Multivariate Methods

Berlin Chen

Graduate Institute of Computer Science & Information Engineering

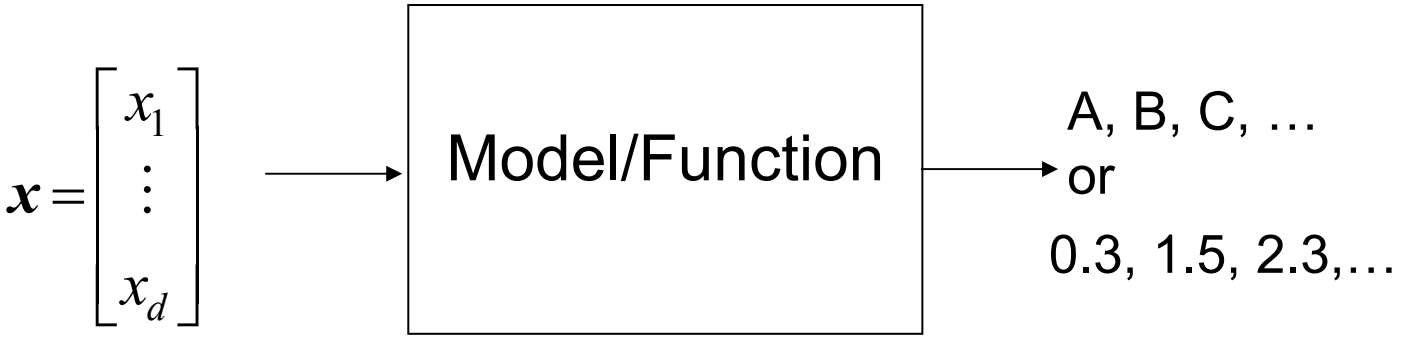
National Taiwan Normal University

References:

1. *Introduction to Machine Learning* , Chapter 5

Multivariate Methods

- Input: a data sample with multiple features (variables/ inputs)
- Output (of Prediction)
  - Classification: class code (discrete variable)
  - Regression: real number (continuous variable)



- Supervised learning
  - Model/function to be trained with labeled training samples

Multivariate Data (1/2)

- Each data sample is represented by an observation vector with d dimensions
  - Each dimension of the vector is termed input/feature/attribute
  - Dimensions with different types and value domains
- The whole data set of size N can be viewed as a data matrix

$$\boldsymbol{X}=\begin{bmatrix} X_1^1 & X_2^1 & \cdots & X_d^1 \\ X_1^2 & X_2^2 & \cdots & X_d^2 \\ \vdots & \vdots & \vdots & \vdots \\ X_1^N & X_2^N & \cdots & X_d^N \end{bmatrix}$$

samples/instances/items

Inputs/features/attributes

Reduced to a naïve Bayes' classifier

  - Features are usually assumed correlated ! $P(\boldsymbol{x}|C_i)=P(x_1,x_2,\cdots,x_d|C_i)\overset{?}{=}\prod_{j=1}^{d}P(x_j|C_i)$
    - Otherwise, there is no need for a multivariate analysis

Multivariate Data (2/2)

- Motivations for multivariate data analysis
  - Simplification
    - Assume that the large body of data can be well summarized by means of relatively few parameters
  - Exploration
    - Predict the value of one variable form the values of other variables
      - Multivariate classification (Discrete)
      - Multivariate regression (Numeric)

Parameter Estimation (1/4)

- Mean of data samples

$$E[\boldsymbol{X}] = \int \boldsymbol{x} \cdot P(\boldsymbol{x}) dx = \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix}$$

μ_1 ← Mean of Column 1 of matrix ***X***

μ_d ← Mean of Column *d* of matrix ***X***

- Covariance matrix of data samples

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{d1} & \cdots & \cdots & \sigma_d^2 \end{bmatrix}$$

σ_i^2 : variance of a variable X_i

σ_{ij} : covariace of two variables X_i and X_j

$$E[X_i X_j] = \int_{x_i} \int_{x_j} x_i x_j P(x_i x_j) dx_i dx_j$$

$$\sigma_{ij} \equiv \mathrm{Cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] = E[X_i X_j] - \mu_i \mu_j$$

if X_i and X_j are independent

$$E[X_i X_j] = E[X_i] E[X_j] = \mu_i \mu_j$$

$$\sigma_{ii} \equiv \mathrm{Cov}(X_i, X_i) = E[(X_i - \mu_i)(X_i - \mu_i)] = E[(X_i)^2] - (\mu_i)^2 = \sigma_i^2$$

