	
6	12	
Total	75	

Questions

3: Consider the following dataset:

Tuple #	Age	CarType	Class Label
1	25	Family	Low
2	20	SUV	High
3	20	Sedan	Low
4	25	SUV	Low
5	45	Family	Low
6	23	SUV	High
7	25	Sedan	High
8	20	Sedan	Low
9	22	Family	High

For decision tree classification, which of the following two splits is better based on gini-index criterion. Note that, $gini(D) = 1 - \sum_{i=1}^{k} p(c_i|D)^2$. (i) $Age \geq 24$ (ii) $CarType = Family$. Answer by computing information gain for the above two cases. (15)

4: Draw the ROC curve from the information in the following table. Clearly mark each of the data points on this curve. If we want to maximize the F-score, what probability threshold should we use to predict the class P? (10+5)

Tuple #	Class	$p(P\|\mathbf{x_i})$
1	P	0.95
2	P	0.85
3	N	0.76
4	P	0.55
5	N	0.50
6	P	0.42
7	N	0.30
8	N	0.22

5: Use naive Bayes to classify the data tuple $(22, Sedan)$ (model the age feature using Normal distribution, the pdf for this is $f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$) (8)

Tuple #	Age	CarType	Class
1	25	Family	Low
2	20	SUV	High
3	45	Family	Low
4	20	Sedan	Low
5	25	SUV	Low
6	45	Family	Low
7	23	SUV	High
8	25	Sedan	High
9	20	Sedan	Low
10	22	Family	High

6:

a. What is generalization? When should we suspect that a classification model is probably overfitting? (6)

b. Discuss bias error and variance error of a classifier. What do you mean by bias-variance tradeoff. (6)