Machine Learning: Homework 3

Laurent HAYEZ

October 23, 2016

Exercise 1. *Invent the two characters Taylor and Robin each with a weight in the range of [63-74] kg and a shoe size in the range of [40-44]. Decide (by hand) with Naïve Bayes using the probability density function whether your Taylor and Robin are female or male based on the following statistics.*

Solution. As we have to deal with numerical attributes, we need to use a probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ is the sample mean and σ is the standard deviation.

In Table 1, we represented the weight and shoe size according to whether the person is male or female, and we computed the sample means and standard deviations.

Let f_1 be a gaussian distribution with $\mu = \mu_1$ and $\sigma = \sigma_1$ representing the distribution of the weight for the males. Let g_1 be the same but with $\mu = \mu_2$ and $\sigma = \sigma_2$ representing the distribution of the weight for the females. In the same fashion we define f_2 and g_2 corresponding to the shoe size. Moreover let $L[\cdot]$ denote the likelihood of an event.

If Taylor weights 73 kgs and has shoe size 41, and Robin weights 67 kgs and has shoe size 44, then define

$$E_1 = \{\text{weight} = 73, \text{shoe size} = 41\}$$

and

$$E_2 = \{\text{weight} = 67, \text{shoe size} = 44\}.$$

Table 1: Table of values for the weights and shoe sizes for men and women

Weight		Shoe size	
Male	Female	Male	Female
82, 80, 77, 72	60, 64, 59, 65	45, 45, 43, 43	39, 41, 40, 41
$\mu_1 = 77.75$ $\sigma_1 = 4.343$	$\mu_2 = 62$ $\sigma_2 = 2.943$	$\mu_3 = 44$ $\sigma_3 = 1.154$	$\mu_4 = 40.25$ $\sigma_4 = 0.957$

We have

$$\begin{aligned} L[\text{Taylor} = \text{male} \mid E_1] &= f_1(73) \cdot f_2(41) \cdot \frac{1}{2} \qquad \text{where } \mathbb{P}[\text{male}] = \frac{1}{2} \\ &= 0.0505 \cdot 0.0118 \cdot \frac{1}{2} \\ &= 0.00029 \end{aligned}$$

Similarly, we have

$$L[\text{Taylor} = \text{female} \mid E_1] = g_1(73) \cdot g_2(41) \cdot \frac{1}{2} = 0.000019.$$

Hence we can compute the probabilities, knowing that $\mathbb{P}[E_1] = 0.00029 + 0.000019 = 0.000318$.

$$\mathbb{P}[\text{Taylor} = \text{male} \mid E_1] = \frac{L[\text{Taylor} = \text{male} \mid E_1]}{\mathbb{P}[E]} = 0.939,$$

$$\mathbb{P}[\text{Taylor} = \text{female} \mid E_1] = \frac{L[\text{Taylor} = \text{female} \mid E_1]}{\mathbb{P}[E]} = 0.0607.$$

Applying the same reasoning for Robin gives us $L[\text{Robin} = \text{male} \mid E_2] = 0.00074$ and $L[\text{Robin} = \text{female} \mid E_2] = 3.112 \cdot 10^{-6}$, hence $\mathbb{P}[E_2] = 0.00074 + 3.112 \cdot 10^{-6}$.

$$\mathbb{P}[\text{Robin} = \text{male} \mid E_2] = \frac{L[\text{Robin} = \text{male} \mid E_2]}{\mathbb{P}[E_2]} = 0.9958,$$

$$\mathbb{P}[\text{Robin} = \text{female} \mid E_2] = \frac{L[\text{Robin} = \text{female} \mid E_2]}{\mathbb{P}[E_2]} = 0.0041.$$

□

Exercise 2. *Build a decision tree (by hand) based on the following data and using*

information theory. Then use WEKA to see if you get the same decision tree (ID3). Use this tree to decide if Edward, the senior student, would buy a new PC with his high income and good credit rating.

Solution. To build the decision tree, we look at each attribute and compute the information gain. The initial information is given by $I(9/14, 5/14) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940$.

