Classification

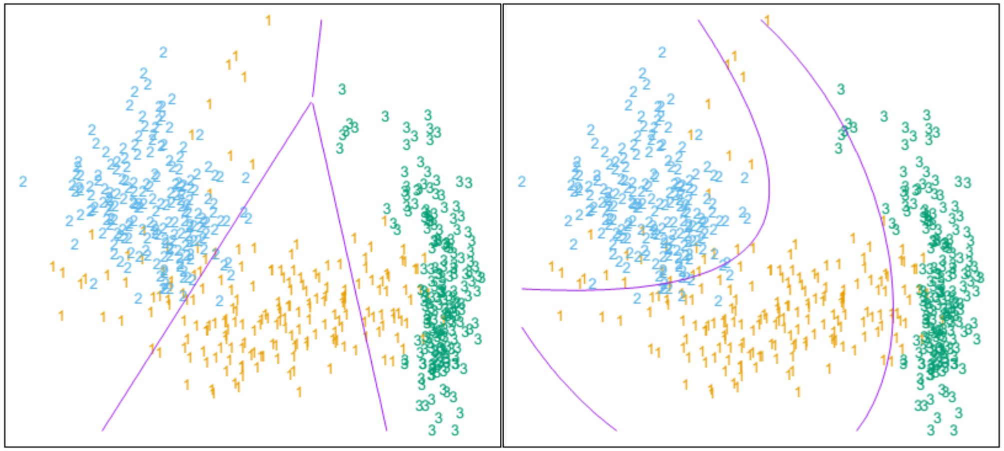
- Finding a separating surface
- Input-output mapping based on training data

  $\{\mathbf{x}_i, d_i\}_{i=1}^{P} \qquad d_i \in \{1, 2, \ldots, k\}$

  – Linear $\quad y(\mathbf{x}) = f(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{N} w_j x_j = \mathbf{w}^T \mathbf{x}$

  – Linear-in the parameter, but non-linear $\quad y(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x})$

  – Nonlinear $\quad y(\mathbf{x}) = f(\mathbf{x}, \mathbf{w})$



Linear classification

- Loss function (squared error) $l(d_i, f(\mathbf{x}_i, \mathbf{w})) = \left(d_i - f(\mathbf{w}, \mathbf{x}_i)\right)^2$
- Cross entropy $L\left(d_i, f\left(\mathbf{x}_i, \mathbf{w}\right)\right) = -\left(d_i \ln f\left(\mathbf{x}_i, \mathbf{w}\right) + \left(1 - d_i\right) \ln\left(1 - f\left(\mathbf{x}_i, \mathbf{w}\right)\right)\right)$
- (empirical) risk functional
  - LS case $R_E(\mathbf{w}) = \frac{1}{P} \sum_{i=1}^{P} \left(d_i - f(\mathbf{w}, \mathbf{x}_i)\right)^2$

$$R_E(\mathbf{w}) = (\mathbf{d} - \mathbf{X}\mathbf{w})^T (\mathbf{d} - \mathbf{X}\mathbf{w}) \qquad \mathbf{d} = \mathbf{X}\mathbf{w} \qquad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_P^T \end{bmatrix}$$

