Extension to M-ary Bayes Classifier

$H_y : \underline{X} \sim p(\underline{x}|y)$, $y = 1, \ldots, M$

priors $\pi_y = P\{Y = y\}$, $y = 1, \ldots, M$

$$\hat{y}_{Bayes} = f_{Bayes}(\underline{x}) = \arg\max_y \pi_y p(\underline{x}|y)$$

$$= \arg\max_y \left(\ln \pi_y + \ln p(\underline{x}|y)\right)$$

f_{Bayes} minimizes P_e $(\equiv Err_{pred})$

Multivariate Gaussian Distribution

Recall: Gaussian random variable X with df

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad X \sim N(\mu, \sigma^2)$$

(μ: mean; σ^2: variance)

Generalization to d-dimensional vector $\underline{X}$:

$$\phi(\underline{x}, \underline{\mu}, C) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|C|}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T C^{-1}(\underline{x}-\underline{\mu})}$$

$\underline{\mu}$: mean of $\underline{X}$; C: covariance matrix of $\underline{X}$

$$\underline{\mu} = E[\underline{X}], \quad C = E\left[(\underline{X}-\underline{\mu})(\underline{X}-\underline{\mu})^T\right]$$

$$\mu_\ell = E[X_\ell], \quad C_{\ell,m} = E\left[(X_\ell - \mu_\ell)(X_m - \mu_m)\right]$$

Compact Notation: $\underline{X} \sim N(\underline{\mu}, C)$

M-ary Classification with Multivariate Gaussians

$$H_y: \; p(\underline{x}|y) = \phi(\underline{x}, \underline{\mu}_y, C_y)$$

$$\hat{y}_{Bayes} = f_{Bayes}(\underline{x}) = \arg\max_y \left(\ln \pi_y + \ln p(\underline{x}|y) \right)$$

$$p(\underline{x}|y) = \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{|C_y|}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_y)^T C_y^{-1} (\underline{x}-\underline{\mu}_y)}$$

$$\ln p(\underline{x}|y) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|C_y| - \frac{1}{2}(\underline{x}-\underline{\mu}_y)^T C_y^{-1}(\underline{x}-\underline{\mu}_y)$$

$$\hat{y}_{Bayes} = \arg\max_y \underbrace{\ln \pi_y - \frac{1}{2}\ln|C_y| - \frac{1}{2}(\underline{x}-\underline{\mu}_y)^T C_y^{-1}(\underline{x}-\underline{\mu}_y)}_{\delta_y(\underline{x}):\ \text{objective function}}$$

If C_y's are different, $\delta_y(\underline{x})$ is quadratic in $\underline{x}$

If $C_y = C$, for all y, then

$$\frac{1}{2}(\underline{x}-\underline{\mu}_y)^T C_y^{-1}(\underline{x}-\underline{\mu}_y) = \frac{1}{2}\left(\underline{x}^T C^{-1} \underline{x} - \underline{\mu}_y^T C^{-1} \underline{x} - \underline{x}^T C^{-1} \underline{\mu}_y + \underline{\mu}_y^T C^{-1} \underline{\mu}_y \right)$$

$\underline{x}^T C^{-1} \underline{x}$: not fn. of y

$-\underline{\mu}_y^T C^{-1} \underline{x}$ and $-\underline{x}^T C^{-1} \underline{\mu}_y$: same

$$\Rightarrow \hat{y}_{Bayes} = \arg\max_y \underbrace{\left[\ln \pi_y + \underline{\mu}_y^T C^{-1} \underline{x} - \frac{1}{2} \underline{\mu}_y^T C^{-1} \underline{\mu}_y \right]}_{\text{linear objective } \delta_y(\underline{x})}$$

Linear Discriminant Analysis (LDA) Classifier

- Assuming $C_y = C$, for all y,

$$f_{Bayes}(x) = \arg\max_y \left[\ln \pi_y + \underline{\mu}_y^T C^{-1} \underline{x} - \frac{1}{2} \underline{\mu}_y^T C^{-1} \underline{\mu}_y \right]$$

- LDA approach replaces $\{\pi_\ell, \underline{\mu}_\ell\}_{\ell=1}^M$ and C by estimates $\{\hat{\pi}_\ell, \hat{\mu}_\ell\}_{\ell=1}^M$ and $\hat{C}$ that are obtained from training data $\tau = \{\underline{x}_i, y_i\}_{i=1}^N$

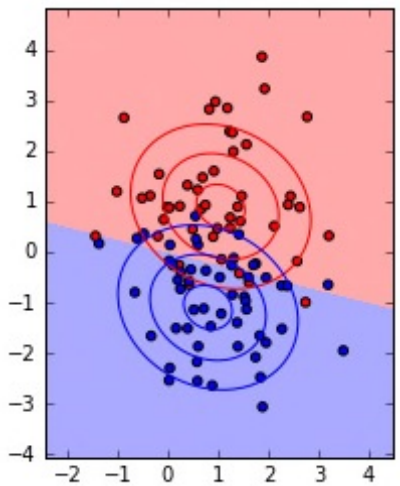
Estimation Let N_ℓ = # training data with class label ℓ

$$\hat{\pi}_\ell = \frac{N_\ell}{N}, \quad \hat{\mu}_\ell = \frac{\sum_{i: y_i = \ell} \underline{x}_i}{N}$$

