Machine Learning Exercise on LDA

In this assignment you will estimate cognitive states from electroencephalogram (EEG) data.

Data Electroencephalography (EEG) data was recorded during a copy-spelling BCI Experiment at the Berlin BCI group (`http://www.bbci.de`).

The data set consists preprocessed EEG data $X \in \mathbb{R}^{5\times62\times5322}$ and stimulus labels $Y \in \mathbb{R}^{2\times5322}$ during a copy-spelling paradigm with a P300 speller. The data matrix X contains 5 selected time windows of EEG activity at 62 electrodes after a visual stimulus was presented on the screen in front of the subject. If the first row of Y is 1, the stimulus was a target stimulus, if the second row of Y is 1, the stimulus was a non-target stimulus.

1. How well can we predict the cognitive state from the EEG data?
   a) Implement a nearest centroid classifier which predicts +1 if

$$\|\mu_+ - x\|_2 < \|\mu_- - x\|_2 \tag{1}$$

   where $\mu_{+/-}$ are the means of the positive (respectively negative) class, that is $\mu_+ = 1/N_+ \sum_i^{N_+} x_{+i}$ and x_{+i} is the ith data point of the + class.

   b) Implement a linear discriminant analysis (LDA) classifier; optimize a vector $w \in \mathbb{R}^{5\cdot62\times1}$ such that

$$\underset{w}{\operatorname{argmax}} \frac{w^\top S_B w}{w^\top S_W w} \tag{2}$$

   where

$$S_B = (\mu_+ - \mu_-)(\mu_+ - \mu_-)^\top \tag{3}$$

$$S_W = 1/N_+ \sum_i (x_{+i} - \mu_+)(x_{+i} - \mu_+)^\top + 1/N_- \sum_j (x_{-j} - \mu_-)(x_{-j} - \mu_-)^\top \tag{4}$$

   Train both classifiers on the $5\cdot62 = 310$ concatenated (across time and electrodes) features to predict target stimuli from the EEG data. Train both classifiers on 70% of the data and test it on 30% of the data. Select the data points for training and test randomly, but make sure they do not overlap!

2. Compare the prediction accuracy for each classifier. For both classifiers: plot histogram of classifier outputs for targets/non-targets using pylab.hist (see fig. 1)

3. Estimate the **generalization performance** of each classifier. Test each classifier on the training set data and on the test data data. Compare the prediction accuracies (see fig.2)

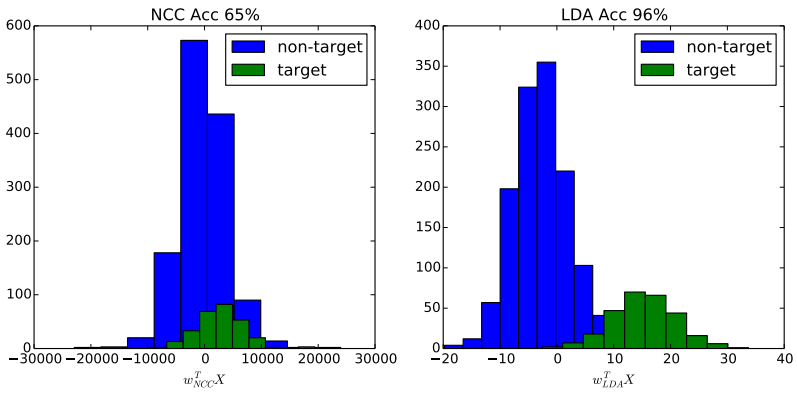

Figure 1: *Left*: Classifier outputs for nearest-neighbor classifier. *Right*: Classifier outputs for linear discriminant classifier.

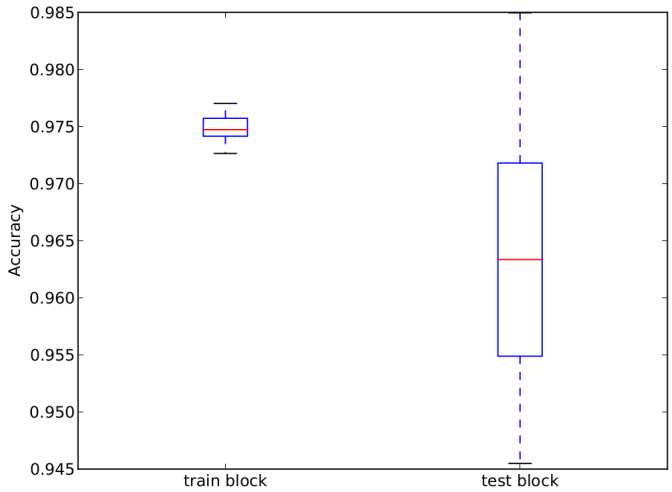

Figure 2: Classification accuracy for training and test blocks in 10-fold cross-validation.