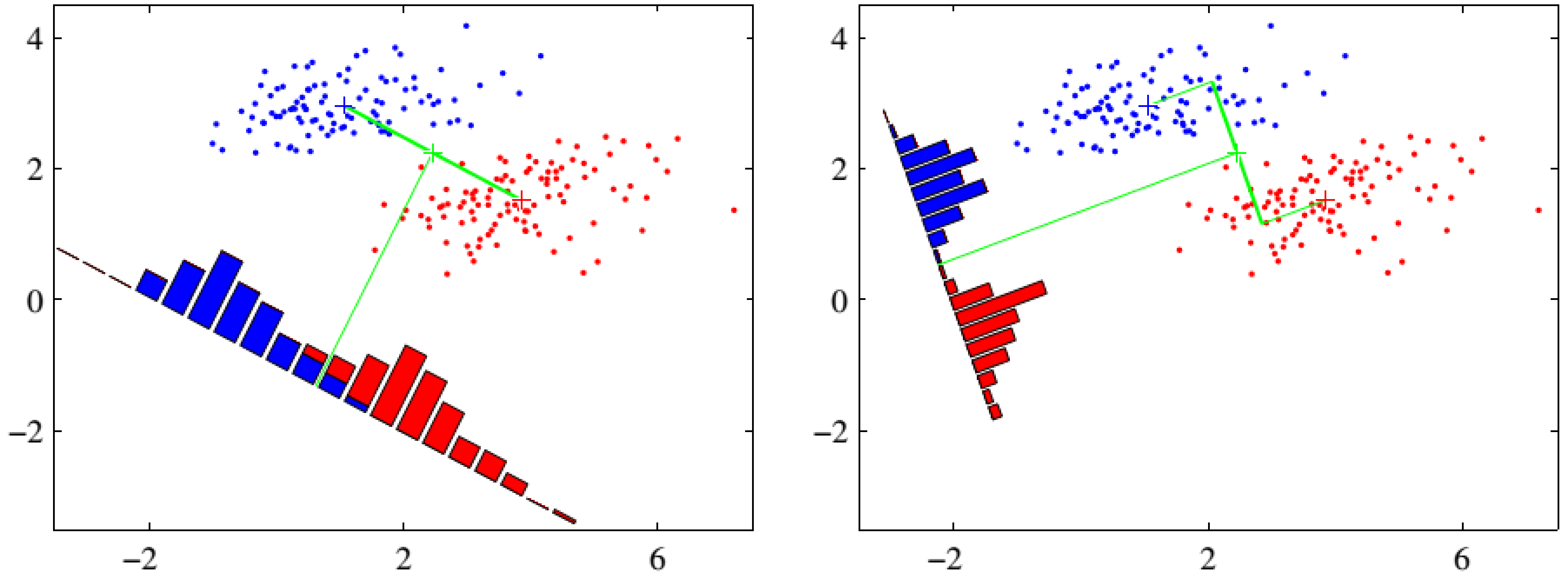
$$\mathbf{w}^* = \mathbf{X}^{\dagger}\mathbf{d} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{d}$$

