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 :

$$sensivity = \frac{TP}{TP + FN}$$

$$specificity = \frac{TN}{TN + FP}$$

$$prevalence = \frac{TP + FN}{TP + FN + TN + FP}$$

and we know that :

$TP = (TP + FN).sensitivity$, $FP = (TN + FP).(1 - specificity)$

$$sensitivity.prevalence = \frac{TP}{TP + FN + TN + FP}$$

$$(1 - specificity).(1 - prevalence) = \frac{FP}{TP + FN + TN + FP}$$

We want to compute : p = Pr(click +|test click +) = $\frac{TP}{TP+FP}$

$$p = \frac{specificity.prevalence}{sensitivity.prevalence + (1 - specificity).(1 - prevalence)}$$

So $p = \frac{10^{-3}.0.99}{10^{-3}.0.99+0.01*0.999} \sim 9\%$.