ICS663: Pattern Recognition

Department of Information and Computer Sciences
University of Hawai`i at Manoa

Kyungim Baek

Announcement

- Homework assignment #1
  - Due: **Monday September 21, by 5:00 PM**
- Project proposal
  - Due: **Wednesday September 23, by 5:00 PM**
  - Presentation (~ 10 minutes)
    - Monday (9/28)
      - BJ, Thomas, Tyson, Sharif, Danny, Jeremy
    - Wednesday (9/30)
      - Tetsuya, Kelly, Nurit

Previously…

- Discriminant function for the minimum-error-rate classification

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

- Case of multivariate normal, $p(\mathbf{x} \mid \omega_i) \sim N(\boldsymbol{\mu}_i, \Sigma_i)$:

$$p(\mathbf{x} \mid \omega_i) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right]$$

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

Previously…

- Case 1: $\Sigma_i = \sigma^2 \mathbf{I}$

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(\omega_i) = \mathbf{w}_i^t \mathbf{x} + w_{i0}$$

where $\mathbf{w}_i = \frac{\boldsymbol{\mu}_i}{\sigma^2}$; $\quad w_{i0} = -\frac{1}{2\sigma^2} \boldsymbol{\mu}_i^t \boldsymbol{\mu}_i + \ln P(\omega_i)$

- The hyperplane separating R_i and R_j (decision surface between R_i and R_j)

$$g_i(\mathbf{x}) = g_j(\mathbf{x}) \Rightarrow \mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0$$

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

Previously…

- Case 2: $\Sigma_i = \Sigma$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln P(\omega_i) = \mathbf{w}_i^t\mathbf{x} + w_{i0}$$

where $\mathbf{w}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$; $\quad w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i)$

- The hyperplane separating R_i and R_j (decision surface between R_i and R_j

$$g_i(\mathbf{x}) = g_j(\mathbf{x}) \Rightarrow \mathbf{w}^t(\mathbf{x}-\mathbf{x}_0) = 0$$

where $\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ and

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{1}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)} \ln\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$$

Lecture 5

- Discriminant Functions for the Normal Density
  - Case 3: Σ_i = Arbitrary
- Bayes Decision Theory for Discrete Features
  - Example: Independent Binary Features
- Parametric Density Estimation
  - Introduction

Case 3: Σ_i = Arbitrary

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

$$\Rightarrow \quad g_i(\mathbf{x}) = \mathbf{x}^t\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^t\mathbf{x} + \omega_{i0}$$

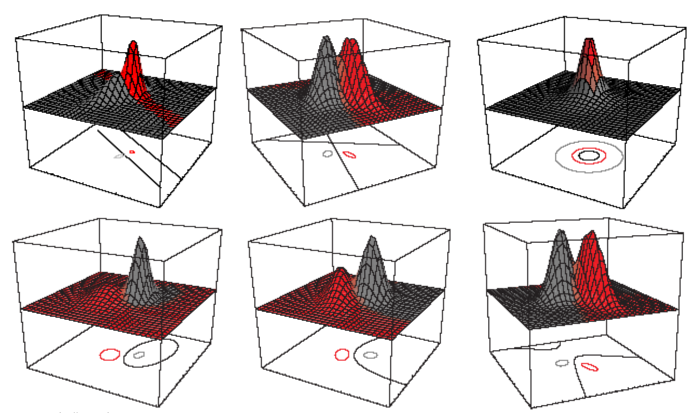
where

$$\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}$$

$$\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

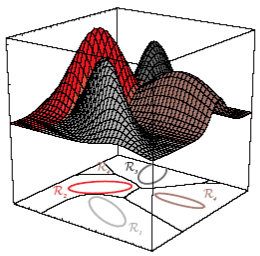
$$\omega_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

Σ_i = Arbitrary: Decision Surfaces



Σ_i = Arbitrary: Decision Surfaces

- Decision regions for four normal distributions



- See Example 1 in the textbook

Σ_i = Arbitrary: Example

- Samples:

$$\omega_1: \ (1,2)^t, (3,1)^t, (5,2)^t, (3,3)^t$$

$$\omega_2: \ (7,6)^t, (8,4)^t, (9,6)^t, (8,8)^t$$

- Compute sample mean and covariance for each class:

$$\boxed{\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \qquad \boldsymbol{\Sigma} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^t}$$

$$\boldsymbol{\mu}_1 = \frac{1}{4}\left(\begin{pmatrix}1\\2\end{pmatrix}+\begin{pmatrix}3\\1\end{pmatrix}+\begin{pmatrix}5\\2\end{pmatrix}+\begin{pmatrix}3\\3\end{pmatrix}\right)=\begin{pmatrix}3\\2\end{pmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix}8/3 & 0\\ 0 & 2/3\end{pmatrix}$$

$$\boldsymbol{\mu}_2 = \frac{1}{4}\left(\begin{pmatrix}7\\6\end{pmatrix}+\begin{pmatrix}8\\4\end{pmatrix}+\begin{pmatrix}9\\6\end{pmatrix}+\begin{pmatrix}8\\8\end{pmatrix}\right)=\begin{pmatrix}8\\6\end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix}2/3 & 0\\ 0 & 8/3\end{pmatrix}$$

