coursework9

Fortunat Mutunda

April 10, 2016

1. Read the article by Domingos: A few useful things to know about machine learning (communications of the ACM, Vol. 55 No. 10, Pages 78-87 doi: 10.1145/2347736.2347755 via ACM Digital library, https://courses.cs.ut.ee/MTAT.03.183/2012_fall/uploads/Main/domingos.pdf). Make a list of key messages with a supporting 1-2 sentence example or clarification of that message (something like short summary of the article)

- Representation + Evaluation + optimization

To me the most important message in this article is:

Combination of performance evaluation and optimization are the mother of learning. Including three consist of classifiers with existing spatial principles, optimization framework that has almost the prediction process, and evaluation distinguishes good bad classifiers.

Generalization: When data are important in the majority of cases, but usually a precise measurement as a function of different classifiers. More new data classifiers of defects which are considered might be satisfactory.

Learning about a problem, and it may be in the form of bias or variance. Cross-validation can help fight overfitting but "no specific method is actually the ultimate solution for more customized" in fact, such as multi test could lead to under-fitting It's the curse of dimensionality USA Blessing of non-uniformity. The first shows how high dimensionality brings good layout problems, while a little more room to explore. technical unit for learning the characteristics of the required specific area .... combination or separately consider two different causes for observation, features sometimes could be better meaning and usefulness combined individually.

More Details vs algorithm: there are more good opportunities to learn when there is more data but the downside of the required time and resources in this century, it is difficult to make good use of large amounts of data.

Many models at a model: the new systems include multiple models instead of sampling the best techniques after examining the various models; so it is very important for many learning models, and to apply, if necessary. Here, bags and stacking methods push through the steps. Correlation and causality was discussed when he does not necessarily have to be as the conclusion which could be considered drawn from the old, even if it is a kind of niche in machine learning and data mining.

2. Draw the ROC curves and calculate the ROC AUC for 4 classifiers based on the following data Attach:roc_data.zip. The data.class is the true class, and the roc1.txt etc are the orders in which different classifiers would classify examples as positive (so some are true positive, some false positive; after certain cutoff there remain false negatives and true negatives).

Hint: you can choose top k % of the ordered observations, "classify" them as positives and calculate the corresponding TPR and FPR. Repeat it for k from 0 to 100 with some step, and these pairs of (TPR, FPR) values will produce a ROC curve. k is known as cutoff point. You can read a nice explanation about ROC curves from here.

3. Characterize the behavior of the 4 classifiers in task 2. Also, provide the "best" cutoff for each of the classifiers.

Hint: If cutoff k=0%, everything is simply predicted as negative class, so each truly positive label is (falsely!) classified as negative, i.e. TPR=0% and each truly negative label is also classified as negative, i.e FPR=0%. As we increase k, we recover more positive samples, i.e. TPR increases, but so does FPR (luckily, at a lower rate, usually!). In the extreme case when k= 100%, everything is predicted positive, so TPR=100%, but FPR will also be 100%. Obviously, we would prefer a classifier with the highest possible TPR and lowest FPR. You can take a difference between them (TPR-FPR), called Youden's index, to find the "best" cutoff point

4. Use the data about housing (http://archive.ics.uci.edu/ml/datasets/Housing) and estimate by regression analysis the last column - report RMSE score.

For this task I used Weka to calculate the RMSE score and this is the result. Because we have more than one value I decided to use simple linear regression to get the value of RMSE as it can be seen down there. And the value is 4.8845

```
=== Run information ===

Scheme:weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8

Relation: housing_data

Instances: 506

Attributes: 14

          CRIM

          ZN

          INDUS

          CHAS

          NOX

          RM

          AGE

          DIS

          RAD

          TAX

          PTRATIO

          B

          LSTAT

          MEDV

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Linear Regression Model

MEDV =
  0.0363 * ZN +

  2.5254 * CHAS +

-12.0975 * NOX +

  4.9887 * RM +

 -1.2719 * DIS +

  0.2223 * RAD +

 -0.0101 * TAX +

 -0.8204 * PTRATIO +

  0.0126 * B +

 -0.5378 * LSTAT +

 21.6793
```

Time taken to build model: 0 seconds

=== Cross-validation === === Summary ===

Correlation coefficient 0.8325

Mean absolute error 3.2424

Root mean squared error 4.8845

Relative absolute error 50.4847 %

Root relative squared error 55.3057 %

Total Number of Instances 452

Ignored Class Unknown Instances 54

5. Estimate every variable one by one using all other attributes in this data set - report RMSE scores for each. What are the most important predictors and what are the most correlated ones? Which variables are "easier" to predict than others? If so, then why?

CRIM ***Root mean squared error 1.2864***

ZN ***Root mean squared error 18.2063***

INDUS ***Root mean squared error 3.7304***

CHAS ***Root mean squared error 0.2486***

NOX ***Root mean squared error 0.272***

Age ***Root mean squared error 16.9941***

DIS ***Root mean squared error 1.0691***

RAD ***Root mean squared error 17.397***

RM ***Root mean squared error 3.2401***

TAX ***Root mean squared error 72.7659***

PTRATION ***Root mean squared error 54.1824***

B ***Root mean squared error 59.1702***

LSTAT ***Root mean squared error 3.8814***

For this task too I had task I also have used weka but the output of everything was too long so I have decided just to use this output instead of everything.

6. (Bonus 2p) Continue the task from ROC examples. Assume there is different cost assigned for different types of mistakes. E.g. cost 20€ for missing a case (false negative) and 15€ for false classification (false positive). Or vice versa. Calculate for each of the 4 classifiers with ROC curve, what would be the optimal cutoff to minimize cost. Provide yourself examples of four such costs based on which you can say for each of the 4 classifiers that exactly that provides the best classification.