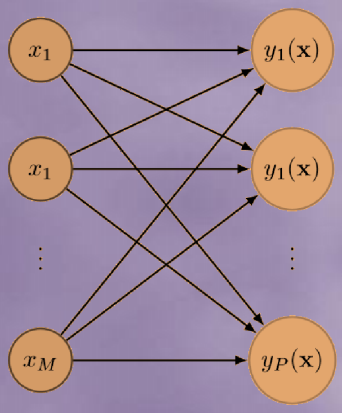

Discriminant Analysis

Daniel Khashabi[1]
KHASHAB2@ILLINOIS.EDU

0.1 Introduction

Discriminant Analysis methods are among methods for linear classification or the dimensionality reduction before classification method. In LDA we assume that conditional probability densities, $f(x|Y = y)$ are normally distributed. Let's assume that for classification posterior distribution we have the followings:

$$\Pr(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l},$$

where π_ks are class membership priors. These priors show the probablity of membership in each of the classes, without having any training data at hand. Assuming multivariate Guassian distribution,

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\left|\mathbf{\Sigma}_k\right|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\mu_k)}$$

The decision rule for deciding what class the data lies in is simply finding the maximum value of the class membership postorior:

$$G(x) = \arg\max_k \Pr\left(G = k|X = x\right). \tag{1}$$

In LDA we assume that all classes have the same covariance matrix $\mathbf{\Sigma}_k = \mathbf{\Sigma}_k,\ \forall k$. It can easily shown that one can write the following expression about the log-ratio of two classes,

$$\log\frac{\Pr\left(G = k|X = x\right)}{\Pr\left(G = l|X = x\right)} = \log\frac{\pi_k}{\pi_l} - \frac{1}{2}\left(\mu_k + \mu_l\right)^T \mathbf{\Sigma}^{-1}\left(\mu_k - \mu_l\right) + \mathbf{x}^T\mathbf{\Sigma}^{-1}\left(\mu_k - \mu_l\right), \tag{2}$$

[1]This is part of my notes; to see the complete list of notes check web.engr.illinois.edu/ khashab2/learn.html.

which is the class margin, for classes l and k, when we have $\Pr(G = l|X = x) = \Pr(G = k|X = x)$. Based on the above equation we can easily see that the margin's equations are linear in terms of $\mathbf{x}$. That is why this method is called "Linear" Discriminant Analysis. By inspection it can be found that the decision rule in equation 1, can be equivalently written as the following equation:

$$G(x) = \arg\max_k \left\{ \mathbf{x}^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k + \log \pi_k \right\}.$$

Note that the approximate values of parameters for each class can be found using training data inside that class:

- Prior prbability of data, being in one class, can be estimated using the the number of data in that class over the number of the whole data: $\hat{\pi}_k = \frac{N_k}{N}$.
- Class distribution mean can be found by averaging over the data inside the class: $\mu_k = \frac{1}{N_k} \sum_{g_i=k} \mathbf{x}_k$.
- Class covariance can be found by calculating the covariance the data inside that class, but because be have assumed in the LDA that the covariance matrix of the all classes are the same, we average the covariance matrices over all the classes, which gives us the *pooled covariance matrix*: $\hat{\mathbf{\Sigma}} = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{g_i=k} (\mathbf{x}_i - \hat{\mu}_k)(\mathbf{x}_i - \hat{\mu}_k)^T$

The cased of *Regularized Discriminant Analysis*(RDA) is a compromise between LDA and QDA. In fact by performing an averaging on LDA and QDA's covariance matrices, we shrink the covariance of QDA towards LDA's. We can define the regularized covariance matrix as following:

$$\hat{\mathbf{\Sigma}}_k(\alpha) = \alpha \hat{\mathbf{\Sigma}}_k + (1 - \alpha) \hat{\mathbf{\Sigma}}_k, \ 0 \geq \alpha \geq 1, \tag{3}$$

in which the parameter α controls the shrinkage level, and RDA's inclination towards LDA or QDA, and $\hat{\mathbf{\Sigma}}_k$ is the *pooled covariance matrix*. Now the quadratic discriminant function can be defined using the shrunken covariance matrix $\hat{\mathbf{\Sigma}}$. The optimum value of α could be evaluated using cross-validation on test data. Another modification to the regularized version would be to shrink the covariance matrix towards an scalar matrix, which is a constant times an identity matrix:

$$\hat{\mathbf{\Sigma}}(\gamma) = \gamma \hat{\mathbf{\Sigma}} + (1 - \gamma) \hat{\sigma}^2 \mathbf{I}, \ 0 \leq \gamma \leq 1, \tag{4}$$

in which γ is a tuning parameter and could be determined using cross-validation on test data. To combine two regularization parameters, simply we replace $\hat{\mathbf{\Sigma}}$ in 3 with $\hat{\mathbf{\Sigma}}(\gamma)$, and name the result $\hat{\mathbf{\Sigma}}(\alpha, \gamma)$.