The plug-in method

Let X be a random variable distributed according to Φ.
Let f be some function of the distribution (summary statistic) so that

$$\theta = f(\Phi).$$

Given a set of observations x_i, they describe an empirical distribution $\hat{\Phi}$, which leads to an empirical summary statistic

$$\hat{\theta} = f(\hat{\Phi}).$$

Example:
Let θ be the mean μ of Φ. Then, the above formula amounts to:

$$\hat{\theta} = E_{\hat{\Phi}}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Assume now that we have a sequence of empirical distributions Φ_k, so that from each we get

$$\theta_k^* = f(\Phi_k), \quad \phi^* = \frac{1}{K}\sum_{k=1}^{K} \phi_k^*.$$

For a given estimator θ, there are two important quantities describing the estimation error:

$$\begin{array}{rc} \textbf{Bias} & \theta - E_\Phi[\theta], \\ \textbf{Variance} & E_\Phi[\theta - E_\Phi(\theta)]^2. \end{array}$$

Using the above sequence of distributions we approximate the bias and variance by:

$$\begin{aligned} \theta - E_\Phi[\theta] &\simeq \hat{\theta} - \theta^* \\ E_{\hat{\Phi}}[\theta - E_{\hat{\Phi}}(\theta)]^2 &\simeq \frac{1}{K}\sum_k (\theta_k^* - \theta^*)^2. \end{aligned}$$

Linear discrimination

Given a set of two clusters $X_1,\ X_2 \in R^n$, let $w \in R^n$ be a projection of that data onto one dimension. We seek w that maximizes the distance between the two cluster centers. Let $m_1,\ m_2$ be the cluster centers in the original space:

$$m_i = \frac{1}{|X_i|} \sum_{x \in X_i} x.$$

From linearity, the cluster center in the projected space is given by

$$\hat{m}_i = w^T m_i.$$

We need to take the variance into account.

Denote the projected data (using w) by $Y_1,\ Y_2$ respectively. Let

$$\hat{s}_i^2 = \sum_{y \in Y_i} (y - \hat{m}_i)^2.$$

Note, we do not normalize the scatter. We define a functional

$$J(w) = \frac{|\hat{m}_1 - \hat{m}_2|^2}{\hat{s}_1^2 + \hat{s}_2^2}.$$

This is a good measure of the separation.

To seek w that maximizes $J(w)$ we define the following:

$$S_i = \sum_{x \in X_i} (x - m_i)(x - m_i)^T$$

$$\begin{array}{rl} \textbf{within class scatter} & S_W = S_1 + S_2, \\ \textbf{between class scatter} & S_B = (m_1 - m_2)(m_1 - m_2)^T. \end{array}$$

$$\hat{s}_1^2 + \hat{s}_2^2 = w^T S_1 w + w^T S_2 w = w^T S_W w.$$

$$\hat{(}m_1 - \hat{m}_2)^2 = w^T(m_1 - m_2)(m_1 - m_2)^T w = w^T S_B w.$$

$J(w)$ becomes:

$$J(w) = \frac{w^T S_B w}{w^T S_W w}.$$

If S_W has an inverse, then w must satisfy the following eigen value probelm:

$$S_W^{-1} S_B w = \lambda w.$$

It follows that

$$w = S_W^{-1}(m_1 - m_2).$$

Otherwise we have the generalized eigen value problem:

$$S_B w = \lambda S_W w.$$