Machine Learning: Homework 2

Laurent Hayez

October 16, 2016

Exercise 1. *Find out (by hand or WEKA) with the simple rule (1R) which attribute best predicts whether a car gets stolen or not.*

> **Solution.** Using the file `ML_hayezl_carRelation.arff` and WEKA, we derive the following one rule:
>
> `price=low` $\Longrightarrow$ `FALSE`, `price=medium` $\Longrightarrow$ `FALSE`, `price=high` $\Longrightarrow$ `TRUE`.
>
> The details can be found in the file `ML_hayezl_carRelationResults.txt`. □

Exercise 2. *Decide (by hand and WEKA) whether Logan is Scottish based on the following attributes and using a Naïve Bayes classifier. Logan likes shortbread, drinks whiskey and eats porridge but doesn't like lager and doesn't watch soccer. Bonus: use a smoothing technique. (+1p)*

> **Solution.** Let $E = \{$shortbread = yes, lager = no, whisky = yes, porridge = yes, soccer = no$\}$. We need to compute $\mathbb{P}[\text{Logan is Scottish} \mid E]$. With the naive Bayes

approach, we need to compute

$$\begin{aligned}\mathbb{P}[\text{Logan is Scottish} \mid E] = \frac{1}{\mathbb{P}[E]} \cdot & \mathbb{P}[\text{shortbread} = \text{yes} \mid yes] \cdot \\ & \mathbb{P}[\text{lager} = \text{no} \mid yes] \cdot \\ & \mathbb{P}[\text{whisky} = \text{yes} \mid yes] \cdot \\ & \mathbb{P}[\text{porridge} = \text{yes} \mid yes] \cdot \\ & \mathbb{P}[\text{soccer} = \text{no} \mid yes] \cdot \\ & \mathbb{P}[\text{being Scottish}]\end{aligned}$$

Computing the conditional probabilities and the probability of being Scottish using Table 1, we obtain

$$\frac{6}{3} \cdot \frac{3}{7} \cdot \frac{4}{7} \cdot \frac{5}{7} \cdot \frac{4}{7} \cdot \frac{7}{13} = 0.046.$$

Computing the same thing but for $\mathbb{P}[\text{Logan is not Scottish} \mid E]$, we obtain 0.0064. Hence

$$\mathbb{P}[E] = 0.046 + 0.0064 = 0.052,$$

$$\mathbb{P}[\text{Logan is Scottish} \mid E] = 0.878,$$

$$\mathbb{P}[\text{Logan is not Scottish} \mid E] = 0.122.$$

Hence Logan is Scottish with probability 0.878.

We can use for example the Laplace estimator as a smoothing technique so that we don't have conditional probabilities being equal to 0. The Laplace estimator add 1 to the count for every attribute value class combination. Hence Table 1 becomes Table 2. The previous computations become $\mathbb{P}[E] = 0.038 + 0.008 = 0.046$,

$$\mathbb{P}[\text{Logan is Scottish} \mid E] = 0.825$$

$$\mathbb{P}[\text{Logan is not Scottish} \mid E] = 0.174$$

so the conclusion does not change.

WEKA gives the same result by training it with the file `ML_hayezl_scottsRelation.arff` and predicting if Logan is Scottish or not with the file `ML_hayezl_scottsRelationPredict.arff`. The results can be found in `ML_hayezl_scottsRelationResults.txt` for the training and in `ML_hayezl_scottsRelationPredictResults.txt` for the prediction. □

Table 1: Probabilities of the attributes given that the person is Scottish or not

Shortbread	yes	no	Lager	yes	no	Whisky	yes	no
yes	6	3	yes	4	3	yes	4	2
no	1	3	no	3	3	no	3	4
yes	6/7	3/6	yes	4/7	3/6	yes	4/7	2/6
no	1/7	3/6	no	3/7	3/6	no	3/7	4/6
Porridge	yes	no	Soccer	yes	no			
yes	5	3	yes	3	4			
no	2	3	no	4	2			
yes	5/7	3/6	yes	3/7	4/6			
no	2/7	3/6	no	4/7	2/6			

Table 2: Probabilities of the attributes given that the person is Scottish or not (with Laplace estimator)

Shortbread	yes	no	Lager	yes	no	Whisky	yes	no
yes	7	4	yes	5	4	yes	5	3
no	2	4	no	4	4	no	4	5
yes	7/9	4/8	yes	5/9	4/8	yes	5/9	3/8
no	2/9	4/8	no	4/9	4/8	no	4/9	5/8
Porridge	yes	no	Soccer	yes	no			
yes	6	4	yes	4	5			
no	3	4	no	5	3			
yes	6/9	4/8	yes	4/9	5/8			
no	3/9	4/8	no	5/9	3/8			