Coursera - Practical Machine Learning - Quiz1

*Jean-Luc BELLIER*

*04 janvier 2017*

Question 1

Which of the following are components in building a machine learning algorithm?

* Artificial intelligence
* Machine learning
* Training and test sets
* Creating features.
* Statistical inference

***Answer :*** The ***Creating features*** are components in building a machine learning algorithm.

Question 2

Suppose we build a prediction algorithm on a data set and it is 100% accurate on that data set. Why might the algorithm not work well if we collect a new data set?

* We have too few predictors to get good out of sample accuracy.
* Our algorithm may be overfitting the training data, predicting both the signal and the noise.
* We may be using a bad algorithm that doesn’t predict well on this kind of data.
* We have used neural networks which has notoriously bad performance.

***Answer :*** ***Our algorithm may be overfitting the training data, predicting both the signal and the noise***.

Question 3

What are typical sizes for the training and test sets?

* 90% training set, 10% test set
* 0% training set, 100% test set.
* 50% in the training set, 50% in the testing set.
* 80% training set, 20% test set

***Answer :*** ***80% training set, 20% test set***.

Question 4

What are some common error rates for predicting binary variables (i.e. variables with two possible values like yes/no, disease/normal, clicked/didn’t click)? Check the correct answer(s).

* Predictive value of a positive
* Correlation
* Root mean squared error
* Median absolute deviation
* R^2

***Answer :*** ***Predictive value of a positive***.

Question 5

Suppose that we have created a machine learning algorithm that predicts whether a link will be clicked with 99% sensitivity and 99% specificity. The rate the link is clicked is 1/1000 of visits to a website. If we predict the link will be clicked on a specific visit, what is the probability it will actually be clicked?

* 99%
* 99.9%
* 50%
* 9%

***Answer :*** By definition we have :

sensivity=TPTP+FNsensivity=TPTP+FN

specificity=TNTN+FPspecificity=TNTN+FP

prevalence=TP+FNTP+FN+TN+FPprevalence=TP+FNTP+FN+TN+FP

and we know that :

TP=(TP+FN).sensitivityTP=(TP+FN).sensitivity, FP=(TN+FP).(1−specificity)FP=(TN+FP).(1−specificity)

sensitivity.prevalence=TPTP+FN+TN+FPsensitivity.prevalence=TPTP+FN+TN+FP

(1−specificity).(1−prevalence)=FPTP+FN+TN+FP(1−specificity).(1−prevalence)=FPTP+FN+TN+FP

We want to compute : p = Pr(click +|test click +) = TPTP+FPTPTP+FP

p=specificity.prevalencesensitivity.prevalence+(1−specificity).(1−prevalence)p=specificity.prevalencesensitivity.prevalence+(1−specificity).(1−prevalence)

So p=10−3.0.9910−3.0.99+0.01∗0.999p=10−3.0.9910−3.0.99+0.01∗0.999 ~ 9%.