Data Mining - Homework 1

Dmitry Donetskov

April 25, 2018

1 Warm-up Exercise

The expected value of lottery prize (denoted as X) for our case is

$$E[X] = \sum_{\{\$5,\$10\}} x p_X(x) = \$5 \times 0.8 + \$10 \times 0.2 = \$6$$

2 Chomsky Disease

Positive means a hampster has the disease. Let's introduce few notations and formalize the conditions:

- y: the true class coded as **0** for 'a hampster does not have the disease' and **1** for 'a hampster has the disease',
- $\hat{y}$: the estimated class for a case,
- N_+: the number of true positives i.e. $\#\{\hat{y} = 1 | y = 1\}$,
- N_-: the number of true negatives i.e. $\#\{\hat{y} = 0 | y = 0\}$,
- the misclassification cost l is :
  - the cost of false negative (FN): $l(\hat{y} = 0 | y = 1) = \$1000$,
  - the cost of false positive (FP): $l(\hat{y} = 1 | y = 0) = \$600$.

2.1 Questions 1-3

The false positive rate (FPR, also known as specificity) is defined as $^{FP}/_{N_-}$. The true positive rate (TPR, also known as sensitivity) is defined as $^{TP}/_{N_+}$. The false negative rate (FNR) is defined as $^{FN}/_{N_+}$ and is equal to 1-TPR:

$$FNR = \frac{FN}{N_+} = \frac{N_+ - TP}{N_+} = \frac{N_+}{N_+} - \frac{TP}{N_+} = 1 - \frac{TP}{N_+} = 1 - TPR$$

Consequently, the values of FPR, TPR, FNR associated with the three points on the ROC curve are

ROC Point	FPR	TPR	FNR
A	0.1	0.2	0.8
B	0.3	0.8	0.2
C	0.5	0.9	0.1

2.2 Question 4

FPR: the mistake that Sara will treat a healthy hamster is associted with the *false positive rate* (FPR). **FNR**: the mistake that Sara will not treat an ill hamster is associted with the *false negative rate* (FNR).

2.3 Question 5

The total cost of mistakes C for our problema consists of two parts which we need to sum up:

- the cost connected with FPR (the type I error),
- the cost connected with FNR (the type II error).

To put it with the formula

$$E[C] = l(\hat{y} = 1|y = 0) \times P\{\hat{y} = 1|y = 0\} + l(\hat{y} = 0|y = 1) \times P\{\hat{y} = 0|y = 1\}$$

According to the formula of expectation, the expected cost of mistakes for each point on the ROC curve is

- The point A: $\$600 \times 0.1 + \$1000 \times 0.8 = \$60 + \$800 = \$860$
- The point B: $\$600 \times 0.3 + \$1000 \times 0.2 = \$180 + \$200 = \$380$
- The point C: $\$600 \times 0.5 + \$1000 \times 0.1 = \$300 + \$100 = \$400$

2.4 Question 6

The AUC has a useful property [1]:

> The AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

So, to answer the question we need to calculate AUC for the classifier used by Sara and subtract it from 1 as we are insterested in that a randomly chosen negative instance will be ranked higher than a randomly chosen positive instance. The AUC due to the ROC curve being a simple one can be manually calculated as $0.9 + 0.05/2 - 0.9 * 0.2 + 0.01 = 0.755$.

The probability Sara will pick up a **wrong** one is $1 - 0.755 = 0.245$.

3 Heart Rate

3.1 Comparing Two Classifing Algorythms

The two classifiers were called from Python by using the documented API with the CV evaluation (k = 10). The *data* variable is Orange.data.Table with the content of heart_disease.tab.

The scoring shows the Naive Bayes classifier is better; classification accuracy is a) Naive Bayes: 0.835, b) Tree: 0.739.

```
learner_bayes = Orange.classification.NaiveBayesLearner()
learner_tree  = Orange.classification.TreeLearner()

cv_res = Orange.evaluation.CrossValidation(data, [learner_bayes, learner_tree], k = 10)

print("Classification␣Accuracy␣(CV,␣k␣=␣10):")
print("Naive␣Bayes:␣%.3f" % round(Orange.evaluation.scoring.CA(cv_res)[0], 3))
print("Tree:␣␣␣␣␣␣␣␣%.3f" % round(Orange.evaluation.scoring.CA(cv_res)[1], 3))
```

3.2 Naive Bayes: Costs of Sending N = {50, 100} People

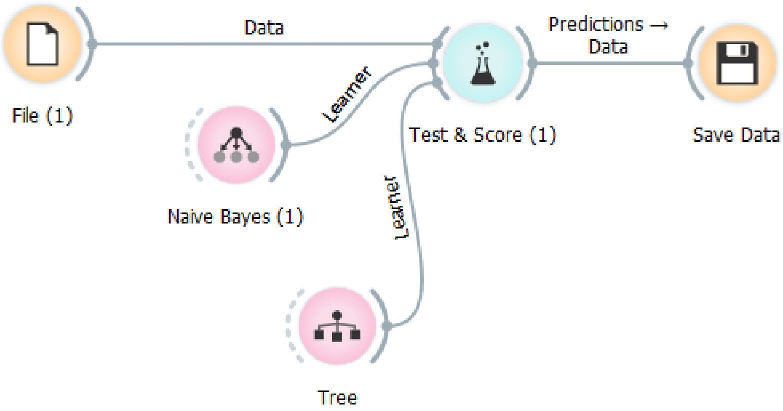
The Naive Bayes classifier's estimation for a person to be sick or not to be sick is used with the default threshold of 0.5. The classifier was called as an instance of Orange.classification.NaiveBayesLearner() with target_class set to 1.

