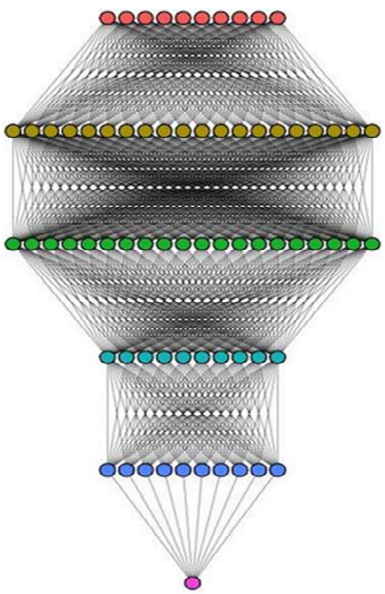
Logistic Regression

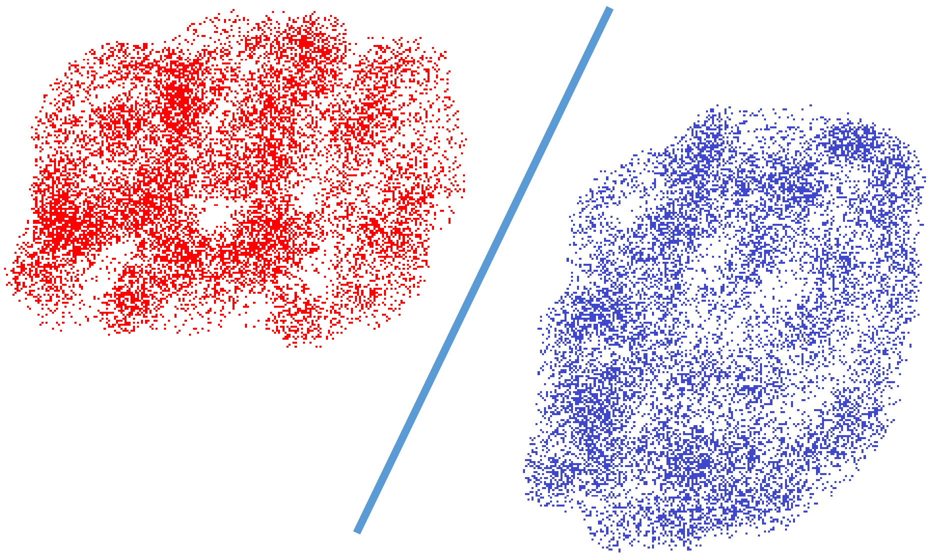
Prithwijit Guha
Dept. of EEE, IIT Guwahati

Supervised Learning



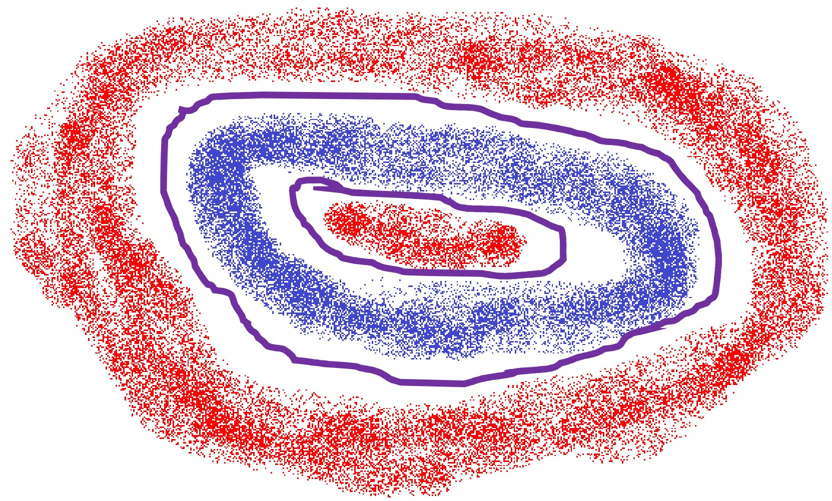
- Bayesian Classification, MAP, Chebyshev Inequality
- Performance Measures, Confusion Matrix, ROC Curves
- Logistic Regression
- Perceptron
- Multi-Layer Perceptron (MLP), ELM
- MLP Architectures, Learning, Interpretations
- Non-parametric Methods and K-NN
- Radial Basis Function Neural Networks
- Data Balancing; SMOTE & Weighted Loss Functions
- Classification & Regression Trees
- Support Vector Machines & Multiple Kernel Learning
- Ensemble Methods, Bagging and Boosting

