Section 4H. Linear Discriminant Analysis (LDA)

Statistics for Data Science

Victor M. Preciado, PhD MIT EECS
Dept of Electrical & Systems Engineering
University of Pennsylvania
preciado@seas.upenn.edu

Recap

Because of the *Curse of Dimensionality*, we use parametric models to estimate the conditional class probabilities $\Pr[Y = k|X = \mathbf{x}]$ from a given dataset $\mathcal{D}$ (instead of local averaging)

Logistic regression is a particular parametric model in which we first pick a random input $x \sim f_X$ and assign a label to this input according to conditional class probabilities of the form:

$$p_1(x; \beta_0, \beta_1) = \Pr(Y = 1|X = x; \beta_0, \beta_1) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

$$p_0(x; \beta_0, \beta_1) = \frac{1}{1 + e^{\beta_0 + \beta_1 x}}$$

However, this in not the only parametric model we can use...

Linear Discriminant Analysis

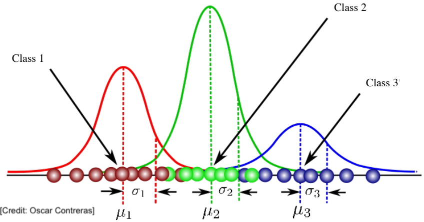
Linear Discriminant Analysis (LDA) assumes data is generated as follows:

1. For each sample point, we randomly draw an output label class k according to the following PMF (called the *prior* probabilities):

$$\Pr(Y = k) = \pi_k \text{ for } k = 1, \ldots, K$$

2. Once an output class is chosen, we generate an input point $\mathbf{x}$ as follows

$$f_k(\mathbf{x}) = \Pr(X = \mathbf{x} | Y = k) = \mathcal{N}(\mu_k, \Sigma_k)$$



[Credit: Oscar Contreras]

Linear Discriminant Analysis (cont.)

- Using Bayes theorem, we can compute the conditional class probabilities

$$\Pr\left(Y = k | X = \mathbf{x}\right) = \frac{\Pr\left(X = \mathbf{x} | Y = k\right) \Pr\left(Y = k\right)}{\sum_{\ell=1}^{K} \Pr\left(X = \mathbf{x} | Y = \ell\right) \Pr\left(Y = \ell\right)} = \frac{f_k\left(\mathbf{x}\right) \pi_k}{\sum_{\ell=1}^{K} f_\ell\left(\mathbf{x}\right) \pi_\ell}$$

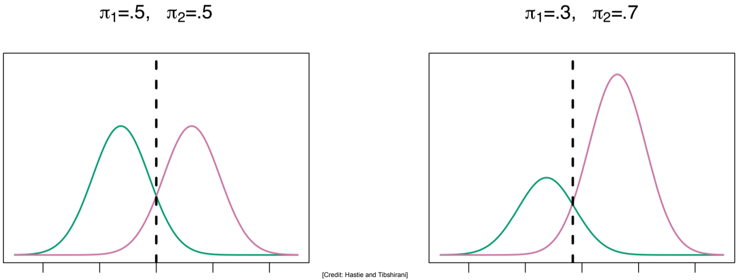
- Hence, we can find the Bayes optimal classifier as follows:

$$\begin{aligned} C\left(\mathbf{x}\right) &= \arg\max_k \Pr\left(Y = k | X = \mathbf{x}\right) \\ &= \arg\max_k \frac{f_k\left(\mathbf{x}\right) \pi_k}{\sum_{\ell=1}^{K} f_\ell\left(\mathbf{x}\right) \pi_\ell} = \arg\max_k f_k\left(\mathbf{x}\right) \pi_k \end{aligned}$$

  Therefore, given an input $\mathbf{x}$, we classify it according to the class that gives the maximum value of $f_k\left(\mathbf{x}\right) \pi_k$

LDA Example

Example: Compare a binary classification problem with different priors



[Credit: Hastie and Tibshirani]

Penn Engineering