Parameter Estimation (2/4)

- Covariance matrix is symmetric
  - Diagonal terms: variances
  - Off-diagonal terms: covariances

$$\Sigma = \mathrm{Cov}(\boldsymbol{X}) = E\left[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T\right] = E\left[\boldsymbol{X}\boldsymbol{X}^T\right] - \boldsymbol{\mu}\boldsymbol{\mu}^T$$

- Correlation between two variables X_i and X_j

$$\mathrm{Corr}(X_i, X_j) = \frac{\sigma_{ij}}{\sigma_i \sigma_j}, \qquad -1 \le \mathrm{Corr}(X_i, X_j) \le 1$$

independent → uncorrelated

  - Two variables independent → covariance =correlation=0
  - But covariance =correlation=0 does not imply two variables are independent (nonlinear dependence)

Parameter Estimation (3/4)

- If X_i and X_j are linear dependent

$$X_j = a \cdot X_i + b \quad (a > 0)$$
$$\Rightarrow E[X_j] = \mu_j = a \cdot [X_i] + b = a \cdot \mu_i + b$$
$$\sigma_j^2 = \mathrm{var}(X_j) = E[(X_j - \mu_j)^2] = a^2 \cdot \mathrm{var}(X_i) = a^2 \cdot \sigma_i^2$$

$$\sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)] = E[(X_i - \mu_i)((a \cdot X_i + b) - (a \cdot \mu_i + b))] = a \cdot E[(X_j - \mu_j)^2] = a \cdot \sigma_i^2$$

$$\Rightarrow \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} = \frac{a \cdot \sigma_i^2}{\sigma_i \cdot (a \cdot \sigma_i)} = 1$$

Similarly, for

$$X_j = -a \cdot X_i + b \quad (a > 0)$$
$$\rho_{ij} = -1$$

Parameter Estimation (4/4)

- Maximum Likelihood Estimators
  - Sample mean as an estimator for mean

$$\bar{\boldsymbol{\mu}} = \boldsymbol{m} = \frac{\sum_{t=1}^{N} \boldsymbol{x}^t}{N} \text{ with } m_i = \frac{\sum_{t=1}^{N} x_i^t}{N}, \text{ i} = 1,\ldots,d$$

  - Sample covariance matrix as an estimator for covariance matrix

$$s_i^2 = \frac{\sum_{t=1}^{N} \left(x_i^t - m_i\right)^2}{N}$$

$$s_{ij} = \frac{\sum_{t=1}^{N} \left(x_i^t - m_i\right)\left(x_j^t - m_j\right)}{N}$$

$$\bar{\boldsymbol{\Sigma}} = \boldsymbol{C} = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1d} \\ s_{21} & s_2^2 & \cdots & s_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ s_{d1} & \cdots & \cdots & s_d^2 \end{bmatrix}$$

  - Sample correlation coefficients

$$r_{ij} = \frac{s_{ij}}{s_i s_j}$$

Multivariate Normal Distribution (1/4)

- A random variable $\boldsymbol{x}$ that is d-dimensional and normal distributed, is denoted as $\boldsymbol{x} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$P(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

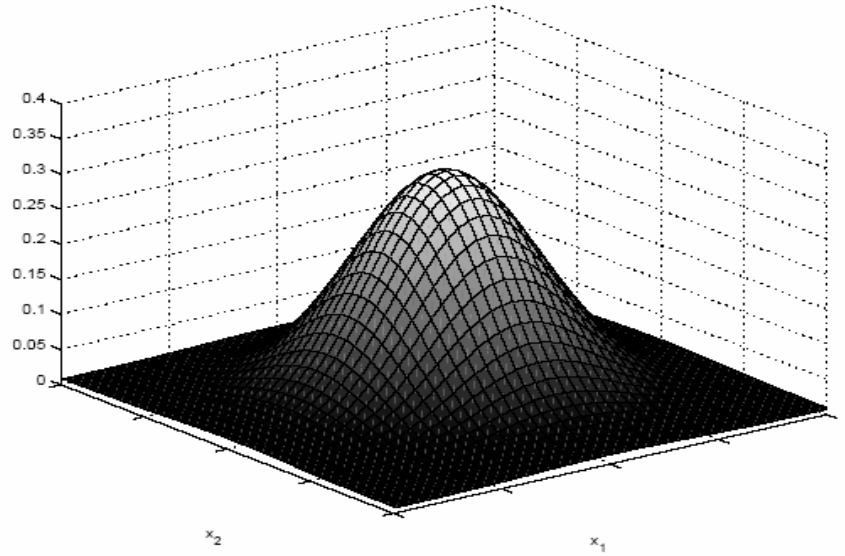
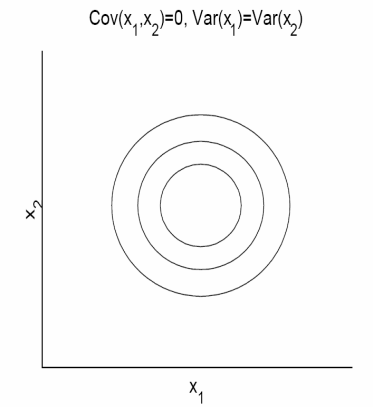


