ECEN 649 Pattern Recognition

Parametric Classification Rules

Ulisses Braga-Neto

ECE Department

Texas A&M University

Basic Idea

- The basic idea of parametric classification is to assume that $\eta(x)$ (or the a-priori probabilities and class-conditional densities) belongs to a class of functions that depend on a parameter vector θ:

$$\eta(x) \in \{\eta(x;\theta) \mid \theta \in \Theta\}$$

- The assumption that $\eta(x)$ is known up to parameter θ is very demanding; what happens when the assumption is violated (as it will almost always be in practice) is the critical engineering question of *robustness*.
- Robust parametric classification rules (such as LDA, which we will discuss in this lecture) are quite useful in practice, even if the parametric assumption is violated.

Parametric Plug-In Rule

- The idea is to find an estimate of θ based on the data

$$\theta_n = \hat{\theta}(S_n)$$

and plug in the assumed form of $\eta(x)$, obtaining an approximation

$$\eta_n(x) = \eta(x; \theta_n)$$

- This yields the plug-in parametric rule

$$\psi_n(x) = \begin{cases} 1, & \eta(x; \theta_n) \geq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

Consistency of Parametric Rule

- Under the assumption that $\eta(x) = \eta(x;\theta^*)$, would consistency of $\hat{\theta}$ as an estimator of θ^*, that is,

$$\theta_n = \hat{\theta}(S_n) \rightarrow \theta^* \quad \text{as } n \rightarrow \infty ,$$

imply consistency of the plug-in classification rule?

- (DGL Theorem 16.1) If $\hat{\theta}$ is consistent and $\eta(x;\theta)$ is continuous in the L_1 sense at $\theta = \theta^*$, i.e.,

$$\theta_n \rightarrow \theta^* \Rightarrow E[|\eta(x;\theta_n) - \eta(x;\theta^*)|] \rightarrow 0$$

then the parametric rule is consistent.

Gaussian Discriminant Analysis

- The most important class of parametric classification rules is sometimes referred to as (Gaussian) *Discriminant Analysis*.
- Here it is assumed that the class-conditional densities are multivariate Gaussian (assuming two classes):

$$p(x|Y=i;\theta) = (2\pi)^{-\frac{d}{2}}|\Sigma_i|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(x-\mu_i\ )^T\Sigma_i^{-1}(x-\mu_i)\right)$$

  for $i = 0, 1$.
- The vector of parameters is given by:

$$\theta\ =\ (\mu_0, \Sigma_0, \mu_1, \Sigma_1, p)$$

  where $p = P(Y = 1)$.

Gaussian Discriminant Analysis - II

- The plug-in rule consists of estimating θ by $\hat{\theta} = (\hat{\mu}_0, \hat{\Sigma}_0, \hat{\mu}_1, \hat{\Sigma}_1, \hat{p})$, where

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n} X_j I_{Y_j=i} , \quad i = 0, 1.$$

are the *sample means*,

$$\hat{\Sigma}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n} (X_j - \hat{\mu}_i)(X_j - \hat{\mu}_i)^T I_{Y_j=i} , \quad i = 0, 1.$$

are the *sample covariance matrices*, and

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} I_{\{y_i=1\}} .$$

- To avoid poor small-sample performance and sensitivity to random sampling, we may set a fixed value $\hat{p} = 1/2$.

Quadratic Discriminant Analysis (QDA)

- The *plug-in discriminant function* is given by

$$g_n(x) = x^T A_n x + b_n^T x + c_n$$

where (assuming $\hat{p} = 1/2$)

$$A_n = -\frac{1}{2}\left(\hat{\Sigma}_1^{-1} - \hat{\Sigma}_0^{-1}\right)$$

$$b_n = \hat{\Sigma}_1^{-1}\hat{\mu}_1 - \hat{\Sigma}_0^{-1}\hat{\mu}_0$$

$$c_n = -\frac{1}{2}\left(\hat{\mu}_1^T\hat{\Sigma}_1^{-1}\hat{\mu}_1 - \hat{\mu}_0^T\hat{\Sigma}_0^{-1}\hat{\mu}_0\right) - \frac{1}{2}\ln\frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_0|}$$

- As we know, the corresponding decision surfaces $g_n(x) = 0$ are *hyperquadrics*. This is called Quadratic Discriminant Analysis (QDA).

Linear Discriminant Analysis (LDA)

- The drawback of QDA is the large number of parameters to be estimated. A simpler and very useful alternative is Linear Discriminant Analysis (LDA).
- Here we make the extra assumption that the covariance matrices are equal. This produces a hyperplane:

$$g_n(x) = a_n^T x + b_n = 0$$

where (assuming $\hat{p} = 1/2$)

$$a_n = \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)$$

$$b_n = -\frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1}(\hat{\mu}_1 + \hat{\mu}_0)$$

- The LDA discriminant $g_n(x)$ above is sometimes called *Anderson's Statistic* and denoted by $W(x)$.