Σ_i = Arbitrary: Example (cont'd)

- Discriminant functions:

$$g_1(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^t\boldsymbol{\Sigma}_1^{-1}\mathbf{x} + \boldsymbol{\mu}_1^t\boldsymbol{\Sigma}_1^{-1}\mathbf{x} + -\frac{1}{2}\boldsymbol{\mu}_1^t\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\ln|\boldsymbol{\Sigma}_1| + \ln P(\omega_1)$$

$$= -\frac{1}{2}\begin{pmatrix}x_1 & x_2\end{pmatrix}\begin{pmatrix}3/8 & 0\\ 0 & 3/2\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix} + \begin{pmatrix}3 & 2\end{pmatrix}\begin{pmatrix}3/8 & 0\\ 0 & 3/2\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix}$$

$$-\frac{1}{2}\begin{pmatrix}3 & 2\end{pmatrix}\begin{pmatrix}3/8 & 0\\ 0 & 3/2\end{pmatrix}\begin{pmatrix}3\\2\end{pmatrix} - \frac{1}{2}\ln\left|\begin{pmatrix}8/3 & 0\\ 0 & 2/3\end{pmatrix}\right| + \ln P(\omega_1)$$

$$= -(3/16)x_1^2 - (3/4)x_2^2 + (9/8)x_1 + 3x_2 - 75/16 - (1/2)\ln(16/9) + \ln P(\omega_1)$$

$$g_2(\mathbf{x}) = -\frac{1}{2}\begin{pmatrix}x_1 & x_2\end{pmatrix}\begin{pmatrix}3/2 & 0\\ 0 & 3/8\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix} + \begin{pmatrix}8 & 6\end{pmatrix}\begin{pmatrix}3/2 & 0\\ 0 & 3/8\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix}$$

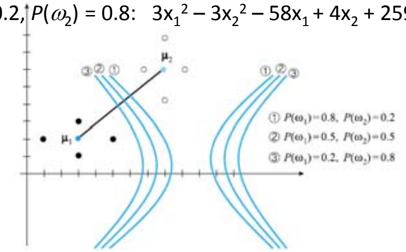
$$-\frac{1}{2}\begin{pmatrix}8 & 6\end{pmatrix}\begin{pmatrix}3/2 & 0\\ 0 & 3/8\end{pmatrix}\begin{pmatrix}8\\6\end{pmatrix} - \frac{1}{2}\ln\left|\begin{pmatrix}2/3 & 0\\ 0 & 8/3\end{pmatrix}\right| + \ln P(\omega_2)$$

$$= -(3/4)x_1^2 - (3/16)x_2^2 + 12x_1 + (18/8)x_2 - 219/4 - (1/2)\ln(16/9) + \ln P(\omega_2)$$

Σ_i = Arbitrary: Example (cont'd)

$$g_{12}(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) = \frac{9}{16}x_1^2 - \frac{9}{16}x_2^2 - \frac{87}{8}x_1 + \frac{3}{4}x_2 + \frac{801}{16} + \ln P(\omega_1) - \ln P(\omega_2)$$

- Decision boundaries: $g_{12}(\mathbf{x}) = 0$
  - $P(\omega_1) = 0.8$, $P(\omega_2) = 0.2$: $3x_1^2 - 3x_2^2 - 58x_1 + 4x_2 + 274.3936 = 0$
  - $P(\omega_1) = 0.5$, $P(\omega_2) = 0.5$: $3x_1^2 - 3x_2^2 - 58x_1 + 4x_2 - 267 = 0$
  - $P(\omega_1) = 0.2$, $P(\omega_2) = 0.8$: $3x_1^2 - 3x_2^2 - 58x_1 + 4x_2 + 259.6064 = 0$



Bayes Decision Theory: Discrete Features

- Components of **x** can take only one of *m* discrete values, $v_1, v_2, \ldots, v_m$
- Case of *independent binary features* in two category problem:
  - Let $\mathbf{x} = (x_1, x_2, \ldots, x_d)^t$ where each x_i is either 0 or 1, with probabilities:

$$p_i = P(x_i = 1 \mid \omega_1) \quad \text{and} \quad q_i = P(x_i = 1 \mid \omega_2)$$

$$P(\mathbf{x} \mid \omega_1) = P(x_1, x_2, \ldots, x_d \mid \omega_1) = \prod_{i=1}^{d} P(x_i \mid \omega_1) = \prod_{i=1}^{d} p_i^{x_i}(1-p_i)^{1-x_i}$$

$$P(\mathbf{x} \mid \omega_2) = P(x_1, x_2, \ldots, x_d \mid \omega_2) = \prod_{i=1}^{d} P(x_i \mid \omega_2) = \prod_{i=1}^{d} q_i^{x_i}(1-q_i)^{1-x_i}$$