Analytic solution
Iterativ solution

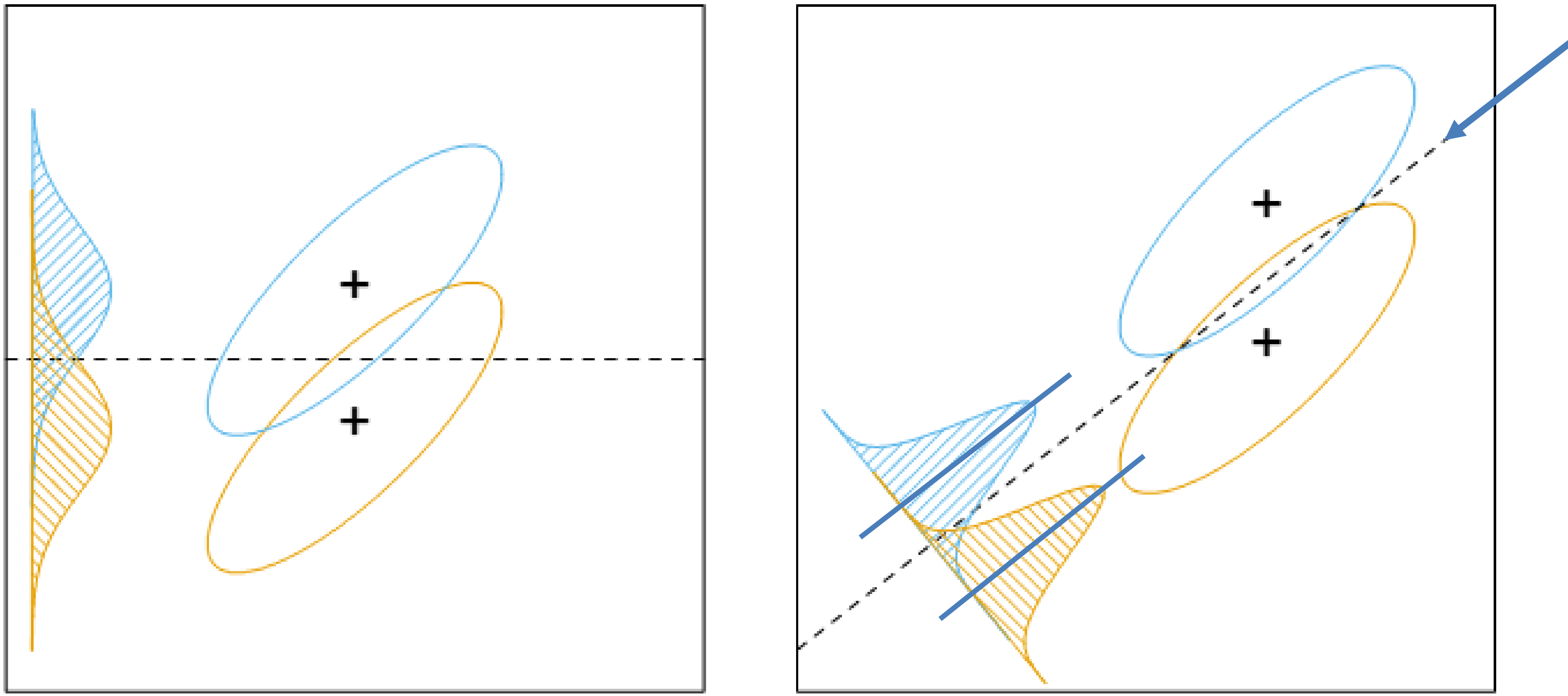
The role of regularization...

Linear binary classification

- Finding an optimal projection direction



LDA optimal projection direction



LDA as a dimension reduction method

LDA - KDA

$$C(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}.$$ to be maximized

$$m_2 - m_1 = \mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1) \qquad \mathbf{m}_i = \frac{1}{l_i}\sum_{n=1}^{l_i} \mathbf{x}_n^i, \qquad s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2$$

$$(m_2 - m_1)^2 = \mathbf{w}^T (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w} = \mathbf{w}^T \mathbf{S}_B \mathbf{w}$$

$$s_1^2 + s_2^2 = \sum_{i \in C^{(1)}} \mathbf{w}^T (\mathbf{x}_i - \mathbf{m}_1)(x_i - \mathbf{m}_1)^T \mathbf{w} + \sum_{i \in C^{(2)}} \mathbf{w}^T (\mathbf{x}_i - \mathbf{m}_2)(x_i - \mathbf{m}_2)^T \mathbf{w} =$$

$$= \mathbf{w}^T \mathbf{S}_{W1} \mathbf{w} + \mathbf{w}^T \mathbf{S}_{W2} \mathbf{w} = \mathbf{w}^T (\mathbf{S}_{W1} + \mathbf{S}_{W2}) \mathbf{w} = \mathbf{w}^T \mathbf{S}_W \mathbf{w}$$

$$C(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}} \mathbf{S}_B \mathbf{w}}{\mathbf{w}^{\mathrm{T}} \mathbf{S}_W \mathbf{w}},$$

LDA optimal parameter vector

$$\mathbf{w} = \arg\min_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} = \frac{\mathbf{w}^T \mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{w}}$$

