Machine Learning Course 2024 Spring: Homework 4

May 19, 2024

1 Problem 1

Suppose you are a computer salesperson. Given the dataset in Table 1 below, please help build a decision tree to decide whether a certain customer will buy a computer.

ID	age	income	student	credit rating	buy computer
1	<30	high	no	fair	no
2	<30	high	no	excellent	no
3	30-40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	30-40	low	yes	excellent	yes
8	<30	medium	no	fair	no
9	<30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<30	medium	yes	excellent	yes
12	30-40	medium	no	excellent	yes
13	30-40	high	yes	fair	yes
14	>40	medium	no	excellent	no

Table 1: Instances for building a decision tree.

1. Which attribute (age, income, student, or credit rating) should be chosen to split the data for the maximum **information gain** at the first time? Please show some key parts

of your calculation. (Hint: $\text{Gain}(D, a) = \text{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|}\text{Ent}(D^v)$, for attribute a splitting root into V branches $D^1, \cdots, D^V$)

2. Which attribute (age, income, student, or credit rating) should be chosen to split the data for the minimum **Gini index** at the first time? Please show some key parts of your calculation. (Hint: $\text{Gini\_index}(D, a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|}\text{Gini}(D^v)$, $\text{Gini}(D) = 1 - \sum_{k=1}^{|\mathcal{Y}|} p_k^2$, for attribute a splitting root into V branches $D^1, \cdots, D^V$)

2 Problem 2

Suppose we have a dataset containing the following several instances, each with "Offers" and "Lottery" two features, and a class label (spam or normal). Now we have a sample, Offers=no, Lottery=yes, which category it might belong to by naive Bayes.

E-mail	"Offers"	"Lottery"	Category
1	yes	yes	spam
2	no	yes	spam
3	yes	no	spam
4	no	no	spam
5	yes	no	spam
6	yes	no	normal
7	no	yes	normal
8	no	no	normal

Table 2: Instances for naive Bayes.

3 Problem 3

Suppose we have a dataset containing 9 instances $\boldsymbol{x}_i$ $(1 \leq i \leq 9)$ shown in Table 3, each with two features (feature_1 and feature_2). Please cluster it using the **DBSCAN** algorithm $(\epsilon = 1, \text{MinPts} = 3)$. We choose the Euclidean distance as the distance metric. Write down the calculation process in detail.

ID	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_6$	$\boldsymbol{x}_7$	$\boldsymbol{x}_8$	$\boldsymbol{x}_9$
feature_1	1	2	2	2	3	5	6	6	6
feature_2	2	1	2	3	2	2	1	2	3

Table 3: Instances for DBSCAN.