Separable Classes



Linearly Separable

Not Linearly Separable



Classification: Input Data & Label

$$\boldsymbol{X_0} = \{\boldsymbol{x}_i^0 : \boldsymbol{x}_i^0 \in \mathbb{R}^D; i = 1, \dots n_0\}$$

$$\boldsymbol{X_1} = \{\boldsymbol{x}_j^1 : \boldsymbol{x}_j^1 \in \mathbb{R}^D; j = 1, \dots n_1\}$$

$$y(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{x} \in \boldsymbol{X_1} \\ 0, & \boldsymbol{x} \in \boldsymbol{X_0} \end{cases}$$

Classification: Input Distribution

$$\boldsymbol{x} \in \boldsymbol{X_0} \Rightarrow \boldsymbol{x} \sim P_0(\boldsymbol{x}; \boldsymbol{\theta}_0)$$

$$\boldsymbol{x} \in \boldsymbol{X_1} \Rightarrow \boldsymbol{x} \sim P_1(\boldsymbol{x}; \boldsymbol{\theta}_1)$$

P_0 and P_1 are the respective Probability Distributions learned from $\boldsymbol{X_0}$ and $\boldsymbol{X_1}$. The respective parameters of these Distributions are $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta_1}$.

Classification: Input Distribution

$$P(\boldsymbol{X_1}|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\boldsymbol{X_1}, \boldsymbol{\theta_1})P(\boldsymbol{X_1})}{P(\boldsymbol{x})}$$

Posterior **Likelihood** **Prior**

$$P(\boldsymbol{X_0}|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\boldsymbol{X_0}, \boldsymbol{\theta_0})P(\boldsymbol{X_0})}{P(\boldsymbol{x})}$$

Evidence

$$P(\boldsymbol{x}) = P(\boldsymbol{x}|\boldsymbol{X_0}, \boldsymbol{\theta_0})P(\boldsymbol{X_0}) + P(\boldsymbol{x}|\boldsymbol{X_1}, \boldsymbol{\theta_1})P(\boldsymbol{X_1})$$

Discriminant Functions & Decision Rule

$$y(\boldsymbol{x}) = \begin{cases} 1, & P(\boldsymbol{X_1}|\boldsymbol{x}) > P(\boldsymbol{X_0}|\boldsymbol{x}) \\ 0, & P(\boldsymbol{X_1}|\boldsymbol{x}) < P(\boldsymbol{X_0}|\boldsymbol{x}) \end{cases}$$

Discriminant Function

$$g_i(\boldsymbol{x}) = \ln\{P(\boldsymbol{X_i}|\boldsymbol{x})\}$$

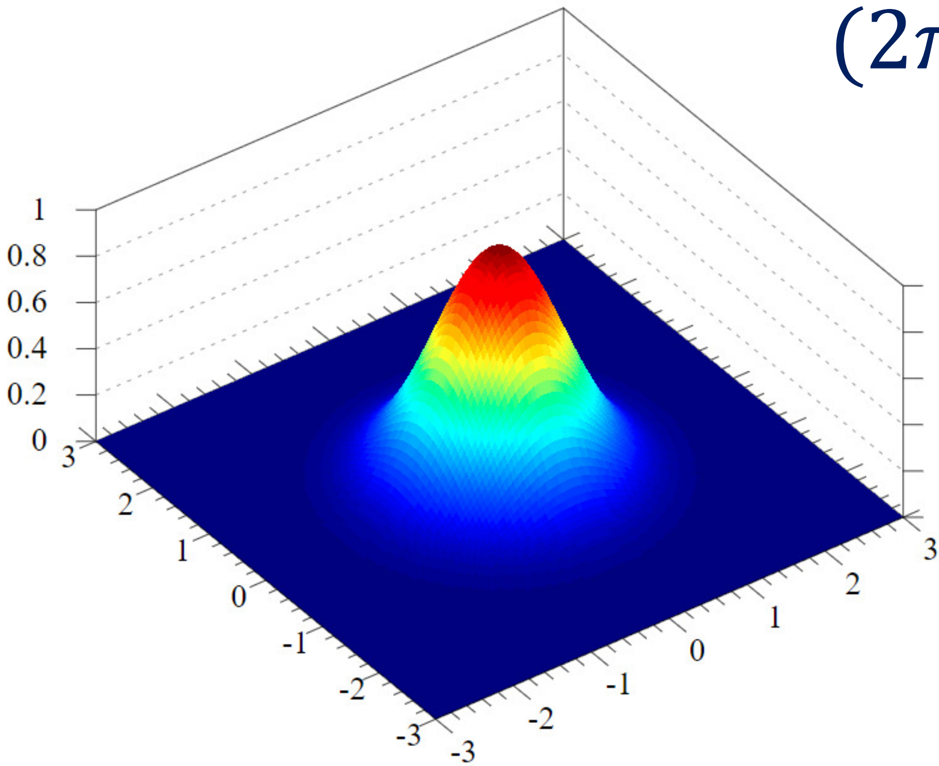
$$g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_0(\boldsymbol{x})$$

