Tutorial on Nested Cross-Validation

Like in the last assignment you will estimate cognitive states from electroencephalogram (EEG) data. This time we will look at cases where we do not have enough data to estimate covariance matrices properly and we will also use probabilistic classifiers to produce uncertainty estimates of our outputs.

Regularized Linear Discriminant Analysis In many real world settings you do not have enough data to compute good estimates of covariance matrices $\mathbf{S}$. This means for elliptical distributions that the large eigenvalues (i.e. the elongation of the ellipsoid along the longer axes) are overestimated while the small eigenvalues (i.e. elongation of the ellipsoids along the shorter axes) are underestimated. We can counteract this systematic bias by **shrinking** $\mathbf{S}$ towards the identity matrix $\mathbf{I}$ by computing a new covariance matrix $\tilde{\mathbf{S}}$ as a convex combination of the two

$$\tilde{\mathbf{S}} = (1-\gamma)\mathbf{S} + \gamma\nu\mathbf{I} \tag{1}$$

where ν is the average eigenvalue of $\mathbf{S}$ and the regularization or shrinkage parameter γ determines the amount of shrinkage (i.e. how much we distrust our covariance estimate $\mathbf{S}$). A straightforward method to optimize γ is to do **nested cross-validation**. For example, for 10-fold nested CV you split the data into 10 non-overlapping folds (the *outer-crossvalidation loop*). Then you specify a number of γ parameters that you want to try. For each of the 10 splits you do *model selection* for γ by testing on the 9 training data blocks how well our LDA classifier can generalize for each γ. The generalization performance is tested using 9-fold crossvalidation within the training folds of the outer crossvalidation loop. This crossvalidation is the *inner crossvalidation loop*. After each round of inner CV you choose the γ that generalizes best on average, train again on all of the (outer CV) training data and test on the (outer CV) test data. The peudocode for nested cross-validation is given in algorithm 1.
Pick the first 500 data points in the data set of assignment 3 and perform nested CV using the 5 γ parameters $[0., 0.0005, 0.005, 0.05, 0.5, 1]$. For each of the inner loops, store the average generalization performance for each γ. Afterwards, plot the grand average generalization performance for each of the different γ, as illustrated in 1.

Algorithm 1 Cross-Validation for Model Selection and Evaluation

Require: Data $(\mathbf{x}_1, y_1)\ldots,(\mathbf{x}_N, y_N)$, parameters $\gamma_1,\ldots,\gamma_S$, Number of CV folds F

\# Split data in F **disjunct** folds
for Outer folds $f_{\text{outer}} = 1,\ldots,F$ **do**
    \# Pick folds $\{1,\ldots,F\} \setminus f_{\text{outer}}$ for Model Selection
    \# **Model Selection**
    for Fold $f_{\text{inner}} = 1,\ldots,F-1$ **do**
        for Parameter $s = 1,\ldots,S$ **do**
            \# Train model on folds $\{1,\ldots,F\} \setminus \{f_{\text{outer}}, f_{\text{inner}}\}$ with parameter γ_s
            \# Compute prediction on fold f_{inner}
        end for
    end for
    \# Pick best parameter γ_s for all f_{inner}
    \# **Model Evaluation**
    \# Train model on folds $\{1,\ldots,F\} \setminus f_{\text{outer}}$ with parameter γ_s
    \# Test model on fold f_{outer}
end for

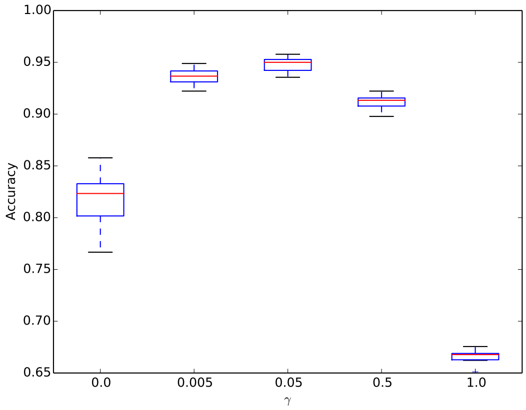

Figure 1: Generalization performance for a number of different shrinkage parameters γ when training a regularized LDA classifier on 500 data points.