Rayleigh quotient

$$\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w} = \lambda \mathbf{w} \quad \text{but} \quad \text{from} \quad \mathbf{w} \mathbf{S}_B \mathbf{w} = \mathbf{w} \left(\mathbf{m}_2 - \mathbf{m}_1 \right) \left(\mathbf{m}_2 - \mathbf{m}_1 \right)^T$$

$$\text{so} \;\; \mathbf{S}_B \mathbf{w} = \left(\mathbf{m}_2 - \mathbf{m}_1 \right) \left(\mathbf{m}_2 - \mathbf{m}_1 \right)^T \mathbf{w}$$

$$\mathbf{S}_W^{-1} \left(\mathbf{m}_2 - \mathbf{m}_1 \right) \propto \mathbf{w}$$

Linear classification

- Probabilistic (generative model)
  - ML solution
  - **Bayes solution** prior $\rightarrow$ posterior

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

$$= \frac{1}{1+\exp(-a)} = \sigma(a)$$

Logistic regression

$$a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \qquad \sigma(a) = \frac{1}{1+\exp(-a)}$$

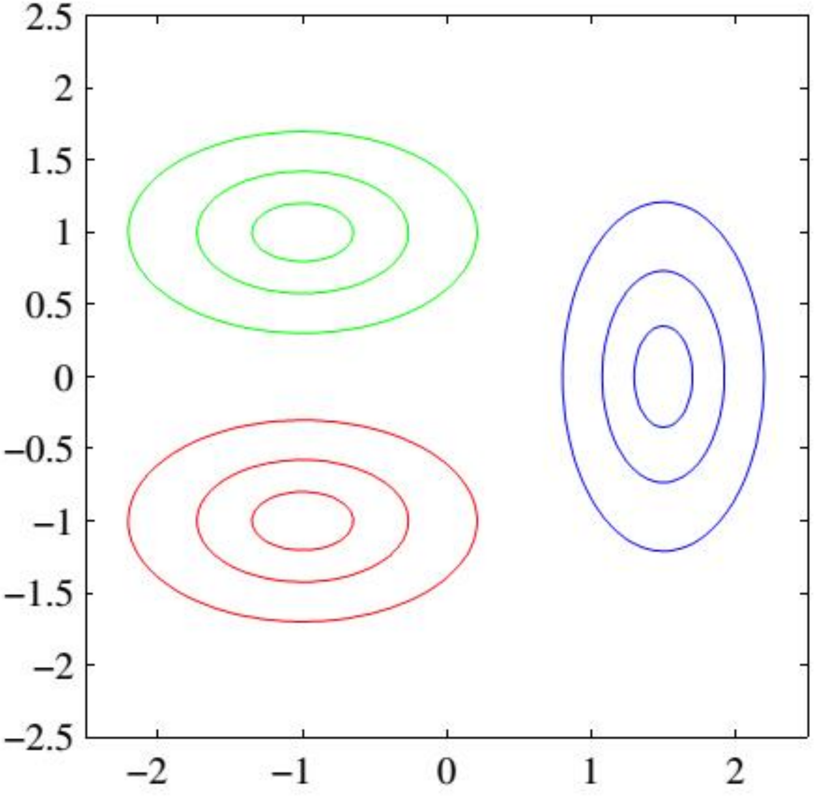
Logistic regression

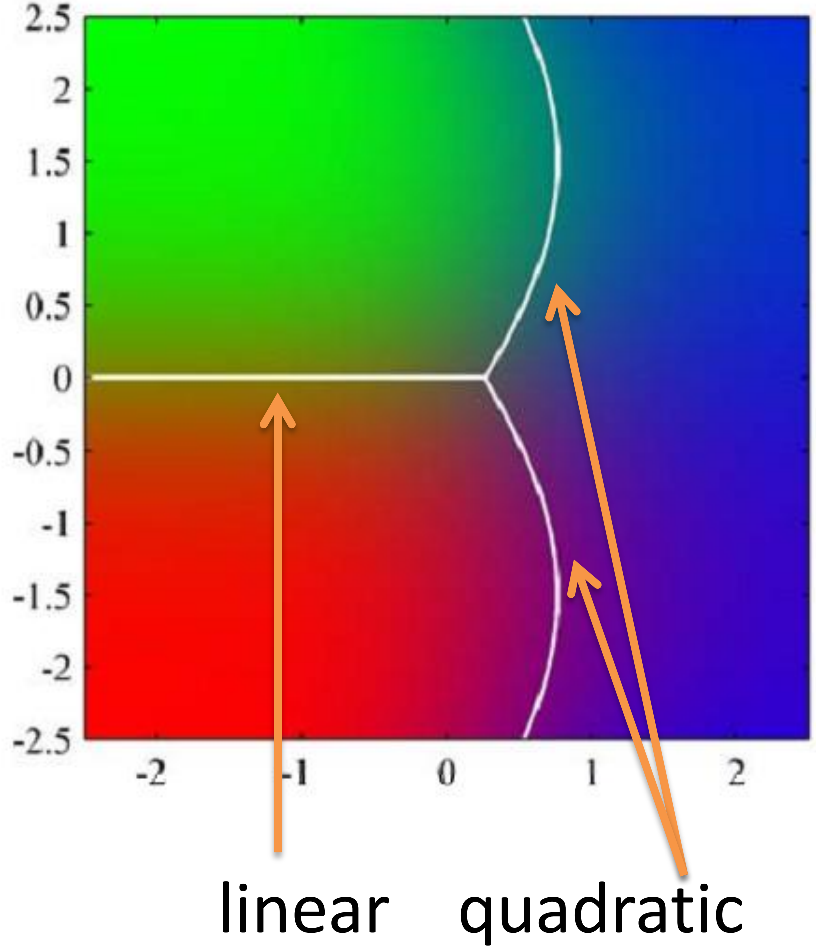
$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}.$$

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0) = \sigma(a)$$

$$\begin{aligned}
\mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}.
\end{aligned}$$

Separating surface for Gaussian distribution





Nonlinear classification

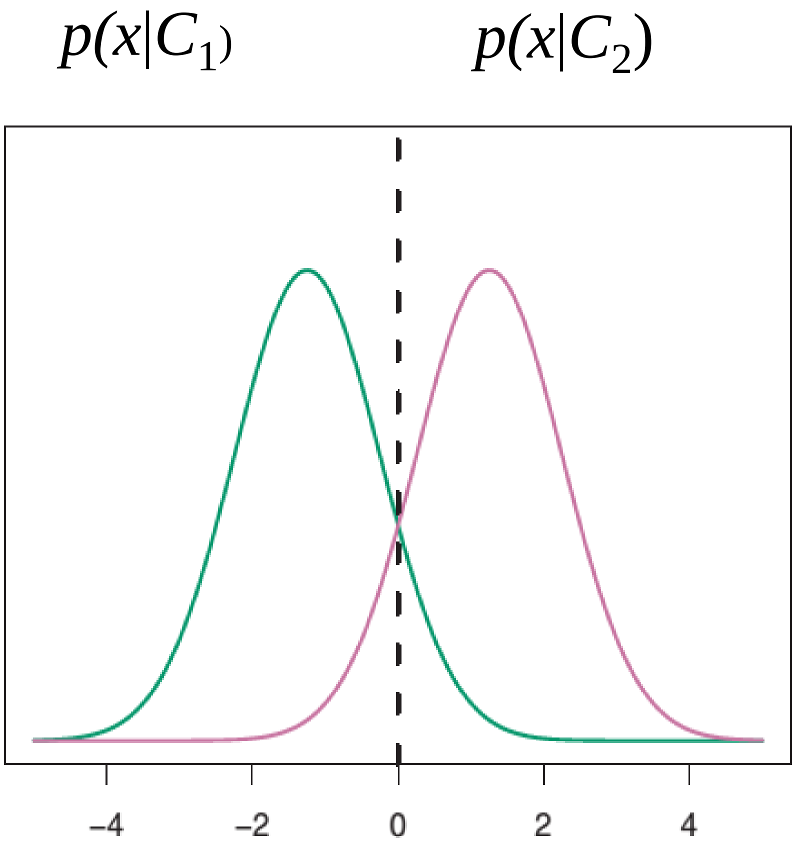
- Basis function approach
- Support vector machines
- Neural networks
  - Classical MLP
  - Deep network