$$y(\boldsymbol{v}) = \begin{cases} 1, & g(\boldsymbol{v}) = g_1(\boldsymbol{v}) - g_0(\boldsymbol{v}) > 0 \\ 0, & g(\boldsymbol{v}) = g_1(\boldsymbol{v}) - g_0(\boldsymbol{v}) < 0 \end{cases}$$

Classification Decision Rule (unseen data $\boldsymbol{v}$)

Discriminant Functions: Gaussian Distribution

$$P(\boldsymbol{x}; \boldsymbol{\theta} = [\boldsymbol{\mu}, \boldsymbol{C}]) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{C}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$



Case-1: $\boldsymbol{C}_1 = \boldsymbol{C}_2 = \sigma^2 \boldsymbol{I}$

Case-2: $\boldsymbol{C}_1 = \boldsymbol{C}_2 = \boldsymbol{C}$

Case-3: $\boldsymbol{C}_1 \neq \boldsymbol{C}_2$

Discriminant Function: Gaussian Distribution

$$g_i(\boldsymbol{x}) = ln\{P(\boldsymbol{X_i}|\boldsymbol{x})\} = ln\left\{\frac{P(\boldsymbol{x}|\boldsymbol{X_i},\boldsymbol{\theta_i})P(\boldsymbol{X_i})}{P(\boldsymbol{x})}\right\}$$

$$= ln\{P(\boldsymbol{x}|\boldsymbol{X}_i,\boldsymbol{\theta_i})\} + ln\{P(\boldsymbol{X_i})\} - ln\{P(\boldsymbol{x})\}$$

$$\boxed{\begin{aligned} g_i(\boldsymbol{x}) = -\frac{n}{2}ln\{2\pi\} - \frac{1}{2}ln\{|\boldsymbol{C}_i|\} - \frac{1}{2}(x-\boldsymbol{\mu}_i)^T\boldsymbol{C}_i^{-1}(x-\boldsymbol{\mu}_i) \\ + ln\{P(\boldsymbol{X_i})\} - ln\{P(\boldsymbol{x})\} \end{aligned}}$$

Discriminant Function: Gaussian Distribution

$$g_i(\boldsymbol{x}) = -\frac{n}{2}ln\{2\pi\} - \frac{1}{2}ln\{|\boldsymbol{C}_i|\} - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{C}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) + ln\{P(\boldsymbol{X_i})\} - ln\{P(\boldsymbol{x})\}$$

$$\boxed{g(\boldsymbol{x}) = -\frac{1}{2}ln\left\{\frac{|\boldsymbol{C}_1|}{|\boldsymbol{C}_0|}\right\} + ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\} - \frac{1}{2}\{(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{C}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) - (\boldsymbol{x}-\boldsymbol{\mu}_0)^T\boldsymbol{C}_0^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)\}}$$

