Machine Learning Course 2024 Spring: Homework 4

May 19, 2024

1 Problem 1

1. **Solution:**

   Firstly, we have the entropy of the root node:

$$\text{Ent}(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k = -\left(\frac{9}{14}\log_2\frac{9}{14} + \frac{5}{14}\log_2\frac{5}{14}\right) = 0.940.$$

   For attribute "age"="<30":

$$\text{Ent}(D^1) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0.971.$$

   For attribute "age"="30-40":

$$\text{Ent}(D^2) = 0.$$

   For attribute "age"=">40":

$$\text{Ent}(D^3) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.971.$$

   Thus, the information gain

$$\text{Gain}(D, \text{"age"}) = \text{Ent}(D) - \left(\frac{5}{14}\text{Ent}(D^1) + \frac{4}{14}\text{Ent}(D^2) + \frac{5}{14}\text{Ent}(D^3)\right) = 0.247.$$

   A similar calculation can be applied to the rest attributes:

$$\text{Gain}(D, \text{"income"}) = 0.029,$$

$$\text{Gain}(D, \text{"student"}) = 0.152,$$

$$\text{Gain}(D, \text{"credit\_rating"}) = 0.048.$$

   So, the attribute "age" should be chosen for the maximum information gain. □

2. **Solution:**

For attribute "age"="<30":

$$\text{Gini}(D^1) = 1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right) = 0.480.$$

For attribute "age"="30-40":

$$\text{Gini}(D^2) = 0.$$

For attribute "age"=">40":

$$\text{Gini}(D^3) = 1 - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right) = 0.480.$$

Thus, the Gini index

$$\text{Gini\_index}(D, \text{``age''}) = \frac{5}{14}\text{Gini}(D^1) + \frac{4}{14}\text{Gini}(D^2) + \frac{5}{14}\text{Gini}(D^3) = 0.343.$$

A similar calculation can be applied to the rest attributes:

$$\text{Gini\_index}(D, \text{``income''}) = 0.440,$$

$$\text{Gini\_index}(D, \text{``student''}) = 0.367,$$

$$\text{Gini\_index}(D, \text{``credit\_rating''}) = 0.429.$$

So, the attribute "age" should be chosen for the minimum Gini index. □

2 Problem 2

First, calculate the prior probabilities of spam and normal mail:

$$P(\text{Spam}) = \frac{5}{8}$$
$$P(\text{Normal}) = \frac{3}{8}$$

Then, calculate the conditional probability of including "Offers" and "Lottery" under spam and normal mail:

$$P(\text{Offers=yes}|\text{Spam}) = \frac{3}{5}$$
$$P(\text{Lottery=no}|\text{Spam}) = \frac{3}{5}$$
$$P(\text{Offers=yes}|\text{Normal}) = \frac{1}{3}$$
$$P(\text{Lottery=no}|\text{Normal}) = \frac{2}{3}$$

Next, calculate the probability that a new message containing "Offers" and "Lottery" is spam versus normal:

$$P(\text{Offers=yes, Lottery=no}|\text{Spam}) = \frac{9}{40}$$
$$P(\text{Offers=yes, Lottery=yes}|\text{Normal}) = \frac{1}{6}$$

Finally, the posterior probability that the new message belongs to spam and normal mail is calculated according to Bayes' theorem:

$$P(\text{Spam}|\text{Offers=yes, Lottery=no}) = \frac{9}{16 \times P(\text{Offers=yes, Lottery=no})}$$
$$P(\text{Normal}|\text{Offers=yes, Lottery=no}) = \frac{1}{4 \times P(\text{Offers=yes, Lottery=no})}$$

So it's more likely spam.

3 Problem 3

Solution:

Table 1 shows the ϵ-neighborhood for every instance, so that we can identify the set of core objects: $\Omega = \{\boldsymbol{x}_3, \boldsymbol{x}_8\}$.

ID	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_6$	$\boldsymbol{x}_7$	$\boldsymbol{x}_8$	$\boldsymbol{x}_9$
neighborhood	$\boldsymbol{x}_3$	$\boldsymbol{x}_3$	$\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_4, \boldsymbol{x}_5$	$\boldsymbol{x}_3$	$\boldsymbol{x}_3$	$\boldsymbol{x}_8$	$\boldsymbol{x}_8$	$\boldsymbol{x}_6, \boldsymbol{x}_7, \boldsymbol{x}_9$	$\boldsymbol{x}_8$

Table 1: ϵ-neighborhood for every instance

Then, we randomly select a core object from as a seed and expand from it to include all density-reachable instances. These instances form a cluster. Suppose the core object $\boldsymbol{x}_3$ is selected as the seed, then the first generated cluster is

$$C_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}.$$

After that, DBSCAN removes all core objects in C_1 from Ω, that is, $\Omega = \Omega \backslash C_1 = \{\boldsymbol{x}_8\}$. Then, the next cluster is generated by selecting another core object from the updated Ω as seed. Then the second generated cluster is

$$C_2 = \{\boldsymbol{x}_6, \boldsymbol{x}_7, \boldsymbol{x}_8, \boldsymbol{x}_9\}.$$

□