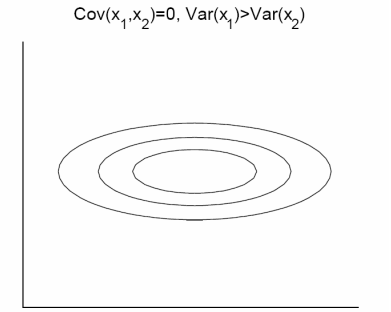
Figure 5.1: Bivariate normal distribution.

Multivariate Normal Distribution (2/4)

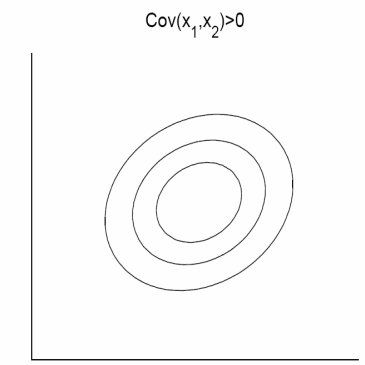
- If each dimension is independent of one another and with same variance value



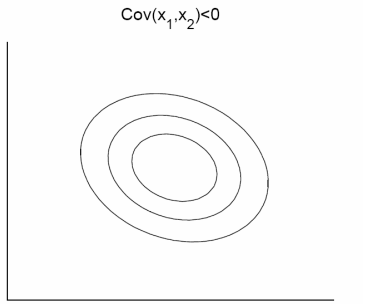
- If each dimension is independent of one another and with different variance value



- If each dimension is dependent of one another with $\mathrm{Corr}(X_1, X_2) > 0$



- If each dimension is dependent of one another with $\mathrm{Corr}(X_1, X_2) < 0$



Multivariate Normal Distribution (3/4)

- If the components of random variable $\boldsymbol{x}$ are independent

$$P(\boldsymbol{x}) = P(x_1, \cdots, x_j, \cdots, x_d) = \prod_{j=1}^{d} P(x_j)$$

$$= \frac{1}{(2\pi)^{d/2} \prod_{j=1}^{d} \sigma_j} \exp\left[-\frac{1}{2} \sum_{j=1}^{d} \left(\frac{x_j - \mu_j}{\sigma_j} \right)^2 \right]$$

Also note that $P(x_j) = \frac{1}{(2\pi)^{1/2} \sigma_j} \exp\left[-\frac{1}{2} \left(\frac{x_j - \mu_j}{\sigma_j} \right)^2 \right]$

Multivariate Normal Distribution (4/4)

- Recall: The projection of a *d*-dimensional normal distribution on a vector $\boldsymbol{w}$ is univariate normal (suppose that $\|\boldsymbol{w}\|=1$)

$$\boldsymbol{w}^T\boldsymbol{x}=w_1x_1+w_2x_2+\cdots+w_dx_d\sim N\left(\boldsymbol{w}^T\boldsymbol{\mu},\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}\right)$$

$$E\left[\boldsymbol{w}^T\boldsymbol{x}\right]=\boldsymbol{w}^TE\left[\boldsymbol{x}\right]=\underbrace{\boldsymbol{w}^T\boldsymbol{\mu}}_{\text{scalar}}$$

$$\begin{aligned}\text{Var}\left(\boldsymbol{w}^T\boldsymbol{x}\right)&=E\left[\left(\boldsymbol{w}^T\boldsymbol{x}-\boldsymbol{w}^T\boldsymbol{\mu}\right)\left(\boldsymbol{w}^T\boldsymbol{x}-\boldsymbol{w}^T\boldsymbol{\mu}\right)^T\right]\\&=E\left[\boldsymbol{w}^T\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^T\boldsymbol{w}\right]=\boldsymbol{w}^TE\left[\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^T\right]\boldsymbol{w}=\underbrace{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}}_{\text{scalar}}\end{aligned}$$