Discriminant Function: $\boldsymbol{C_1} = \boldsymbol{C_0} = \sigma^2 \boldsymbol{I}$

$$g(\boldsymbol{x}) = -\frac{1}{2} ln\left\{\frac{|\boldsymbol{C}_1|}{|\boldsymbol{C}_0|}\right\} + ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\} - \frac{1}{2}\{(\boldsymbol{x}-\boldsymbol{\mu}_1)^T \boldsymbol{C}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) - (\boldsymbol{x}-\boldsymbol{\mu}_0)^T \boldsymbol{C}_0^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)\}$$

$$= -\frac{1}{2} ln\left\{\frac{\sigma^2}{\sigma^2}\right\} + ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\} - \frac{1}{2}\{(\boldsymbol{x}-\boldsymbol{\mu}_1)^T(\sigma^{-2}\boldsymbol{I})\,(\boldsymbol{x}-\boldsymbol{\mu}_1) - (\boldsymbol{x}-\boldsymbol{\mu}_0)^T(\sigma^{-2}\boldsymbol{I})(\boldsymbol{x}-\boldsymbol{\mu}_0)\}$$

$$= 0 + ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\} - \frac{1}{2\sigma^2}\{(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\,(\boldsymbol{x}-\boldsymbol{\mu}_1) - (\boldsymbol{x}-\boldsymbol{\mu}_0)^T\,(\boldsymbol{x}-\boldsymbol{\mu}_0)\}$$

$$= -\frac{1}{2\sigma^2}\{(\boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{\mu}_1^T\,\boldsymbol{x} + \boldsymbol{\mu}_1^T\boldsymbol{\mu}_1) - (\boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{\mu}_0^T\,\boldsymbol{x} + \boldsymbol{\mu}_0^T\boldsymbol{\mu}_0)\} + ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\}$$

$$= -\frac{1}{2\sigma^2}\{-2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\boldsymbol{x} + (\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T\boldsymbol{\mu}_0)\} + ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\}$$

Discriminant Function: $\boldsymbol{C_1} = \boldsymbol{C_0} = \sigma^2 \boldsymbol{I}$

$$g(\boldsymbol{x}) = -\frac{1}{2\sigma^2}\{-2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{x} + (\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T\boldsymbol{\mu}_0)\} + ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\}$$

$$g(\boldsymbol{x}) = \frac{1}{\sigma^2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{x} + ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\} - \frac{1}{2\sigma^2}(\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T\boldsymbol{\mu}_0)$$

$$g(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x} + b$$

$$\boldsymbol{a} = \frac{1}{\sigma^2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$$

$$b = ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\} - \frac{1}{2\sigma^2}(\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T\boldsymbol{\mu}_0)$$

Discriminant Function: $\boldsymbol{C_1 = C_0 = C}$

$$g(\boldsymbol{x}) = -\frac{1}{2} ln\left\{\frac{|\boldsymbol{C}_1|}{|\boldsymbol{C}_0|}\right\} + ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\} - \frac{1}{2}\{(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{C}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) - (\boldsymbol{x}-\boldsymbol{\mu}_0)^T\boldsymbol{C}_0^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)\}$$

$$= -\frac{1}{2} ln\left\{\frac{|\boldsymbol{C}|}{|\boldsymbol{C}|}\right\} + ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\} - \frac{1}{2}\{(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) - (\boldsymbol{x}-\boldsymbol{\mu}_0)^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)\}$$

$$= -\frac{1}{2}\{(\boldsymbol{x}^T\boldsymbol{C}^{-1}\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{C}^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^T\boldsymbol{C}^{-1}\boldsymbol{\mu}_1) - (\boldsymbol{x}^T\boldsymbol{C}^{-1}\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{C}^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\mu}_0^T\boldsymbol{C}^{-1}\boldsymbol{\mu}_0)\}$$
$$+ ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\}$$

$$= -\frac{1}{2}\{-2\boldsymbol{x}^T\boldsymbol{C}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0) + (\boldsymbol{\mu}_1^T\boldsymbol{C}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T\boldsymbol{C}^{-1}\boldsymbol{\mu}_0)\} + ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\}$$

