ECEN 649 Pattern Recognition

Plug-In Classification Rules

Ulisses Braga-Neto

ECE Department

Texas A&M University

Plug-In Classification Rules

- The most intuitive idea for obtaining a classification rule is to try to approximate the (unknown) Bayes classifier somehow.
- The Bayes classifier is given by:

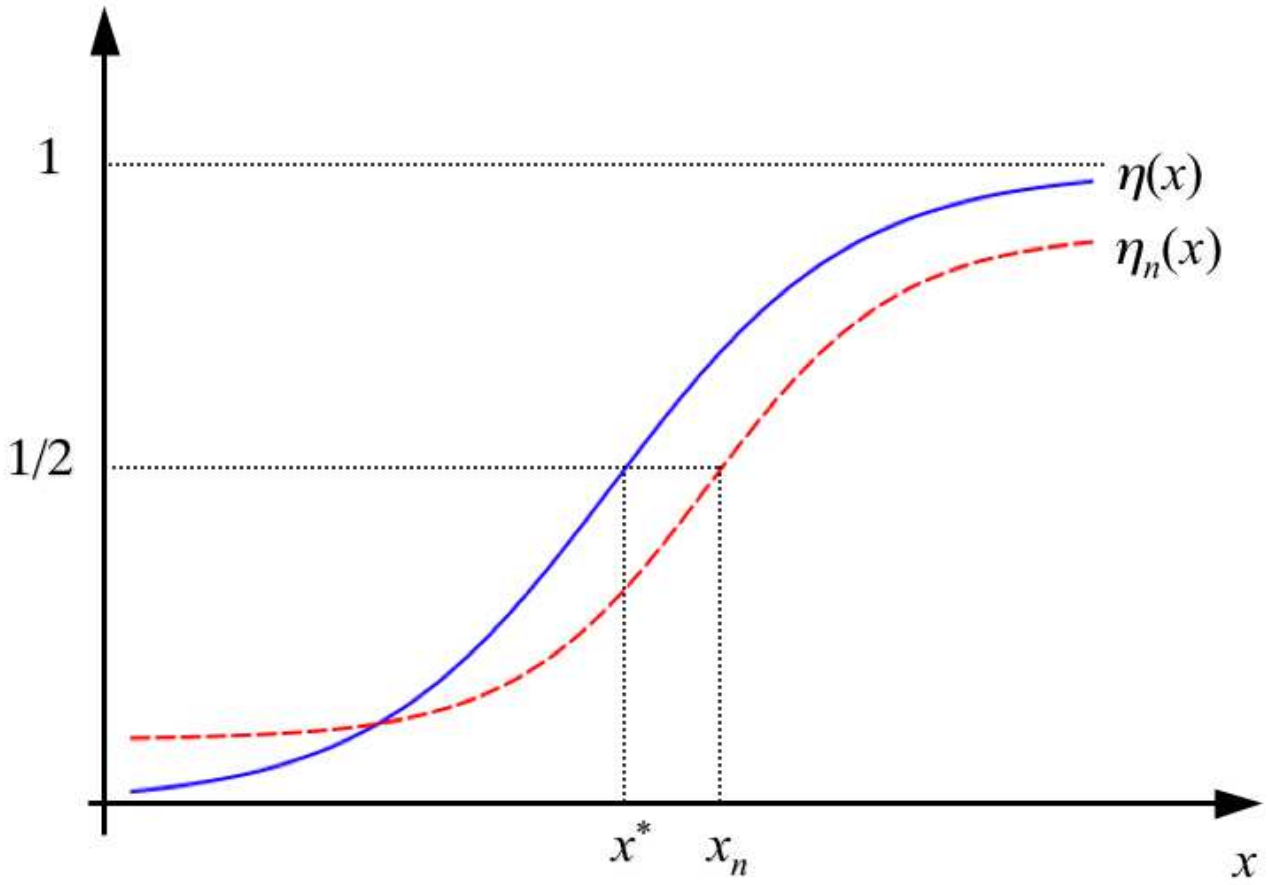
$$\psi^*(x) = \begin{cases} 1, & \eta(x) \geq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

so the idea is to approximate the unknown $\eta(x)$ by an estimate $\eta_n(x)$ based on the data S_n, and let

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

Plug-In Classification Rules - II

- Plug-in classifiers are in general sub-optimal. For instance, we can see from the 1-D example below that $\psi_n(x) \neq \psi^*(x)$ for $x^* < x < x_n$.



