

Data Mining (Mineria de Dades)

Lluís A. Belanche
belanche@lsi.upc.edu

Soft Computing Research Group

Dept. de Llenguatges i Sistemes Informàtics (Software department)

Universitat Politècnica de Catalunya

2011-2012

Data Mining (Part II)

Contents

1. Association Rules
2. Bayesian decision theory
3. Bayesian classifiers: LDA/QDA
4. Bayesian classifiers: Naïve Bayes
5. Bayesian classifiers: Nearest Neighbours
6. Neural Networks: Multilayer Perceptrons
7. Support Vector Machines
8. Ensemble Learning

Bayesian decision theory (II)

Normally distributed features

A continuous r.v. X is normally distributed, written $X \sim N(\mu, \sigma^2)$, when its pdf is:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$

where μ is the *mean* and σ^2 is the *variance*:

- $E[X] = \int_{\mathbb{R}} xp(x) dx = \mu$
- $E[(X - \mu)^2] = \int_{\mathbb{R}} (x - \mu)^2 p(x) dx = \sigma^2$

Bayesian decision theory (II)

Normally distributed features

A continuous n -variate random vector $X = (X_1, \dots, X_n)^t$ is normally distributed, written $X \sim N(\boldsymbol{\mu}, \Sigma)$, when its joint pdf is:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

where $\boldsymbol{\mu}$ is the *mean vector* and $\Sigma_{n \times n} = (\sigma_{ij}^2)$ is the (real symmetric and p.s.d.) *covariance matrix*.

- $E[X] = \boldsymbol{\mu}$ and $E[(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^t] = \Sigma$.
- $CoVar[X_i, X_j] = \sigma_{ij}^2$ and $Var[X_i] = \sigma_{ii}^2$

Note that if $X \sim N(\boldsymbol{\mu}, \Sigma)$, then X_i, X_j are statistically independent $\iff CoVar[X_i, X_j] = 0$ (in general, only the left-to-right implication holds).

Bayesian decision theory (II)

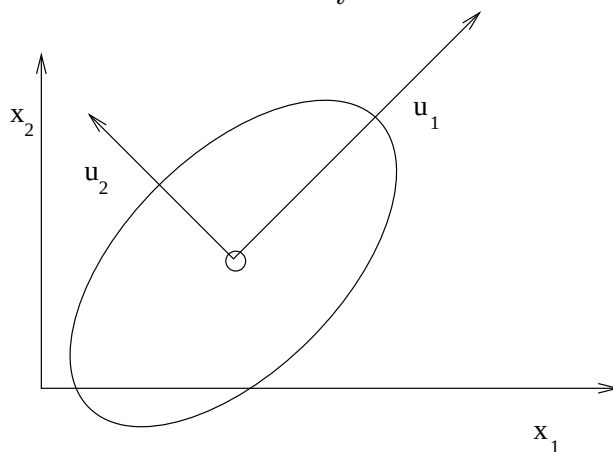
Normally distributed features

- The shape of the decision regions defined by a Bayes classifier depends on the assumed form for $p(\mathbf{x}|\omega_k)$.
- What is behind the choice of a multivariate normal? Examples from a class are noisy versions of an ideal class member (a *prototype*):
 - Prototype: modeled by the mean vector
 - Noise: modeled by the covariance matrix
- Very important: the number of parameters of the distribution is $\frac{n(n+1)}{2} + n$.

Bayesian decision theory (II)

Normally distributed features

- The quantity $d(x, \mu) = \sqrt{(x - \mu)^t \Sigma^{-1} (x - \mu)}$ is called the *Mahalanobis distance* from x to the distribution to which it belongs.
- The surfaces of equal probability $d(x, \mu) = ct$ are *hyperellipsoids*
- The principal directions (PC) of the hyperellipsoids are given by the *eigenvectors* u_i of Σ , which satisfy $\Sigma u_i = \lambda_i u_i$.
- The lengths of the hyperellipsoids along these axes are proportional to $\sqrt{\lambda_i}$, where λ_i are the *eigenvalues* associated with u_i .



Bayesian decision theory (II)

Properties of the Gaussian (normal) distribution

- **Simplified forms** are useful: e.g., if the X_i are statistically independent, then $p(\mathbf{x}) = \prod_{i=1}^n p(x_i)$, Σ is diagonal and the PCs are aligned with the coordinate axes.
- **Simple analytical properties**, e.g. any moment can be expressed as a function of μ and Σ .
- **Central limit theorem**, the mean of M random variables tends to a normal distribution, in the limit of infinite M .
- **Linear transformation invariance**, the distribution of a linear transformation of the coordinate system remains normal.
- **Marginal densities** (integrating out some variables) and **conditional densities** (setting some variables to fixed values) are also normal.
- **Nature processes?**

Bayesian decision theory (II)

Discriminant functions for the Gaussian density

- We showed that the Bayes rule minimizing the probability of error could be formulated in terms of a family of discriminant functions:

“assign the feature vector $\mathbf{x}$ to the class ω_k whenever $g_k(\mathbf{x}) > g_i(\mathbf{x})$, $\forall k \neq i$ ”

- For normally distributed classes, i.e., $X_{|\Omega=\omega_k} \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, this family can be reduced to very simple expressions.