Discriminant Function: $\boldsymbol{C_1} = \boldsymbol{C_0} = \boldsymbol{C}$

$$g(\boldsymbol{x}) = -\frac{1}{2}\{-2\boldsymbol{x}^T\boldsymbol{C}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) + (\boldsymbol{\mu}_1^T\boldsymbol{C}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T\boldsymbol{C}^{-1}\boldsymbol{\mu}_0)\} + ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\}$$

$$g(\boldsymbol{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{C}^{-1}\boldsymbol{x} + ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\} - \frac{1}{2}(\boldsymbol{\mu}_1^T\boldsymbol{C}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T\boldsymbol{C}^{-1}\boldsymbol{\mu}_0)$$

$$g(\boldsymbol{x}) = \boldsymbol{p}^T\boldsymbol{x} + q$$

$$\boldsymbol{p} = \boldsymbol{C}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$$

$$q = ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\} - \frac{1}{2}(\boldsymbol{\mu}_1^T\boldsymbol{C}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T\boldsymbol{C}^{-1}\boldsymbol{\mu}_0)$$

Discriminant Function: $\boldsymbol{C_1} \neq \boldsymbol{C_0}$

$$g(\boldsymbol{x}) = -\frac{1}{2} ln\left\{\frac{|\boldsymbol{C}_1|}{|\boldsymbol{C}_0|}\right\} + ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\} - \frac{1}{2}\{(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{C}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) - (\boldsymbol{x}-\boldsymbol{\mu}_0)^T\boldsymbol{C}_0^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)\}$$

$$= -\frac{1}{2}\{(\boldsymbol{x}^T\boldsymbol{C}_1^{-1}\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{C}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{\mu}_1) - (\boldsymbol{x}^T\boldsymbol{C}_0^{-1}\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{C}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\mu}_0^T\boldsymbol{C}_0^{-1}\boldsymbol{\mu}_0)\}$$
$$+ ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\} - \frac{1}{2} ln\left\{\frac{|\boldsymbol{C}_1|}{|\boldsymbol{C}_0|}\right\}$$

$$= -\frac{1}{2}\{\boldsymbol{x}^T(\boldsymbol{C}_1^{-1} - \boldsymbol{C}_0^{-1})\boldsymbol{x} - 2\boldsymbol{x}^T(\boldsymbol{C}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{C}_0^{-1}\boldsymbol{\mu}_0) + (\boldsymbol{\mu}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T\boldsymbol{C}_0^{-1}\boldsymbol{\mu}_0)\} + ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\}$$
$$- \frac{1}{2} ln\left\{\frac{|\boldsymbol{C}_1|}{|\boldsymbol{C}_0|}\right\}$$

Discriminant Function: $\boldsymbol{C_1} \neq \boldsymbol{C_0}$

$$g(\boldsymbol{x}) = -\frac{1}{2}\{\boldsymbol{x}^T(\boldsymbol{C}_1^{-1} - \boldsymbol{C}_0^{-1})\boldsymbol{x} - 2\boldsymbol{x}^T(\boldsymbol{C}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{C}_0^{-1}\boldsymbol{\mu}_0) + (\boldsymbol{\mu}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T\boldsymbol{C}_0^{-1}\boldsymbol{\mu}_0)\}$$
$$+ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\} - \frac{1}{2}ln\left\{\frac{|\boldsymbol{C}_1|}{|\boldsymbol{C}_0|}\right\})$$

$$g(\boldsymbol{x}) = \left[ln\left\{\frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})}\right\} - \frac{1}{2}(\boldsymbol{\mu}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T\boldsymbol{C}_0^{-1}\boldsymbol{\mu}_0) - \frac{1}{2}ln\left\{\frac{|\boldsymbol{C}_1|}{|\boldsymbol{C}_0|}\right\}\right] + \boldsymbol{x}^T(\boldsymbol{C}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{C}_0^{-1}\boldsymbol{\mu}_0)$$
$$-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{C}_1^{-1} - \boldsymbol{C}_0^{-1})\boldsymbol{x}$$

Discriminant Function: $\boldsymbol{C_1} \neq \boldsymbol{C_0}$

$$g(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} + \boldsymbol{b}^T \boldsymbol{x} + c$$

$$\boldsymbol{A} = \boldsymbol{C}_1^{-1} - \boldsymbol{C}_0^{-1}$$

$$\boldsymbol{b} = \boldsymbol{C}_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{C}_0^{-1} \boldsymbol{\mu}_0$$

$$c = \left[ln \left\{ \frac{P(\boldsymbol{X_1})}{P(\boldsymbol{X_0})} \right\} - \frac{1}{2} (\boldsymbol{\mu}_1^T \boldsymbol{C}_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T \boldsymbol{C}_0^{-1} \boldsymbol{\mu}_0) - \frac{1}{2} ln \left\{ \frac{|\boldsymbol{C}_1|}{|\boldsymbol{C}_0|} \right\} \right]$$