- The projection of a *d*-dimensional normal distribution to a *k*-dimensional space is *k*-variate normal

$$\boldsymbol{W}^T\boldsymbol{x}\sim N_k\left(\underbrace{\boldsymbol{W}^T\boldsymbol{\mu}}_{k\text{-dim vector}},\underbrace{\boldsymbol{W}^T\boldsymbol{\Sigma}\boldsymbol{W}}_{k\text{x}k\text{ matrix}}\right)\qquad \boldsymbol{W}=\left[\boldsymbol{w}_1\quad\boldsymbol{w}_2\quad\cdots\quad\boldsymbol{w}_k\right]$$

Multivariate Classification (1/2)

- Normal density $N_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ as the class-conditional probability $P(\boldsymbol{x}|C_i)$ of random variable $\boldsymbol{x} \in \Re^d$

$$P(\boldsymbol{x}|C_i) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right]$$

- Define the discriminant function as

$$g_i(\boldsymbol{x}) = \log P(\boldsymbol{x}|C_i) + \log P(C_i)$$

$$= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) + \log P(C_i)$$

Multivariate Classification (2/2)

- Maximum likelihood (ML) training of classifiers
  - Given a set of labeled samples

$$\boldsymbol{X} = \left\{\boldsymbol{x}^t, \boldsymbol{r}^t\right\}, \; r_i^t = 1 \text{ if } \boldsymbol{x}^t \in C_i \text{ and } \; 0 \; \text{otherwise}$$

$$\hat{P}(C_i) = \frac{\sum_{t=1}^{N} r_i^t}{N}$$

$$\hat{\boldsymbol{\mu}}_i = \boldsymbol{m}_i \; (\text{sample mean}) = \frac{\sum_{t=1}^{N} r_i^t \boldsymbol{x}^t}{\sum_{t=1}^{N} r_i^t}$$

$$\hat{\boldsymbol{\Sigma}}_i = \boldsymbol{S}_i \; (\text{sample covariance matrix}) = \frac{\sum_{t=1}^{N} r_i^t \left(\boldsymbol{x}^t - \boldsymbol{m}_i\right)\left(\boldsymbol{x}^t - \boldsymbol{m}_i\right)^T}{\sum_{t=1}^{N} r_i^t}$$

Quadratic Discriminant (1/2)

- The discriminant function with normal class-conditional density can be expressed as a quadratic discriminant

$$g_i(\boldsymbol{x}) = \log P(\boldsymbol{x}|C_i) + \log P(C_i)$$

$$= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{S}_i| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_i)^T \boldsymbol{S}_i^{-1}(\boldsymbol{x}-\boldsymbol{m}_i) + \log P(C_i)$$

$$= -\frac{1}{2}\left(\boldsymbol{x}^T\boldsymbol{S}_i^{-1}\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{S}_i^{-1}\boldsymbol{m}_i + \boldsymbol{m}_i^T\boldsymbol{S}_i^{-1}\boldsymbol{m}_i\right) - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{S}_i| + \log P(C_i)$$

$$= \boldsymbol{x}^T\boldsymbol{W}_i\boldsymbol{x} + \boldsymbol{w}_i^T\boldsymbol{x} + w_{i0}$$

quadratic discriminant

Recall : second - order polynomial for scalar variable

$g_i(x) = w_{i2}x^2 + w_{i1}x + w_{i0}$

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

$$\boldsymbol{w}_i = \boldsymbol{S}_i^{-1}\boldsymbol{m}_i$$

$$w_{i0} = -\frac{d}{2}\log 2\pi - \frac{1}{2}\boldsymbol{m}_i^T\boldsymbol{S}_i^{-1}\boldsymbol{m}_i - \frac{1}{2}\log|\boldsymbol{S}_i| + \log P(C_i)$$

Quadratic Discriminant (2/2)