Using Bayes rule and the natural log, the discriminant function for class ω_k becomes:

$$g_k(\mathbf{x}) = \ln \{P(\omega_k)p(\mathbf{x}|\omega_k)\} = \ln P(\omega_k) - \ln \{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}\} - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^t \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)$$

Eliminating constant terms:

$$g_k(\mathbf{x}) = \ln P(\omega_k) - \frac{1}{2} (\ln |\boldsymbol{\Sigma}_k| + (\mathbf{x} - \boldsymbol{\mu}_k)^t \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k))$$

- This expression is called a **quadratic discriminant function**, and the **decision boundaries** $g_i(\mathbf{x}) = g_j(\mathbf{x})$ are general quadratic forms in n -dimensional space

Bayesian decision theory (II)

Discriminant functions for the Gaussian density

If we assume that all class-conditional distributions $p(\mathbf{x}|\omega_k)$ have the *same* covariance matrix Σ , we get:

$$g_k(\mathbf{x}) = \ln P(\omega_k) + \boldsymbol{\mu}_k^t \Sigma^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_k^t \Sigma^{-1} \boldsymbol{\mu}_k$$

Reorganizing terms we obtain:

$$g_k(\mathbf{x}) = \mathbf{w}_k^t \mathbf{x} + w_{k0}$$

where $\mathbf{w}_k = \Sigma^{-1} \boldsymbol{\mu}_k$ and $w_{k0} = -\frac{1}{2} \boldsymbol{\mu}_k^t \Sigma^{-1} \boldsymbol{\mu}_k + \ln P(\omega_k)$.

- These are **linear discriminant functions** (linear in $\mathbf{x}$) and the **decision boundaries** $g_i(\mathbf{x}) = g_j(\mathbf{x})$ are hyperplanes in n -dimensional space.

Bayesian decision theory (II)

Discriminant functions for the Gaussian density

If we further assume that all the X_i, X_j are statistically independent, that is $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$, we get:

$$g_k(\mathbf{x}) = \ln P(\omega_k) - \frac{1}{2} \sum_{i=1}^n \frac{(\mu_{ki} - x_i)^2}{\sigma_i^2}$$

If we further assume that all the X_i have the same variance σ^2 , that is $\Sigma = \sigma^2 I_n$, we get:

$$g_k(\mathbf{x}) = \ln P(\omega_k) - \frac{1}{2\sigma^2} \|\boldsymbol{\mu}_k - \mathbf{x}\|^2$$

If we further assume that all the classes have the same prior $P(\omega_k) = \frac{1}{c}$, we get:

$$g_k(\mathbf{x}) = -\|\boldsymbol{\mu}_k - \mathbf{x}\|^2$$

Bayesian decision theory (II)

Discriminant functions for the Gaussian density

In all cases, we have a *minimum-distance* classifier:

- In the general case (some covariance matrices are different), the classifier uses the Mahalanobis distance (a fully-weighted Euclidean distance) from x to all the classes.
The technique is called **quadratic discriminant analysis** (QDA)
- In case all covariance matrices are equal, the classifier uses the Mahalanobis distance (a partially-weighted Euclidean distance) from x to all the classes.
The technique is called **linear discriminant analysis** (LDA)
- In case all covariance matrices are diagonal, the classifier uses a simply-weighted Euclidean distance from x to all the classes.
- In case all covariance matrices are a multiple of the identity matrix, the classifier uses an unweighted Euclidean distance from x to all the classes.

Bayesian decision theory (II)

Computations in practice (I)

In practical situations, only an i.i.d data sample S is available. Let $S_k \subset S$ be the subset of observations known to belong to class ω_k . Then $S_1, \dots, S_c$ is a partition of S .

We have unbiased estimates for the vector means and for the class priors:

$$\hat{\mu}_k = \frac{1}{|S_k|} \sum_{\mathbf{x} \in S_k} \mathbf{x}; \quad \hat{P}(\omega_k) = \frac{|S_k|}{|S|}$$

1. If we know (or assume) that covariance matrices are different (wish to use QDA):

$$\hat{\Sigma}_k = \frac{1}{|S_k| - 1} \sum_{\mathbf{x} \in S_k} (\mathbf{x} - \hat{\mu}_k)(\mathbf{x} - \hat{\mu}_k)^t$$

2. If we know (or assume) that covariance matrices are equal (wish to use LDA):

$$\hat{\Sigma}_{\text{pooled}} = \frac{1}{|S| - c} \sum_{k=1}^c (|S_k| - 1) \hat{\Sigma}_k$$