Log Odds and Logit Transform

$\boldsymbol{X_0} = \{\boldsymbol{x}_i^0 : \boldsymbol{x}_i^0 \in \mathbb{R}^D; i = 1, \ldots n_0\}$ $\quad$ $\boldsymbol{X_1} = \{\boldsymbol{x}_j^1 : \boldsymbol{x}_j^1 \in \mathbb{R}^D; j = 1, \ldots n_1\}$

$\boldsymbol{1 - y} = P(\, C_0 \mid \boldsymbol{x}\,)$ $\quad$ $\boldsymbol{y} = P(\, C_1 \mid \boldsymbol{x}\,)$

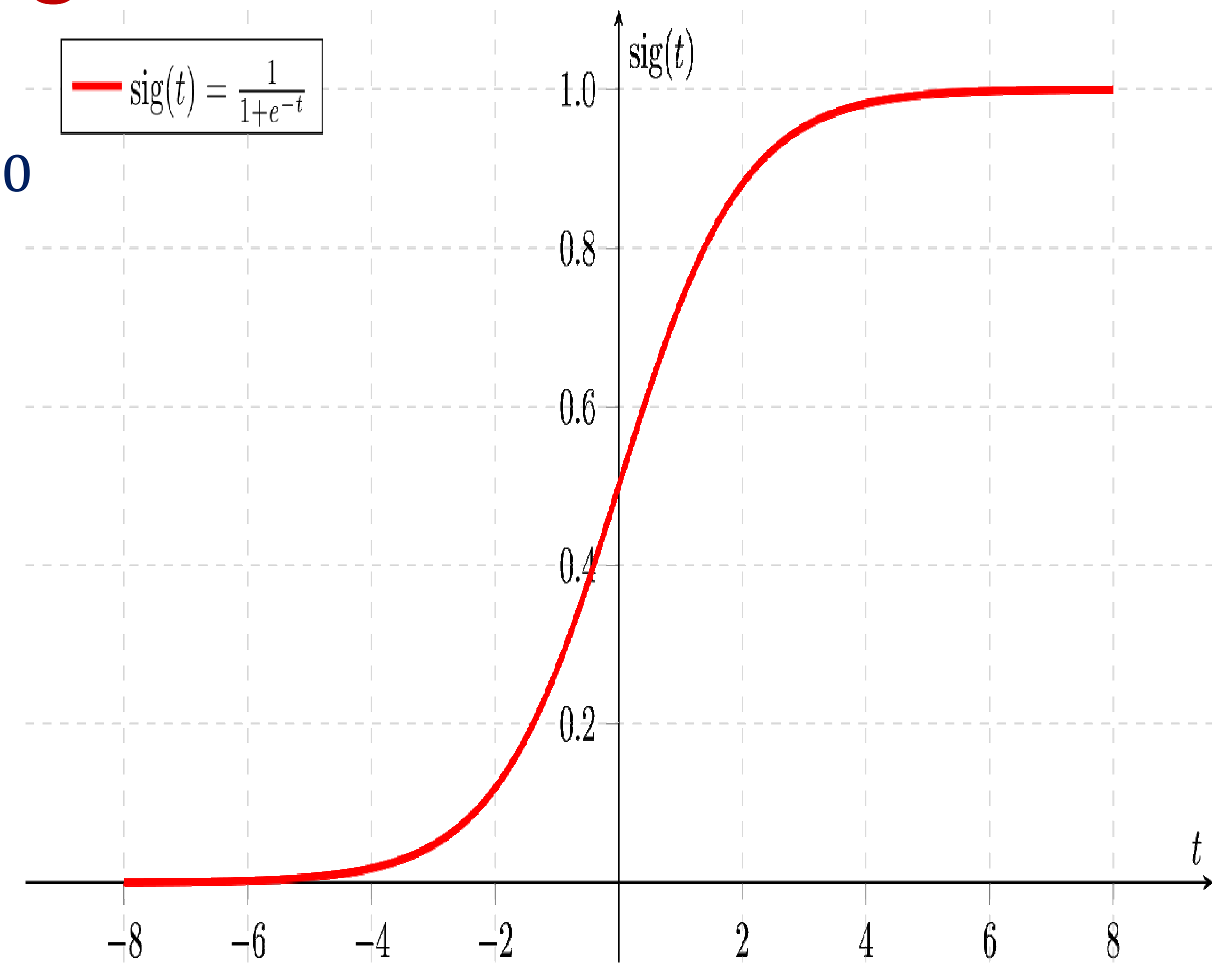
$$P(\, C_1 \mid \boldsymbol{x}\,) \Rightarrow y > 0.5 \Rightarrow \frac{y}{1-y} > 1 \Rightarrow log\left\{\frac{y}{1-y}\right\} > 0$$

