# Machine Learning - Homework 5

Parita Pooj (psp2133)

November 28, 2016

## Problem 1

1. Classifiers: Boosted decision tree, Neural Networks

2. Citations included in bibliography [1-3]

3. For neural networks, training with single level network is faster and better than multi-level network. For multi-level network with too many nodes, the network tends to overfit. Thus, a single hidden layer with about 60-80 nodes gives the best error rate for the neural network.
   For boosted decision tree, training with a shallow tree of depth less than 5 or 6 gives high error rates. Increasing the depth, the decision tree gives reasonable error rates. On an average, about 100 estimators work the best empirically.
   The evaluation was done based on cross-validation error rates.

4. The final hyperparameters for 1/5th training data used:
   **Neural Network**
   Hidden Layers: 1
   Nodes: 70

   **Boosted Decision Tree**
   Depth: 8, Estimators: 125

5. The error rates are as below:
   **Neural Network**
   Training error rate: 10.03 %
   Cross-validation error rates: 11.68 %, 11.25 %, 11.98 %, 11.03 %, 11.76 %
   Test error rate: 11.37 %

   **Boosted Decision Tree**

Training error rate: 14.28 %
Cross-validation error rates: 14.91 %, 15.89 %, 14.87 %, 14.24 %, 15.34 %
Test error rate: 15.03 %

# Problem 2

(a) $\frac{1}{2} \ln \frac{\eta}{1-\eta}$

(b) MAE = 0.0390

# Problem 3

(a)

1. True
   Justification: Since, the distribution for $S$ is specified as $P_{(w_*, \sigma_*^2)}$ where $\sigma_*^2$ is the variance.

2. False
   Justification: Since, $E(w_*|A) = w_*$, the covariance will be 0.

3. True
   Justification:

$$cov(w_*|A) = E((w_* - E(w_*|A))(w_* - E(w_*|A))^T|A)$$

$$cov(w_*|A) = E((w_* - w_{ols})(w_* - w_{ols})^T|A)$$

$$A^T A cov(w_*|A) = E((Aw_* - Aw_{ols})(Aw_* - Aw_{ols})^T|A)$$

$$A^T A cov(w_*|A) = E((E(y|A) - \hat{y})(E(y|A) - \hat{y})^T|A)$$

$$cov(w_*|A) = \sigma_*^2 (A^T A)^{-1}$$

4. True
   Justification:
   $E(\hat{y}|A) = E(A\hat{w_{ols}}|A) = AE(\hat{w_{ols}}|A) = Aw_*$

5. True
   Since, $Aw_{ols}$ is the orthogonal projection of $y$ in the span of A. We know that subtracting the orthogonal projection from the vector is the orthogonal projection of the vector in the transpose of its null space.

6. True
   Since 5. is true, we can write:
   $A^T r = 0$
   Since, the first column of A is 1, the first row of $A^T$ will be 1.
   Let this first column be $a_1$
   Therefore, $a_1^T r = 0$
   $\sum_{i=1}^{n} r_i = 0$

(b)
1. Test averaged square loss of OLS estimator: 24.4066
2. Test averaged square loss of the sparse linear predictor (OMP): 36.0452
   Other methods tried: Lasso, LARS, CoSaMP - all of which give a higher error rate
3. Non-zero sparse entries observed for - CHAS, PTRATIO, LSTAT
4. Toolbox from `https://www.mathworks.com/matlabcentral/fileexchange/32402-cosamp-and-omp-for-sparse-recovery`

# References


[1] `http://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html#sklearn.neural_network.MLPClassifier`

[2] `http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html#sklearn.ensemble.AdaBoostRegressor`

[3] `http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html#sklearn.tree.DecisionTreeRegressor`

[4] `Toolboxfromhttps://www.mathworks.com/matlabcentral/fileexchange/32402-cosamp-and-omp-for-sparse-recovery`