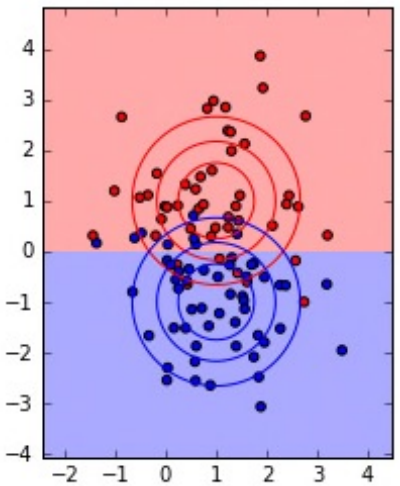
$$\hat{C} = \frac{\sum_{\ell=1}^M \sum_{i: y_i = \ell} (\underline{x}_i - \hat{\mu}_\ell)(\underline{x}_i - \hat{\mu}_\ell)^T}{(N - M) \leftarrow \text{or } N}$$

Let $\delta_\ell(\underline{x}) = \ln \hat{\pi}_\ell + \hat{\mu}_\ell^T \hat{C}^{-1} \underline{x} - \frac{1}{2} \hat{\mu}_\ell \hat{C}^{-1} \hat{\mu}_\ell$

Then $\hat{y}_{LDA} = f_{LDA}(\underline{x}) = \arg\max_\ell \delta_\ell(\underline{x})$



LDA



Bayes

Linear Classifiers

- Linear classifiers (e.g. LDA) are classifiers for which the objective $\delta_\ell(\underline{x})$ is linear

$$\delta_\ell(\underline{x}) = \underline{a}_\ell^T(\underline{x}) + b_\ell, \quad \ell = 1, 2, \ldots, M$$

For LDA, $\underline{a}_\ell^T = \hat{\underline{\mu}}_\ell^T \hat{C}^{-1}$, $b_\ell = \ln \hat{\pi}_\ell - \frac{1}{2} \hat{\underline{\mu}}_\ell \hat{C}^{-1} \hat{\underline{\mu}}_\ell$

$$\hat{y}_{lin} = f_{lin}(\underline{x}) = \arg\max_\ell \left(\underline{a}_\ell^T(\underline{x}) + b_\ell \right)$$

- Decision between classes ℓ and m:

$$\underline{a}_\ell^T \underline{x} + b_\ell \underset{m}{\overset{\ell}{\gtrless}} \underline{a}_m^T \underline{x} + b_m$$

$$\equiv \underline{w}_{\ell m}^T \underline{x} \underset{m}{\overset{\ell}{\gtrless}} \beta_{\ell m} \qquad \underline{w}_{\ell m} = \underline{a}_\ell - \underline{a}_m, \quad \beta_{\ell m} = b_m - b_\ell$$

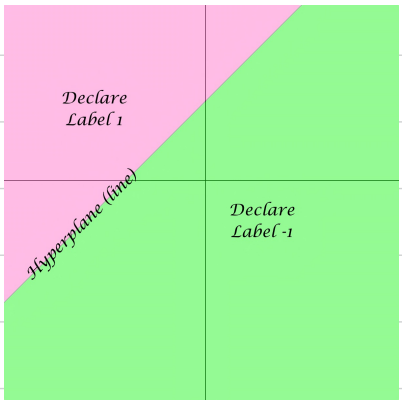
- Define $g_{\ell m}(\underline{x}) = \underline{w}_{\ell m}^T \underline{x} - \beta_{\ell m} = -g_{m\ell}(\underline{x})$ ← linear discriminant function

Then $g_{\ell m}(x) \underset{m}{\overset{\ell}{\gtrless}} 0$. Note: $g_{\ell m}(\underline{x}) = \delta_\ell(\underline{x}) - \delta_m(\underline{x})$

- $g_{\ell m}(\underline{x}) = 0$ defines a hyperplane in $\mathbb{R}^d$.

Binary Linear Classifier

$$\hat{y}_{lin} = f_{lin}(x) = \begin{cases} 1 & \text{if } \underline{w}^T \underline{x} \geq \beta \\ -1 & \text{if } \underline{w}^T \underline{x} < \beta \end{cases}$$



Example $M = 3$, $d = 2$, $\underline{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$$\delta_1(\underline{x}) = x_1 + 2x_2 + 1$$
$$\delta_2(\underline{x}) = 2x_1 - x_2 + 2$$
$$\delta_3(\underline{x}) = -x_1 + 2x_2 + 1$$

Linear discriminant functions:

$$g_{12}(\underline{x}) = -x_1 + 3x_2 - 1$$
$$g_{23}(\underline{x}) = 3x_1 - 3x_2 + 1$$
$$g_{31}(\underline{x}) = -2x_1$$

Suppose we want to classify $\underline{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

1. Using $\delta_\ell(\underline{x})$:

$$\delta_1(\underline{x}) = 4, \quad \delta_2(\underline{x}) = 3, \quad \delta_3(x) = 0$$
$$\Rightarrow \quad \hat{y} = \arg\max_{\ell} \delta_\ell(\underline{x}) = 1$$

2. Using $g_{\ell m}(\underline{x})$:

$$g_{12}(\underline{x}) = 2 > 0$$
$$g_{23}(\underline{x}) = 1 > 0$$
$$g_{31}(\underline{x}) = -2 < 0 \Rightarrow g_{13}(\underline{x}) > 0$$

$$\Rightarrow \quad \hat{y} = 1$$