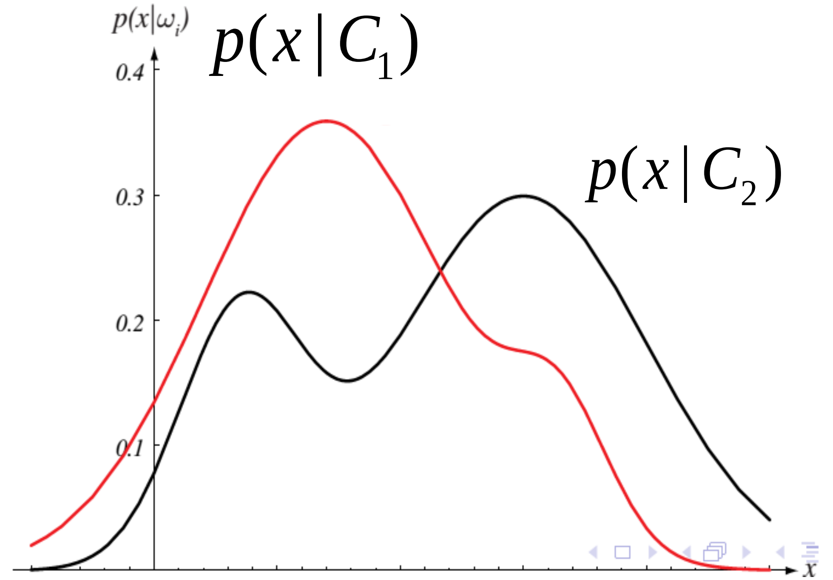
Decision theory
(Probabilistic classification)

- Many different application fields (e.g. Medical diagnostic) Decision rule based on measurements:
  - Likelihood function Conditional density function of the measurements
  - Two-class classification $P(x|C_1), \quad P(x|C_2)$
  - Input space: space of measurement (observation) data
  - Decision rule: separation of the measurement (observation) space Single-dimenzional case comparing to a threshold value
  - Bayes only on the likelihood functions