The Log Odds and Sigmoid

$$log\left\{\frac{y}{1-y}\right\} = \boldsymbol{\omega}^T x + \omega_0$$

$$\frac{1-y}{y} = \mathrm{e}^{-(\boldsymbol{\omega}^T x + \omega_0)}$$

$$y = \frac{1}{1 + \mathrm{e}^{-(\boldsymbol{\omega}^T x + \omega_0)}}$$



The Likelihood Function

$P(y = 1 \mid \boldsymbol{x}) = h(\boldsymbol{x})$

$P(\, y = 0 \mid \boldsymbol{x}\,) = 1 - h(\boldsymbol{x})$

$$P(\, y \mid \boldsymbol{x}\,) = \{h(\boldsymbol{x})\}^{y}\{1 - h(\boldsymbol{x})\}^{1-y}$$

The Log-Likelihood

$$P(\, y \mid \boldsymbol{x}; \boldsymbol{\omega}, \omega_0 \,) = \{h(\boldsymbol{x})\}^{y}\{1 - h(\boldsymbol{x})\}^{1-y}$$

$$\boldsymbol{S} = \{(\boldsymbol{x}_i, y_i); i = 1, \dots n\}$$

i.i.d.

$$l(\boldsymbol{\omega}, \omega_0) = \prod_{i=1}^{n} P(\, y_i \mid \boldsymbol{x}_i; \boldsymbol{\omega}, \omega_0 \,)$$

The Log-Likelihood

$$l(\boldsymbol{\omega}, \omega_0) = \prod_{i=1}^{n} \{h(\boldsymbol{x}_i)\}^{y_i} \{1 - h(\boldsymbol{x}_i)\}^{1-y_i}$$

Log Likelihood

$$\mathrm{L}(\boldsymbol{\omega}, \omega_0) = \sum_{i=1}^{n} [y_i log\{h(\boldsymbol{x}_i)\} + \{1 - y_i\} log\{1 - h(\boldsymbol{x}_i)\}]$$

Maximizing the Log-Likelihood

$$\boldsymbol{\omega}_{\mathrm{j}}^{(\mathrm{k+1})} = \boldsymbol{\omega}_j^{(k)} + \eta_{jk} \frac{\partial \mathrm{L}(\boldsymbol{\omega}, \omega_0)}{\partial \boldsymbol{\omega}_j}$$

Gradient Ascent

$$\omega_0^{(\mathrm{k+1})} = \omega_0^{(k)} + \eta_k \frac{\partial \mathrm{L}(\boldsymbol{\omega}, \omega_0)}{\partial \omega_0}$$

Maximizing the Log-Likelihood

$$\mathrm{L}(\boldsymbol{\omega}, \omega_0) = \sum_{i=1}^{n} [y_i log\{h(\boldsymbol{x}_i)\} + \{1 - y_i\} log\{1 - h(\boldsymbol{x}_i)\}]$$

$$= \sum_{i=1}^{n} [y_i log\{h(\boldsymbol{x}_i)\} + log\{1 - h(\boldsymbol{x}_i)\} - y_i log\{1 - h(\boldsymbol{x}_i)\}]$$

$$= \sum_{i=1}^{n} \left[y_i log \left\{ \frac{h(\boldsymbol{x}_i)}{1 - h(\boldsymbol{x}_i)} \right\} + log\{1 - h(\boldsymbol{x}_i)\} \right]$$