Independent Binary Features

- Likelihood ratio:

$$\frac{P(\mathbf{x} \mid \omega_1)}{P(\mathbf{x} \mid \omega_2)} = \prod_{i=1}^{d} \left(\frac{p_i}{q_i}\right)^{x_i} \left(\frac{1-p_i}{1-q_i}\right)^{1-x_i}$$

- For minimum error-rate decision, the discriminant function in this case is:

$$g(\mathbf{x}) = \ln \frac{P(\mathbf{x} \mid \omega_1)}{P(\mathbf{x} \mid \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

$$= \sum_{i=1}^{d} \left[x_i \ln \frac{p_i}{q_i} + (1-x_i) \ln \frac{1-p_i}{1-q_i} \right] + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

Independent Binary Features

- The discriminant function is linear in the x_i:

$$g(\mathbf{x}) = \sum_{i=1}^{d} w_i x_i + w_0$$

where:

$$w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad i = 1, \ldots, d$$

$$w_0 = \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

decide ω_1 if $g(\mathbf{x}) > 0$ and ω_2 if $g(\mathbf{x}) \le 0$

Parametric Density Estimation

Recall: Bayes' Theorem

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

- Interpretation: $posterior = \dfrac{likelihood \cdot prior}{evidence}$
- Requirements for Bayesian classification
  - $P(\omega_i)$: prior probability distributions
  - $p(\mathbf{x} \mid \omega_i)$: class-conditional densities (likelihoods)

Why Learn?

- Bayesian decision theory assumes complete knowledge about the probabilistic structure of the problem
- But, in practice we have access only to *partial information*
- Forms of partial information
  - **Training data** D_n: often noisy, partially missing
  - **Prior knowledge** about probability distributions
  - **Dependence structure** in data
- Objective: Estimate $P(\omega_i \mid \mathbf{x}) \leftarrow P_n(\omega_i \mid \mathbf{x})$

from R. Meir

Estimation Problem

- **Data**: collection of samples to be classified into *c* classes $\omega_1, \omega_2, \ldots, \omega_c$, and divided into data sets (according to class) as follows $D_1, D_2, \ldots, D_c$ (assume supervised learning)
- **Needed**:
  - Prior probabilities $P(\omega_i)$
  - Class-conditional probabilities $p(\mathbf{x} \mid \omega_i)$

Estimation Problem (cont'd)

- **Estimation problem**:
  - Estimating $P(\omega_i)$ is not very difficult in the context of supervised learning. For example, assume

$$P(\omega_i) = \frac{\#\ of\ times\ class\ i\ appears\ in\ sample}{sample\ size}$$

  - Estimating $p(\mathbf{x} \mid \omega_i)$ is hard, especially in high-dimensional feature spaces and when the number of training samples available seems to be too small
    - Assumption: $p(\mathbf{x} \mid \omega_i) = p(\mathbf{x} \mid \omega_i, \theta)$ – *known* parametric form, $\exists$ a 'true' θ_0

Parametric Classification

Model + Data
↓
Parameter estimate
↓
Posterior probability distribution
↓
Classifier

$$p(\mathbf{x}\,|\,\omega_i,\theta) + D_n \xrightarrow{(1)} \theta_n \xrightarrow{(2)} P_n(\omega_i\,|\,\mathbf{x}) \xrightarrow{(3)} \alpha_n(\mathbf{x})$$

(1) θ_n parameter estimate
(2) $P_n(\omega_i|\mathbf{x})$ posterior estimate
(3) $\alpha_n(\mathbf{x})$ empirical classifier

from R. Meir

Parameter Estimation

- Simplifying assumptions:
  – Feature independence
  – Samples have been drawn independently (i.e. samples follow the i.i.d. model – independent and identically distributed random variables)

Estimation Methods

- **Maximum-Likelihood (ML) estimation**: *parameters of probabilistic distributions are fixed* but unknown values
  – Parameters are unknown constants to be identified through training
  – Best estimate of parameter values is achieved by maximizing the probability of obtaining the samples observed
- **Bayesian estimation**: *parameters are random variables* that follow a known (e.g. Gaussian) distribution
  – Parameters as random variables having some known prior distribution
  – By observing the behavior of the training samples, the posterior probabilities inferred help revising the parameter values
  – The more the training samples, the better the chances of refining the posterior probabilities, and subsequently, to peak the parameter values

from S. Iliescu

Estimation Methods (cont'd)

- Differences between ML and Bayesian estimation
  – Conceptual (philosophical): are parameters constants or random variables?
  – Different approaches, but usually leading to the same results (when sufficient training samples are available)

from S. Iliescu