Quiz 1

mgestal

Thursday, May 28, 2015

Question 1

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

Solution

- Data mining
- Artificial intelligence
- Training and test sets
- **Asking the right question.**

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?

Solution

- **Our algorithm may be overfitting the training data, predicting both the signal and the noise.**
- We are not asking a relevant question that can be answered with machine learning.
- We have too few predictors to get good out of sample accuracy.
- We may be using bad variables that don't explain the outcome.v

Question 3

What are typical sizes for the training and test sets?

Solution

- 80% training set, 20% test set
- 100% training set, 0% test set.
- 10% test set, 90% training set
- **60% in the training set, 40% in the testing 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)?

Solution

- **Sensitivity**
- Correlation
- Root mean squared error
- R^2

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?

Solution

- **9%**

```
# Positive <=> clicked

# Rate link is clicked = prevalance = 1/1000 = 0.1%
# We can assume that there are 100000 visits. Prevalence is 0,1% ==> 100 clicks
# sensitivity = 99%, so TP = 99 and FN = 1
TP <- 99
FN <- 1

# N=100000 => FP+TN= N-100 = 99900
# specificity = 99%, so TN = 0.99 * 99000 = 98901
TN <- 0.99 * 99900
FP <- 99900 - TN
FP
```

```
## [1] 999
```

```
# Question relies about positive predictive value (PPV) ==> probability to get
# a click in a specific visit, if the predicted outcome is positive.  PPV = TP/(TP+FP)
PPV <- TP / (TP + FP)
PPV
```

```
## [1] 0.09016393
```