Bayesian decision theory (II)

Computations in practice (II)

- Once we use sample statistics instead of population parameters, we lose optimality!
- If the class-conditional densities are far from being Gaussian, the model will be poor
- Even when the class-conditional densities are Gaussian, the sample statistics need to be estimated reliably (this is particularly acute for QDA)
- If we use QDA, we need at least $|S_k| > n$ for the matrix inversions, $1 \leq k \leq c$

Bayesian decision theory (II)

Pros & Cons

- (↓) Assumption of gaussianity may be far from true
- (↓) Needs a lot of data per class if will to use QDA
- (↓) Requires matrix inversion(s), which may be costly or numerically delicate
- (↑) Adaptable to all class-conditional distributions (not only gaussian), even with mixed variables
- (↑) Very resistant to overfitting the data sample
- (↑) Accepts class priors and (asymmetric) losses for misclassifications

Bayesian decision theory (II)

Computing the probability of error

We give now the exact formula for the true probability of error for the case of two-class normally distributed features with equal covariance matrix Σ and unequal class priors $P(\omega_1), P(\omega_2)$:

$$p(error) = P(\omega_1) \operatorname{erf} \left(\frac{\alpha - D/2}{\sqrt{D}} \right) + P(\omega_2) \left[1 - \operatorname{erf} \left(\frac{\alpha + D/2}{\sqrt{D}} \right) \right]$$

where:

- $\operatorname{erf}(t) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^t \exp(-t^2/2) dt$
- $D = (\mu_1 - \mu_2)^t \Sigma^{-1} (\mu_1 - \mu_2)$
- $\alpha = \ln P(\omega_2) - \ln P(\omega_1)$

Bayesian decision theory (II)

A numerical example (I)

Derive a linear discriminant function for the two-class classification problem defined by the following Gaussian class-conditional densities:

$$\boldsymbol{\mu}_1 = (0, 0, 0)^t, \boldsymbol{\mu}_2 = (1, 1, 1)^t, \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \text{diag}(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}), P(\omega_2) = 2P(\omega_1)$$

Solution: since all the X_i, X_j are statistically independent ($i \neq j$) and have the same variance $\sigma^2 = \frac{1}{4}$, that is $\boldsymbol{\Sigma} = \frac{1}{4}I$, we get:

$$g_1(\mathbf{x}) = \ln P(\omega_1) - \frac{1}{2} \frac{\|\boldsymbol{\mu}_1 - \mathbf{x}\|^2}{\sigma^2} = \ln \frac{1}{3} - \frac{1}{2} \frac{\|(0, 0, 0) - \mathbf{x}\|^2}{\frac{1}{4}}$$
$$g_2(\mathbf{x}) = \ln P(\omega_2) - \frac{1}{2} \frac{\|\boldsymbol{\mu}_2 - \mathbf{x}\|^2}{\sigma^2} = \ln \frac{2}{3} - \frac{1}{2} \frac{\|(1, 1, 1) - \mathbf{x}\|^2}{\frac{1}{4}}$$

Bayesian decision theory (II)

A numerical example (II)

Then we can build an optimal *dichotomizer*:

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) \begin{matrix} \omega_1 \\ > \\ \omega_2 \\ < \end{matrix} 0, \quad \mathbf{x} = (x_1, x_2, x_3)^t$$

$$g(\mathbf{x}) = -\ln 2 - 2\|\mathbf{x}\|^2 + 2\|(1, 1, 1) - \mathbf{x}\|^2 \begin{matrix} \omega_1 \\ > \\ \omega_2 \\ < \end{matrix} 0$$

Which results in: $(x_1 + x_2 + x_3) \begin{matrix} \omega_2 \\ > \\ \omega_1 \\ < \end{matrix} \frac{3}{2} - \frac{1}{4} \ln 2$

The prediction for the test example $\mathbf{x}^* = (0,1,0,7,0,8)^t$ is $\mathbf{x}^* \in \omega_2$, given that $0,1 + 0,7 + 0,8 = 1,6 > \frac{3}{2} - \frac{1}{4} \ln 2 \approx 1,32$

Bayesian decision theory (II)

Discussion

- The Bayes classifier is the best possible classifier when the class-conditional densities and priors are known
- The Bayes classifier for normally distributed classes is a quadratic classifier
- The Bayes classifier for normally distributed classes with equal covariance matrices is a linear classifier
- These classifiers can be derived from decision-theoretic principles and some simplifying assumptions
- Therefore, using a specific (Euclidean or Mahalanobis) distance function (metric) implicitly corresponds to certain statistical assumptions
- The question whether these assumptions hold can rarely be answered in practice; in most cases we are limited to posing and answering the question “does this classifier give satisfactory predictions or not?”