CS 418: Introduction to Data Science
Homework Assignment 03
Fall 2019

Instructions

This assignment is due Friday, October 25, at 11:59PM (Central Time).

Answers for this assignment must be entered on *Blackboard*. You have 1 attempt to submit the assignment. This attempt is not timed. The assignment will close automatically at the due date and no late submissions will be accepted.

This assignment is individual. Offering or receiving any kind of unauthorized or unacknowledged assistance is a violation of the University's academic integrity policies, will result in a grade of zero for the assignment, and will be subject to disciplinary action.

Part I: Classification (100 pt.)

Problem 1 (10 pt.)

Match each statement with the corresponding classification technique:

- **Easy to understand and interpret.**
- **Provides probabilities that quantify the uncertainty in the predictions.**
- **Produces decision boundaries of any shape.**
- **Produces rectilinear decision boundaries.**
- **Produces both linear and nonlinear decision boundaries.**
- **Assumes that attributes are conditionally independent.**
- **Training is computationally expensive.**
- **Testing is computationally expensive.**

a. **Decision Trees**

b. **k-Nearest Neighbors**

c. **Naïve Bayes**

d. **Support Vector Machines**

Problem 2 (15 pt.)

Given the following dataset with the true class and the class predicted by a classifier:

True class	+	-	+	-	-	+	-	+	+	-	-	-	-	+	+	-	+	-	-	+
Predicted class	+	-	+	-	-	-	+	+	-	-	-	+	-	+	+	-	-	-	-	+

Answer the following questions:

a. **(5 pts.) Complete the following confusion matrix for the classifier, where the columns indicate the predicted class and the rows indicate the true class:**

		Predicted class	
		$y = +$	$y = -$
True class	$y = +$	$\boldsymbol{A}$	$\boldsymbol{B}$
	$y = -$	$\boldsymbol{C}$	$\boldsymbol{D}$

b. (10 pts.) Compute the accuracy, error, precision, recall, and F1 score of the classifier.

Problem 3 (35 pt.)

Given the following dataset:

A	B	C	Class
0	0	1	-
1	0	1	+
0	1	0	-
1	0	0	-
1	0	1	+
0	0	1	+
1	1	0	-
0	0	0	-
0	1	0	+
1	1	1	+

Suppose that you are building a decision tree from this dataset. Answer the following questions:

a. (2 pts.) What is the entropy of the root node?

b. (10 pts.) What is the entropy of each child and what is the information gain if you split the root node of the tree on attribute A, B, or C?

c. (2 pts.) Which of the attributes would you select to split the root node of the tree?

d. (3 pts.) Suppose that you build a decision tree by splitting the root node of the tree on the attribute selected in Problem 3.c. What is the class label predicted using this tree for an observation with $A = 0$, $B = 0$, and $C = 1$?

Another commonly used impurity measure in decision trees is the Gini index. The Gini index of a node v of the tree is given by:

$$Gini\ index(v) = 1 - \sum_{k} p(k|v)^2$$

where $p(k|v)$ is the proportion of observations from class k at node v. The gain in the Gini index resulting from splitting a node v of the tree on an attribute X is given by:

$$Gain(v, X) = Gini\ index(v) - \sum_{h=1}^{m} \frac{n_h}{n} \cdot Gini\ index(v_h)$$

where n is the number of observations at node v, n_h is the number of observations at node v_h, and nodes v_h with $h = 1, \dots, m$ are the children of node v in the tree.

e. **(2 pts.) What is the Gini index of the root node?**

f. **(10 pts.) What is the Gini index of each child and what is the gain in the Gini index if you split the root node of the tree on attribute A, B, or C?**

g. **(2 pts.) Does your answer for Problem 3.c change if you use the Gini index instead of entropy as the impurity measure?**

h. **(4 pts.) For any binary classification problem, what is the maximum and minimum value of the Gini index of a node of the tree?**

Problem 4 (20 pt.)

Given the dataset in Problem 3, answer the following questions:

a. **(4 pts.) Find $P(A=0|+)$, $P(A=0|-)$, $P(B=0|+)$, $P(B=0|-)$, $P(C=1|+)$, $P(C=1|-)$, $P(+)$, and $P(-)$.**

b. **(8 pts.) Use the Naïve Bayes approach to estimate the probability that an observation is positive and the probability that an observation is negative given that $A=0$, $B=0$, and $C=1$.**

c. **(3 pts.) What is the class label predicted using the Naïve Bayes approach for an observation with $A=0$, $B=0$, and $C=1$?**

d. **(5 pts.) What is the class label predicted using the Naïve Bayes approach for an observation with $A=0$, $B=1$, and $C=0$?**

Problem 5 (20 pt.)

Given the following dataset:

A	B	C	Class
1.5	2.0	3.0	+
3.0	1.0	2.0	-
3.5	2.5	1.5	-
2.5	2.0	2.5	+
1.0	0.5	0.5	-
1.5	0.5	1.0	+

Suppose that you want to classify an observation $z=(2.5,\ 0.5,\ 2.5)$ using k-Nearest Neighbors with Euclidean distance as the proximity metric. Answer the following questions:

a. **(8 pts.) What is the distance between z and every observation in the dataset?**

b. **(3 pts.) What is the predicted class label for z if $k=1$?**

c. **(3 pts.) What is the predicted class label for z if $k=2$?**

d. **(3 pts.) What is the predicted class label for z if $k=3$?**

e. **(3 pts.) What is the predicted class label for z if $k=4$?**