605.449 — Introduction to Machine Learning

Correction and Explanation of the Fisher Score

In project 2, you are being asked to evaluate the clusters for feature selection using a variation of Fisher's LDA score. In fact, the function provided is incomplete. Here is the full, corrected function:

$$\mathcal{L} = \frac{\mathbf{w}^\top \left(\sum_{i=1}^{k} (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^\top \right) \mathbf{w}}{k \mathbf{w}^\top \left(\sum_{i=1}^{k} \Sigma_i \right) \mathbf{w}}$$

This requires some clarification, especially where the equation changed.

- First I want to call your attention to the change from $(\mathbf{m}_i - \mathbf{m}_j)^2$ to $(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^\top$. This was the first main error in the original equation. In a discussion post, I said that the squaring of the difference in the mean vectors was a component-wise squaring. Wrong. It is actually an outer product of the vectors and yields a matrix. But now that we have a matrix, we have some work to do.

- Second, notice that I removed one of the sums and replaced $\mathbf{m}_j$ with just $\mathbf{m}$. Here, rather than doing pairwise differences, we are just going to compare each cluster mean vector with the mean of the mean vectors. So for this,

  $$\mathbf{m} = \frac{1}{k} \sum_{i=1}^{k} \mathbf{m}_i.$$

  And yes, here we are doing component-wise means.

- Third, notice the denominator has changed. Well in reality it hasn't. I am just using more standard notation. For this, each Σ_i is the covariance of the respective cluster. Recall that you calculate the covariance matrix as follows. For each feature pair x_i and x_j, compute the following over all of the data points assigned to the cluster:

  $$\Sigma_c(x_i, x_j) = \frac{1}{n_c} \sum_{d=1}^{n_c} (x_i(d) - m_{c,i})(x_j(d) - m_{c,j})$$

  where $m_{c,i}$ is the ith component of the mean vector for cluster c and n_c is the number of examples assigned to cluster c.

- Fourth, notice I have added weight vectors $\mathbf{w}$ into the function. That's because it is the feature set that is used to guide the clustering, and in LDA the weight vector was essentially your feature set weighted by importance. So for this, $\mathbf{w}$ is just a vector of ones and zeros corresponding to the original features in your data set. If a component in the vector gets a one, then that feature is being used. If it gets a zero, then it is not being used. Thus you set $\mathbf{w}$ with the feature selection part of the algorithm.

- Fifth, notice we are now dividing by k. This probably is not all that important (since it is a constant), but what it does is compute a mean between scatter value for all of the clusters.

- Finally, notice what happens. Because we have a vector times a matrix times a vector in both the numerator and the denominator, we now have a scalar divided by a scalar. So we no longer need to worry about finding the length of the resulting vector because there is no vector.

When considering the impact of this change on your project, it is my hope that you all have everything working (except for having a correct evaluation function) and that all you would need to do is replace the evaluation function and re-run the experiments. Basic SFS, k-means, and HAC are fast, but the GA selection is not. Therefore, I am granting two additional days to complete the assignment.