$$P(x|C_1) \underset{C_1}{\overset{C_2}{\lesseqgtr}} P(x|C_2)$$

ML decision





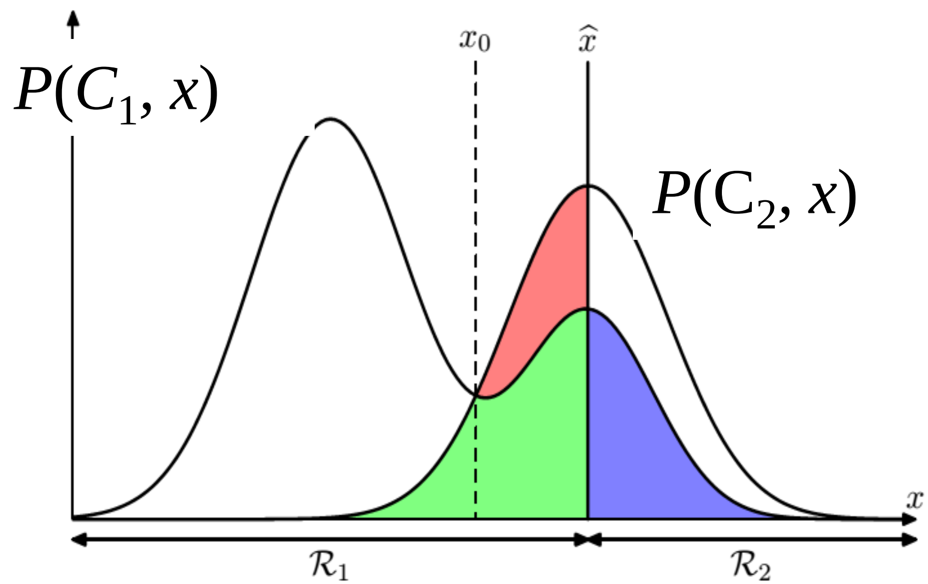
Bayes decision

- Bayes decision (based on the posteriors)

$$P(C_1|x) \underset{C_2}{\overset{C_1}{\gtreqless}} P(C_2|x) \qquad P(C_1|x) = \frac{p(x|C_1)P(C_1)}{p(x)} = \frac{p(x,C_1)}{p(x)} = \frac{p(x|C_1)P(C_1)}{\sum_{i=1,2} p(x|C_i)P(C_i)}$$

- Bayes rule

$$\frac{p(x|C_1)P(C_1)}{p(x)} \underset{C_2}{\overset{C_1}{\gtreqless}} \frac{p(x|C_2)P(C_2)}{p(x)}$$



The validation of a decision

- The probability of the incorrect decisions

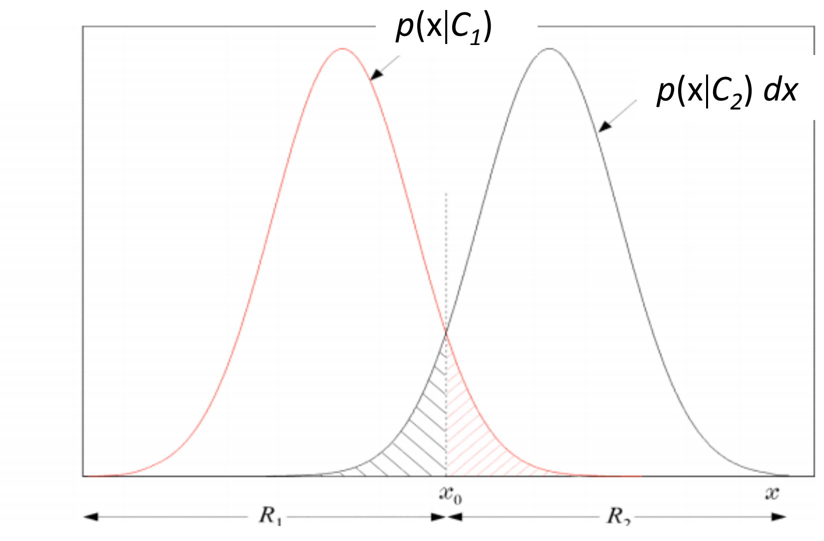
$$P(\text{error} \mid C_2) = \int_{-\infty}^{x_0} p(x|C_2)\, dx \quad \text{and} \quad P(\text{error} \mid C_1) = \int_{x_0}^{\infty} p(x|C_1)\, dx$$

If $P[C_1] = P[C_2]$

$$P_e \triangleq P(\text{error})$$

$$= P[C_2]\, P(\text{error} \mid C_2) + P[C_1]\, P(\text{error} \mid C_1)$$

$$= \frac{1}{2}\int_{-\infty}^{x_0} p(x|C_2)\, dx + \frac{1}{2}\int_{x_0}^{+\infty} p(x|C_1)\, dx$$



Statistical decision

Bayesian cost

$$\mathfrak{R} \quad \mathscr{R} = K_{11}P_1\int_{R_1} p(x|C_1)dx + K_{12}P_2\int_{R_1} p(x|C_2)dx +$$

$$+K_{21}P_1\int_{R_2} p(x|C_1)dx + K_{22}P_2\int_{R_2} p(x|C_2)dx$$

Using ...