Logistic Regression

- A well-known example in Statistics is based on the "logit" function

$$\text{logit}(p) \;=\; \ln\left(\frac{p}{1-p}\right)\;, \quad 0 < p < 1$$

- We may assume (model) the posterior probability as linear in "logit" space (i.e. assume linear log-odds):

$$\text{logit}(\eta(x)) \;=\; \ln\left(\frac{\eta(x)}{1-\eta(x)}\right) \;=\; a + b^T x$$

so that

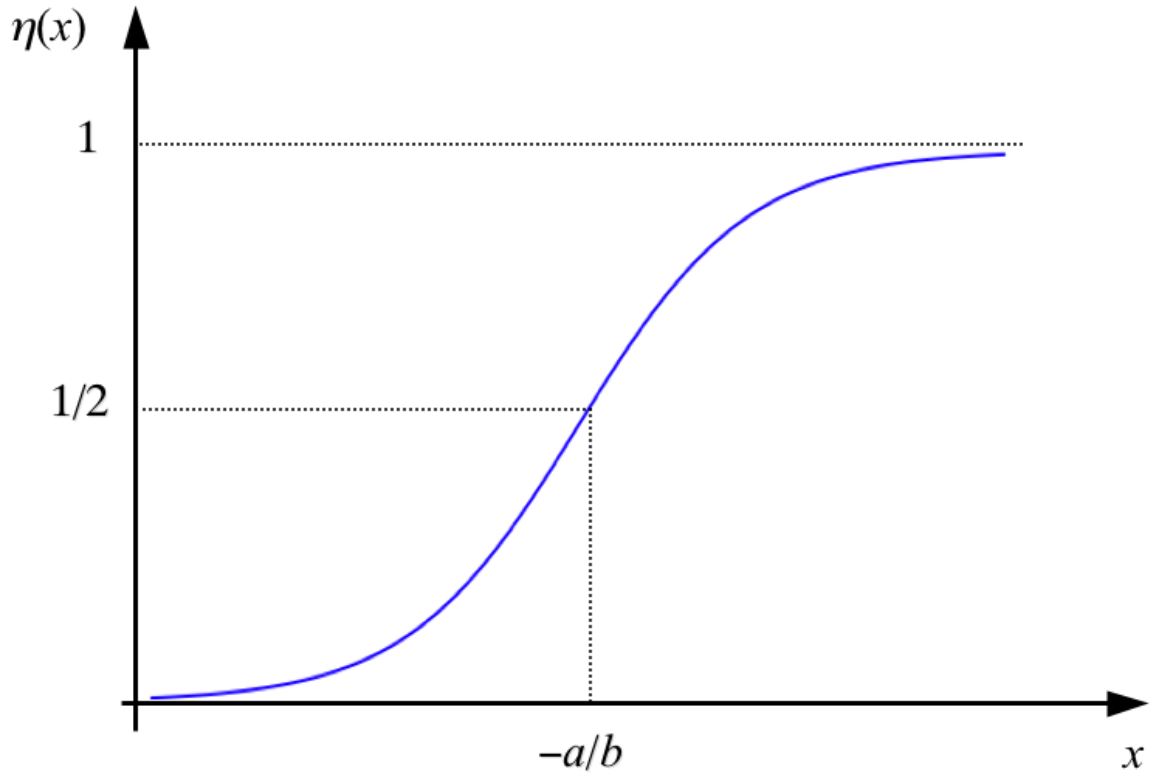
$$\eta(x) \;=\; \frac{e^{a+b^T x}}{1+e^{a+b^T x}} \;=\; \frac{1}{1+e^{-(a+b^T x)}}$$

Logistic Regression - II

- The function

$$\eta(x) = \frac{1}{1 + e^{-(a+b^T x)}}$$

is called the *logistic curve* with parameters a, b.



Logistic Regression - III

- In practice, one estimates coefficients a_n and b_n and plug them in the model, yielding:

$$\eta_n(x) = \frac{1}{1+e^{-(a_n+b_n^T x)}}$$

- One common way to do this is to estimate a_n and b_n by *maximum likelihood*, that is, pick a_n and b_n that maximizes the likelihood L_n (or its log) of seeing the data S_n under the model:

$$L_n = \Pi_{i=1}^n P_n(Y=y_i|X=x_i) = \Pi_{i=1}^n \eta_n(x_i)^{y_i}(1-\eta_n(x_i))^{1-y_i}$$

- This is a *parametric* classification rule. This rule is not consistent in general; we will see later sufficient conditions for consistency of parametric rules.

K-Nearest Neighbors

- The K-nearest neighbor classification rule is a well-known and useful methodology, which we will discuss in detail later.
- Here we just point out that it is a plug-in rule, with

$$\eta_n(x) = \frac{1}{K} \sum_{i=1}^{K} Y_{(i)}(x)$$

where $Y_{(i)}(x)$ indicates the label of the i-th closest sample point to x. Clearly, $\eta_n(x) \geq \frac{1}{2}$ if and only if the majority label among the K closest neighbors to x is 1, while $\eta_n(x) < \frac{1}{2}$ if and only if the majority label is 0.

- This is an example of *nonparametric* classification rule (more on this later). It is a universally consistent rule, if K is allowed to increase with n, at a certain rate.

Approximation Theorem

- One would like to guarantee that if $\eta_n(x)$ is close to $\eta(x)$ in some sense, then ϵ_n should be close to ϵ^*.

