Practical Machine Learning Overview

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

Practical Machine Learning Content

- Prediction study design
- Types of Errors
- Cross validation
- The caret package
- Plotting for prediction
- Preprocessing
- Predicting with regression
- Predicting with trees
- Boosting
- Bagging
- Model blending
- Forecasting

Basic terms

In general, **Positive** = identified and **negative** = rejected. Therefore:

- **True positive** = correctly identified
- **False positive** = incorrectly identified
- **True negative** = correctly rejected
- **False negative** = incorrectly rejected

Medical testing example:

- **True positive** = Sick people correctly diagnosed as sick
- **False positive**= Healthy people incorrectly identified as sick
- **True negative** = Healthy people correctly identified as healthy
- **False negative** = Sick people incorrectly identified as healthy.

http://en.wikipedia.org/wiki/Sensitivity_and_specificity

Correlated predictors

```
library(caret)
library(kernlab)
data(spam)
inTrain <- createDataPartition(y = spam$type, p = 0.75, list = FALSE)
training <- spam[inTrain, ]
testing <- spam[-inTrain, ]

M <- abs(cor(training[, -58]))
diag(M) <- 0
which(M > 0.8, arr.ind = T)
```

```
##        row col
## num415  34  32
## direct  40  32
## num857  32  34
## num857  32  40
```

Basic idea behind boosting

1. Start with a set of classifiers $h_1, \ldots, h_k$
   - Examples: All possible trees, all possible regression models, all possible cutoffs.
2. Create a classifier that combines classification functions: $f(x) = \text{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$.
   - Goal is to minimize error (on training set)
   - Iterative, select one h at each step
   - Calculate weights based on errors
   - Upweight missed classifications and select next h

Adaboost on Wikipedia

http://webee.technion.ac.il/people/rmeir/BoostingTutorial.pdf