Age: youth: We have 3 no and 2 yes, hence $I(2/5, 3/5) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.971$ bits.

Middle-aged: We have 4 yes, hence $I(1, 0) = 0$ bits.

Senior: We have 3 yes and 2 no, hence $I(3/5, 2/5) = 0.971$ bits.

With these values, we can compute the information that "Age" gives us and the information gain. We have $\text{Info}(Age) = \frac{5}{14}\cdot 0.971 + \frac{4}{14}\cdot 0 + \frac{5}{14}\cdot 0.971 = 0.693$ bits. Hence

$$\text{gain}(Age) = 0.940 - 0.693 = 0.247.$$

Applying the exact same method to the attributes "Income", "Student" and "Credit rating", we obtain

$$\text{gain}(Income) = 0.015$$

$$\text{gain}(Student) = 0.359$$

$$\text{gain}(Credit\ rating) = 0.048.$$

Hence the attribute Student gives us the most information, so it will be the first node of our decision tree.

We need to compute the information gain of "Age", "Income" and "Credit rating" given that Student=yes on the one hand, and Student=no on the other hand. The new initial information when Student=yes is $I(8/9, 1/9) = 0.503$ bits. We have

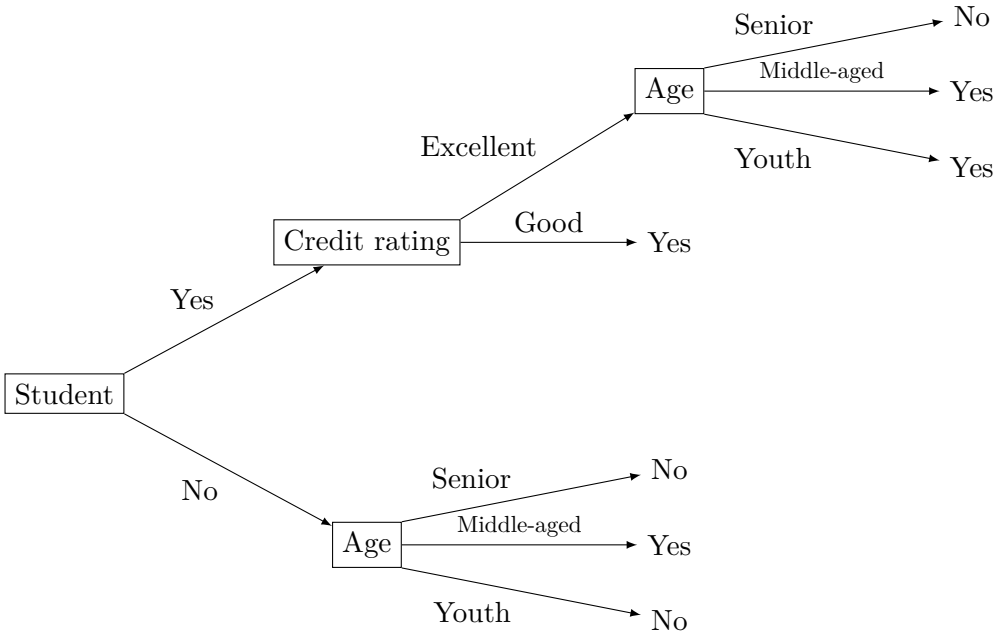
$$\text{gain}(Age) = 0.048$$

$$\text{gain}(Income) = 0.048$$

$$\text{gain}(Credit\ rating) = 0.197.$$

Hence the second node of the decision tree will be given by the Credit rating. We keep on computing the information gain with the same method when Student=yes and Credit rating=excellent, because now we know that if the person is a student and has a good credit rating, he will buy the computer.

We need to apply the same strategy when Student=no. In the end we computed the following decision tree:



We conclude that, as Edward is a student and has a good credit rating, he would by the PC.

From the file `ML_hayezl_computerRelation.arff` we computed the decision tree in WEKA with ID3 and we obtained the same decision tree (result is saved in the file `ML_hayezl_computerRelationResults.txt`). Using the prediction with WEKA, we also obtained that Edward will buy the PC. □