- The following famous distribution-free result shows that this is true if $\eta_n(x)$ is close to $\eta(x)$ in an "L_1-sense."

- (DGL Theorem 2.2): For any feature-label distribution F_{XY}, given the data S_n,

$$\begin{aligned} \epsilon_n - \epsilon^* &= 2E\left[|\eta(X) - 1/2|\, I_{\{\psi_n(X) \neq \psi^*(X)\}} \,\middle|\, S_n\right] \\ &\leq 2E\left[|\eta(X) - \eta_n(X)|\,|S_n\right] \end{aligned}$$

Consistency of Plug-In Rule

- As an easy corollary of the previous theorem, the plug-in rule is (resp. strongly) consistent if

$$E\left[|\eta(X) - \eta_n(X)|\,|S_n\right] = \int |\eta(x) - \eta_n(x)| p(x)\,dx \;\rightarrow\; 0$$

  in probability (resp. with probability 1).

- The plug-in rule is universally consistent if the previous convergence holds for all F_{XY}.

- This is a basic step in most proofs of consistency of classification rules.

Regularized Discriminant Analysis

- The most important parametric plug-in classification rules are the Gaussian discriminants: QDA, LDA, DLDA, and NMC. Among these, QDA demands the most data, followed by LDA, DLDA, and NMC.
- LDA regularizes QDA by *shrinking* the covariance matrix estimates into a single pooled estimate.
- The shrinkage from QDA to LDA can be controlled by introducing a parameter $0 \leq \alpha \leq 1$ and setting

$$\hat{\Sigma}_i^R(\alpha) = \frac{n_i(1-\alpha)\hat{\Sigma}_i + n\alpha\hat{\Sigma}}{n_i(1-\alpha) + n\alpha}, \quad i = 0, 1$$

where $\hat{\Sigma}$ is the pooled sample covariance matrix, and $\hat{\Sigma}_i$ and n_i are the individual sample covariance matrices and sample sizes; $\alpha = 0$ gives QDA and $\alpha = 1$ gives LDA.

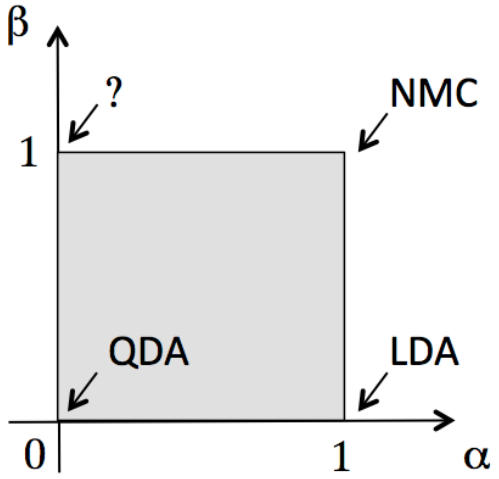
Regularized Discriminant Analysis - II

- To get more regularization while not overly increasing bias, one can further shrink $\hat{\Sigma}_i^R(\alpha)$ towards its average eigenvalue multiplied by the identity matrix, by introducing a further parameter $0 \leq \beta \leq 1$.

$$\hat{\Sigma}_i^R(\alpha, \beta) = (1-\beta)\hat{\Sigma}_i^R(\alpha) + \beta \frac{\text{trace}(\hat{\Sigma}_i(\alpha))}{d} I_d\,, \quad i = 0, 1$$

- Note that $\alpha = \beta = 1$ gives NMC. Hence, this rule ranges from QDA to LDA to NMC, and intermediate cases, depending on the selected values of α and β.
- This is called *regularized discriminant analysis*. Picking α and β is a process of *model selection* (more on this later), which can be based on availability of samples or optimization of some error criterion.

Regularized Discriminant Analysis - III



$\alpha = 0$ and $\beta = 0 \Rightarrow$ QDA

$\alpha = 1$ and $\beta = 0 \Rightarrow$ LDA

$\alpha = 1$ and $\beta = 1 \Rightarrow$ NMC

- If $\alpha = 0$ and $\beta = 1$ then $\hat{\Sigma}_0^R = m_0 I_d$ and $\hat{\Sigma}_1^R = m_1 I_d$, where $m_i = \text{trace}(\hat{\Sigma}_i)/d \geq 0$, for $i = 0, 1$.
- It can be shown that this leads to a *spherical* decision boundary for $m_0 \neq m_1$ ($m_0 = m_1$ yields the plain NMC):

$$\left|\left| X - \frac{m_1 \hat{\mu}_0 - m_0 \hat{\mu}_1}{m_1 - m_0} \right|\right|^2 = \frac{m_1 m_0}{m_1 - m_0} \left(\frac{||\hat{\mu}_1 - \hat{\mu}_0||^2}{m_1 - m_0} + d \ln \frac{m_1}{m_0} \right).$$