INF 4300
04.12.13

Repetition - classification
Anne Solberg (anne@ifi.uio.no)

Bayes rule for a classification problem

- Suppose we have J, j=1,...J classes. ω is the class label for a pixel, and ***x*** is the observed feature vector).
- We can use Bayes rule to find an expression for the class with the highest probability:

$$P(\omega_j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_j) P(\omega_j)}{p(\mathbf{x})}$$

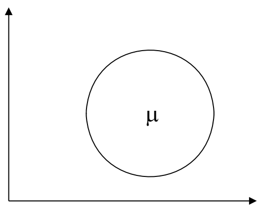
$$\text{posterior probability} = \frac{\textit{likelihood} \times \text{prior probability}}{\text{normalizing factor}}$$

- $P(\omega_j)$ is the prior probability for class ω_j. If we don't have special knowledge that one of the classes occur more frequent than other classes, we set them equal for all classes. ($P(\omega_j)=1/J$, j=1.,,,J).

Euclidean distance vs. Mahalanobis distance

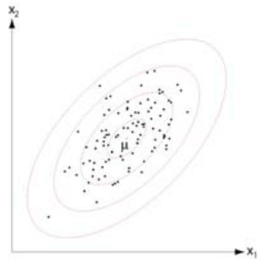
- Euclidean distance between point x and class center μ:

$$(x-\mu)^T(x-\mu) = \|x-\mu\|^2$$



- Mahalanobis distance between x and μ:

$$r^2 = (x-\mu)^T \Sigma^{-1} (x-\mu)$$



Discriminant functions for the normal density

- We saw that the minimum-error-rate classification can computed using the discriminant functions

$$g_i(\mathbf{x}) = \ln p(\mathbf{x} \mid \omega_i) + \ln P(\omega_i)$$

- With a multivariate Gaussian we get:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

- Let ut look at this expression for some special cases:

Case 1: $\Sigma_j=\sigma^2 I$

- An equivalent formulation of the discriminant functions:

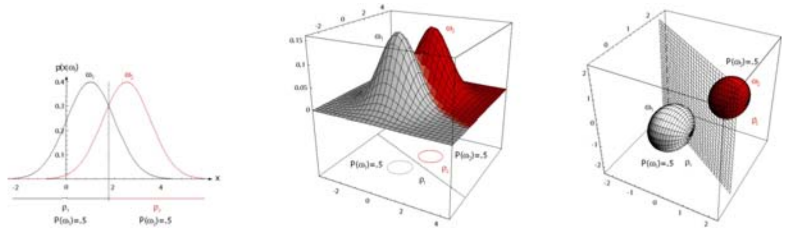
$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_i0$$
$$\text{where } \mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i \text{ and } wi0 = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln P(\omega_i)$$

- The equation $g_i(\mathbf{x})=g_j(\mathbf{x})$ can be written as

$$\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0)=0$$
$$\text{where } \mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$$
$$\text{and } x_0 = \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}\ln\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

- **w**=μ_i-μ_j is the vector between the mean values.
- This equation defines a hyperplane through the point x_0, and orthogonal to **w**.
- If $P(\omega_i)=P(\omega_j)$ the hyperplane will be located halfway between the mean values.

A simple model, $\Sigma_j=\sigma^2 I$



- The distributions are spherical in *d* dimensions.
- The decision boundary is a generalized hyperplane of *d-1* dimensions
- The decision boundary is perpendicular to the line separating the two mean values
- This kind of a classifier is called a linear classifier, or a linear discriminant function
  - Because the decision function is a linear function of ***x***.
- If $P(\omega_i)= P(\omega_j)$, the decision boundary will be half-way between μ_i and μ_j

Case 2: Common covariance, $\Sigma_j= \Sigma$

- If we assume that all classes have the same shape of data clusters, an intuitive model is to assume that their probability distributions have the same shape
- By this assumption we can use all the data to estimate the covariance matrix
- This estimate is common for all classes, and this means that also in this case the discriminant functions become linear functions

$$g_j(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \ln P(\omega_j)$$
$$= -\frac{1}{2(\sigma^2 I)}(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - 2\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \ln P(\omega_j)$$

Common for all classes, no need to compute
Since $\boldsymbol{x}^T\boldsymbol{x}$ is common for all classes, $g_j(\boldsymbol{x})$ again reduces to a linear function of ***x***.

Case 2: Common covariance, $\Sigma_j= \Sigma$

- An equivalent formulation of the discriminant functions is

$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + wi_0$$
$$\text{where } \mathbf{w}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$$
$$\text{and } wi_0 = -\frac{1}{2}\boldsymbol{\mu}_i{}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i)$$

- The decision boundaries are agian hyperplanes.
- Because $\mathbf{w}_i= \boldsymbol{\Sigma}^{-1}(\mu_i- \mu_j)$ is not in the direction of $(\mu_i- \mu_j)$, the hyperplan wil not be orthogonal to the line between the means.

Case 3:, Σ_j=arbitrary

- The discriminant functions will be quadratic:

$$g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + wi_0$$

$$\text{where } \mathbf{W}_i = -\frac{1}{2}\mathbf{\Sigma}_i^{-1}, \quad \mathbf{w}_i = \mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

$$\text{and } wi_0 = -\frac{1}{2}\boldsymbol{\mu}_i{}^t\mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln\left|\mathbf{\Sigma}_i\right| + \ln P(\omega_i)$$

- The decision surfaces are hyperquadrics and can assume any of the general forms:
  - hyperplanes
  - hypershperes
  - pairs of hyperplanes
  - hyperellisoids,
  - hyperparaboloids
  - hyperhyperboloid

k-Nearest-Neighbor classification

- A very simple classifier.
- Classification of a new sample x_j is done as follows:
  - Out of N training vectors, identify the *k* nearest neighbors (measure by Euclidean distance) in the training set, irrespectively of the class label. *k* should be odd.
  - Out of these *k* samples, identify the number of vectors k_i that belong to class ω_i , *i:1,2,....M* (if we have *M* classes)
  - Assign x_j to the class ω_i with the maximum number of k_i samples.
- *k* must be selected a priori.