For the first 50 people who are the most likely to be sick, there is one person who is not sick. There are no other mistakes, the total cost of mistake is $500.

For the first 100 people who are the most likely to be sick, there are nine person who are not sick. There are no other mistakes. The total cost of mistakes is $4500

3.3 Two Classifiers: Costs of Sending N People

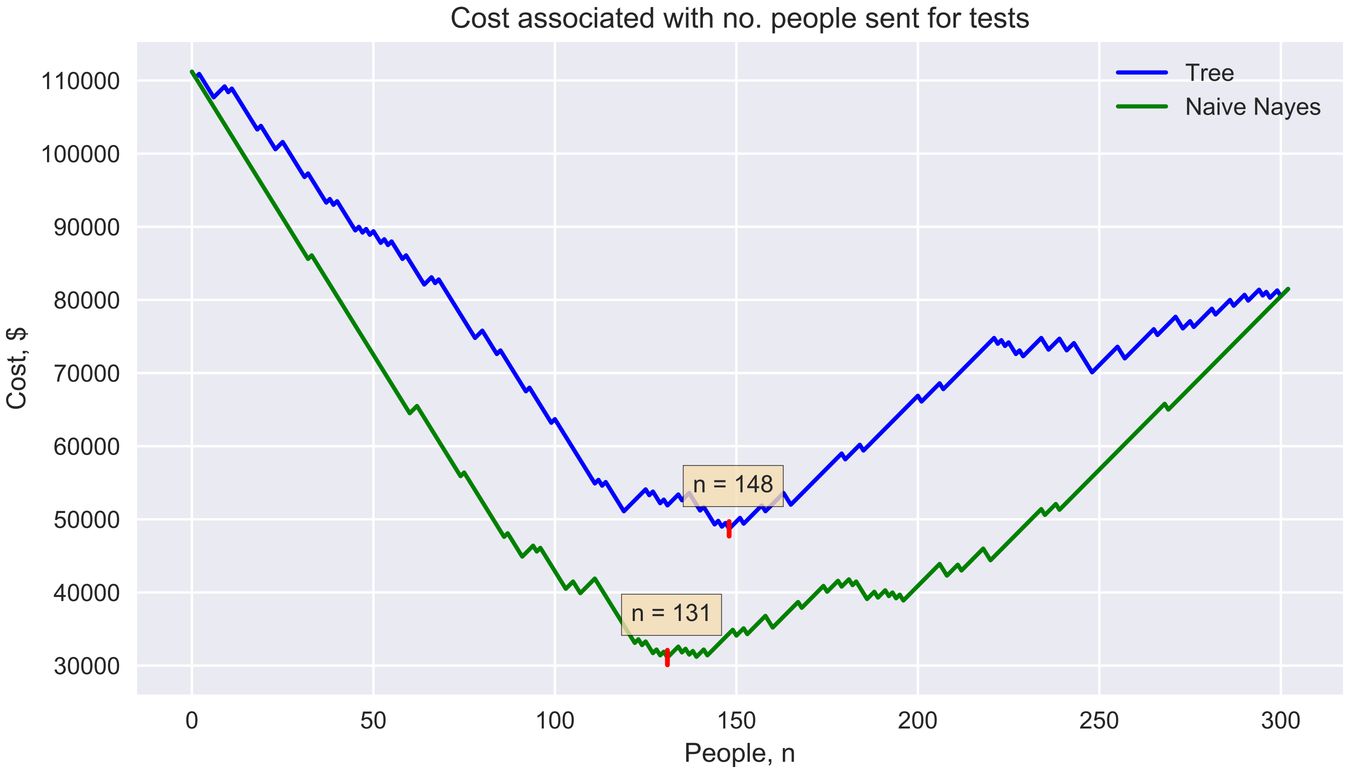
The Test&Score widget has been used to get the scoring for both the classifiers with the following workflow:



To find out the optimal number of patients sent to tests we evaluate both the classifiers as follows

- sort the Test&Score results for Tree in the descending order of being positive (by 'Tree (1)'),
- sort the Test&Score results for Naive Bayes in the descending order being positive (by 'Naive Bayes (1)')
- for n running from 1 to N (the sample size)
  - take the n first people as positive and the rest as negative,
  - calculate the cost of FP and FN.

The optimal number of people to send to tests are shown on the graph below. In terms of choosing the classifier between these two, we may conclude that Naive Bayes works better, we'll lose less money with it.



References

[1] Tom Fawcett. An introduction to ROC analysis. *Pattern Recogn. Lett.*, 27(8):861–874, June 2006.