Linear Discriminant Analysis (LDA) - II

- The single covariance matrix is estimated by the *pooled* sample covariance matrix:

$$\hat{\Sigma} = \frac{(n_0 - 1)\hat{\Sigma}_0 + (n_1 - 1)\hat{\Sigma}_1}{n_0 + n_1 - 2}$$

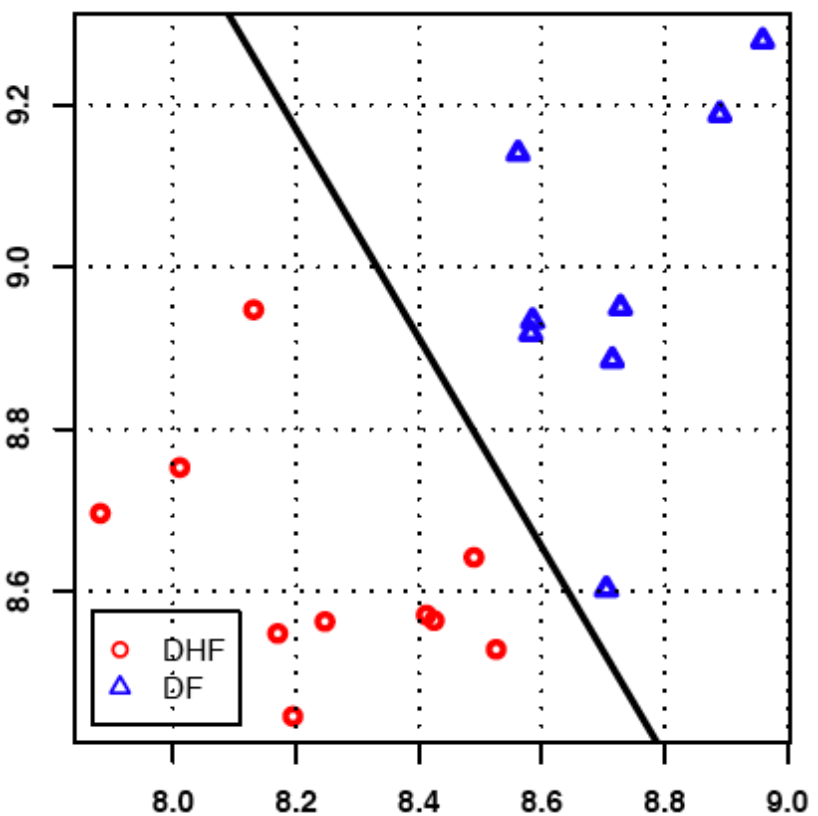
  Note that, if $n_0 = n_1$, then $\hat{\Sigma} = \frac{1}{2}(\hat{\Sigma}_0 + \hat{\Sigma}_1)$.

- For historical reasons, LDA is sometimes referred to as *Fisher's Discriminant Analysis*.

- LDA has shown to be robust to small deviations from Gaussianity. On the other hand, if the distributions are multi-modal or the covariance matrices are very dissimilar, then LDA will in general fail.

Linear Discriminant Analysis (LDA) - III

Example of LDA classifier for distinguishing benign dengue fever (DF) from dengue hemorrhagic fever (DHF) based on expression of two genes.



Diagonal LDA (DLDA)

- A special case of LDA is obtained by constraining the pooled sample covariance matrix $\hat{\Sigma}$ to be diagonal. In this case, one employs the sample variances $\hat{\sigma}_{i,j}^2$ of variable i in class j to estimate $\hat{\Sigma}$:

$$\hat{\Sigma}_{ii} = \frac{(n_0 - 1)\hat{\sigma}_{i,0}^2 + (n_1 - 1)\hat{\sigma}_{i,1}^2}{n_0 + n_1 - 2}$$

  with $\hat{\Sigma}_{ij} = 0$ for $i \neq j$. As before, if $n_0 = n_1$, then $\hat{\Sigma}_{ii} = \frac{1}{2}(\hat{\sigma}_{i,0}^2 + \hat{\sigma}_{i,1}^2)$.

- DLDA will generally perform better than plain LDA when the number of samples is very small.

Nearest Mean Classifier

- If we further assume that $\Sigma = \sigma^2 I_d$, and assume $\hat{p} = 1/2$, then the discriminant is a function of the mean vectors only:

$$g_n(x) = \frac{1}{2}\left(||x - \hat{\mu}_0||^2 - ||x - \hat{\mu}_1||^2\right)$$

  so that a new sample point x is classified to the nearest sample mean $\hat{\mu}_i$.

- This produces a hyperplane decision boundary that passes through the middle point between $\hat{\mu}_0$ and $\hat{\mu}_1$ and is perpendicular to $\hat{\mu}_1 - \hat{\mu}_0$.