Maximizing the Log-Likelihood

$$h(\boldsymbol{x}_i) = \frac{1}{1+e^{-u_i}} \Rightarrow \frac{h(\boldsymbol{x}_i)}{1-h(\boldsymbol{x}_i)} = e^{u_i}$$

$$\mathrm{L}(\boldsymbol{\omega}, \omega_0) = \sum_{i=1}^{n} \left[y_i log \left\{ \frac{h(\boldsymbol{x}_i)}{1-h(\boldsymbol{x}_i)} \right\} + log\{1-h(\boldsymbol{x}_i)\} \right]$$

$$\Downarrow$$

$$\mathrm{L}(\boldsymbol{\omega}, \omega_0) = \sum_{i=1}^{n} [y_i u_i + log\{1-h(x_i)\}]$$

$$\boxed{u_i = \sum_{r=1}^{d} \boldsymbol{\omega}_r x_{ir} + \omega_0}$$

Maximizing the Log-Likelihood

$$\mathrm{L}(\boldsymbol{\omega}, \omega_0) = \sum_{i=1}^{n} [y_i u_i + log\{1 - h(x_i)\}]$$

$$\Downarrow$$

$$\frac{\partial \mathrm{L}(\boldsymbol{\omega}, \omega_0)}{\partial \boldsymbol{\omega}_j} = \sum_{i=1}^{n} \left[y_i \frac{\partial u_i}{\partial \boldsymbol{\omega}_j} - \frac{1}{1 - h(\boldsymbol{x}_i)} \frac{\partial h(\boldsymbol{x}_i)}{\partial \boldsymbol{\omega}_j} \right]$$

Maximizing the Log-Likelihood

$$u_i = \sum_{r=1}^{d} \boldsymbol{\omega}_r x_{ir} + \omega_0 \quad \Rightarrow \quad \frac{\partial u_i}{\partial \boldsymbol{\omega}_j} = \boldsymbol{x}_{ij}$$

$$h(\boldsymbol{x}_i) = \frac{1}{1 + e^{-u_i}}$$

$$\frac{\partial h(\boldsymbol{x}_i)}{\partial \boldsymbol{\omega}_j} = \frac{\partial h(\boldsymbol{x}_i)}{\partial u_i} \times \frac{\partial u_i}{\partial \boldsymbol{\omega}_j} = -(1 + e^{-u_i})^{-2}(-e^{-u_i})(\boldsymbol{x}_{ij})$$

$$\frac{\partial h(\boldsymbol{x}_i)}{\partial \boldsymbol{\omega}_j} = \frac{e^{-u_i}}{(1 + e^{-u_i})^2}\left(\boldsymbol{x}_{ij}\right) = h(\boldsymbol{x}_i)\{1 - h(\boldsymbol{x}_i)\}(\boldsymbol{x}_{ij})$$

Maximizing the Log-Likelihood

$$\frac{\partial \mathrm{L}(\boldsymbol{\omega}, \omega_0)}{\partial \boldsymbol{\omega}_j} = \sum_{i=1}^{n} \left[y_i \frac{\partial u_i}{\partial \boldsymbol{\omega}_j} - \frac{1}{1 - h(\boldsymbol{x}_i)} \frac{\partial h(\boldsymbol{x}_i)}{\partial \boldsymbol{\omega}_j} \right]$$

$$\frac{\partial \mathrm{L}(\boldsymbol{\omega}, \omega_0)}{\partial \boldsymbol{\omega}_j} = \sum_{i=1}^{n} \left[y_i \boldsymbol{x}_{ij} - \frac{1}{1 - h(\boldsymbol{x}_i)} h(\boldsymbol{x}_i)\{1 - h(\boldsymbol{x}_i)\}(\boldsymbol{x}_{ij}) \right]$$

Maximizing the Log-Likelihood

$$\frac{\partial \mathrm{L}(\boldsymbol{\omega}, \omega_0)}{\partial \boldsymbol{\omega}_j} = \sum_{i=1}^{n} \{y_i - h(\boldsymbol{x}_i)\}(\boldsymbol{x}_{ij})$$

$$\frac{\partial \mathrm{L}(\boldsymbol{\omega}, \omega_0)}{\partial \omega_0} = \sum_{i=1}^{n} \{y_i - h(\boldsymbol{x}_i)\}$$

Summary

- Recapitulating Discriminant Functions
- From Log Odds To Sigmoid Function
- Logistic Regression
- Maximum Likelihood Formulation

Thank You