$$\mathfrak{R} = K_{21}P_1 + C_{22}P_2 + (K_{12}-K_{22})P_2\int_{R_1} p(x|C_2)dx - (K_{21}-K_{11})P_1\int_{R_1} p(x|C_1)dx$$

Decision areas: to minimize the average cost

$$R_1 = \arg\min \int_{R_1} \left\{(K_{12}-K_{22})P_2\, p(x|C_2) - (K_{21}-K_{11})P_1\, p(x|C_1)\right\}dx$$

$$(K_{12}-K_{22})P_2\, p(x|C_2) \underset{C_1}{\overset{C_2}{\substack{> \\ = \\ <}}} (K_{21}-K_{11})P_1\, p(x|C_1)$$

$$\frac{p(x|C_1)}{p(x|C_2)} \underset{\geq}{\leq} \frac{(K_{12}-K_{22})P_2}{(K_{21}-K_{11})P_1}$$

Statisztikai döntés

- Likelihood ratio test

$$\Lambda(x) = \frac{P(C_1)}{P(C_2)} \underset{C_2}{\overset{C_1}{\substack{> \\ = \\ <}}} \eta$$

Naiv decision $\eta = 1$

$$\Lambda(x) = \frac{p(x|C_1)}{p(x|C_2)} \underset{C_2}{\overset{C_1}{\substack{> \\ = \\ <}}} \eta$$

Based on the Likelihood function $\eta = 1$

$$\Lambda(x) = \frac{p(x|C_1)}{p(x|C_2)} \underset{C_2}{\overset{C_1}{\substack{> \\ = \\ <}}} \eta = \frac{P(C_2)}{P(C_1)}$$

Bayes decision $\eta = \frac{P(C_2)}{P(C_1)}$

$$\Lambda(x) = \frac{p(x|C_1)}{p(x|C_2)} \underset{C_2}{\overset{C_1}{\substack{> \\ = \\ <}}} \eta = \frac{(K_{12} - K_{22})P(C_2)}{(K_{21} - K_{11})P(C_1)}$$

A Bayes decision using cost values

$$\eta = \frac{(K_{12} - K_{22})P(C_2)}{(K_{21} - K_{11})P(C_1)}$$

Validation of a binary classification

- Possible outcomes (medical diagnosis)

reality / decision	healthy	sick
healthy	True negative (TN)	False negativ (FN) (missed)
sick	False positive (FP) (False alarm)	True positive (TP)

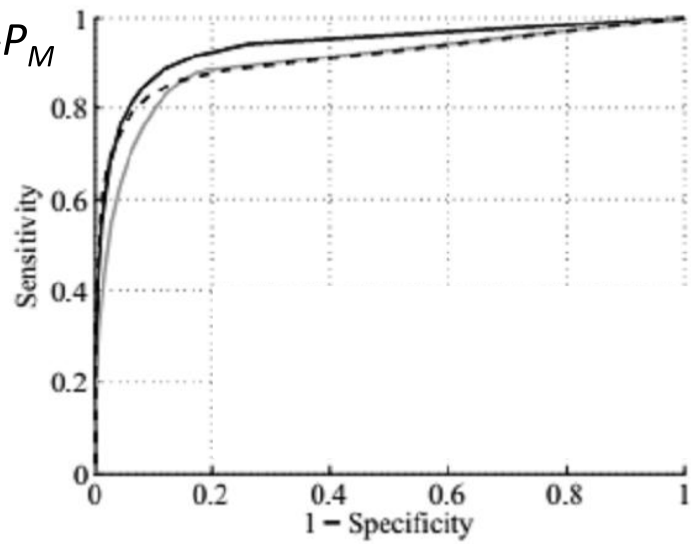
ROC (Receiver Operating Characteristic)

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

P_F ; P_M False alarm, Missed

1-P_M



P_F

AUC