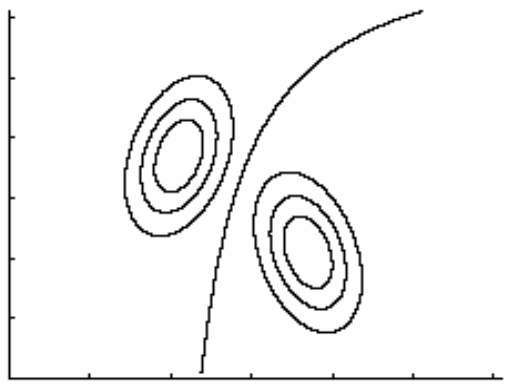
likelihood

posterior

$$g_1(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{W}_1 \boldsymbol{x} + \boldsymbol{w}_1^T \boldsymbol{x} + w_{10}$$

$$g_2(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{W}_2 \boldsymbol{x} + \boldsymbol{w}_2^T \boldsymbol{x} + w_{20}$$

$$\begin{aligned}\tilde{g}(\boldsymbol{x}) &= g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}) \\ &= \boldsymbol{x}^T (\boldsymbol{W}_1 - \boldsymbol{W}_2)\boldsymbol{x} + (\boldsymbol{w}_1^T - \boldsymbol{w}_2^T)\boldsymbol{x} \\ &\quad + (w_{10} - w_{20}) \\ &= \boldsymbol{x}^T W' \boldsymbol{x} + \boldsymbol{w}' \boldsymbol{x} + w_0'\end{aligned}$$



likelihood

Figure 5.3: Classes have different covariance matrices.

Linear Discriminant (1/2)

- The discriminant function with normal class-conditional density sharing the same covariance matrix can be expressed as a linear discriminant

$$\boldsymbol{S} = \sum_{i=1}^{K} \hat{P}(C_i) \cdot \boldsymbol{S}_i$$

All classes share the same covariance matrix

$$g_i(\boldsymbol{x}) = \log P(\boldsymbol{x}|C_i) + \log P(C_i)$$

$$= -\frac{1}{2}\left(\boldsymbol{x}^T \boldsymbol{S}^{-1} \boldsymbol{x} - 2\boldsymbol{x}^T \boldsymbol{S}^{-1} \boldsymbol{m}_i + \boldsymbol{m}_i^T \boldsymbol{S}^{-1} \boldsymbol{m}_i\right) - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{S}| + \log P(C_i)$$

$$\Rightarrow \tilde{g}_i(\boldsymbol{x}) = \boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}$$

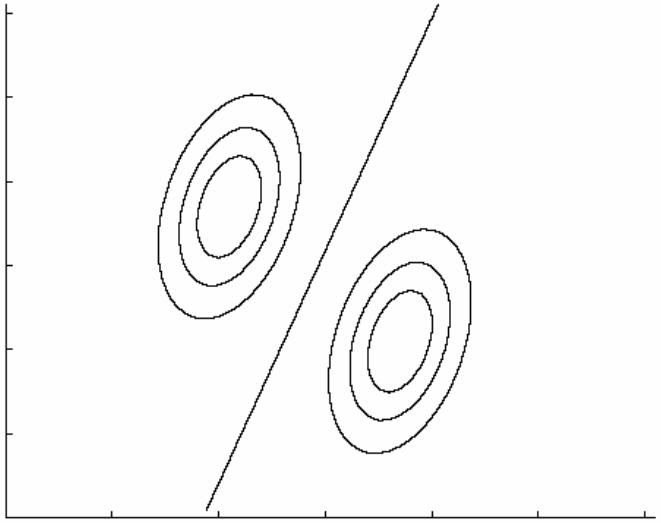
$$w_i = \boldsymbol{S}^{-1} \boldsymbol{m}_i$$

$$w_{i0} = -\frac{d}{2}\log 2\pi - \frac{1}{2}\boldsymbol{m}_i^T \boldsymbol{S}^{-1} \boldsymbol{m}_i - \frac{1}{2}\log|\boldsymbol{S}| + \log P(C_i)$$

Linear Discriminant (2/2)

$$g_1(\boldsymbol{x}) = \boldsymbol{w}_1^T \boldsymbol{x} + w_{10}$$
$$g_2(\boldsymbol{x}) = \boldsymbol{w}_2^T \boldsymbol{x} + w_{20}$$
$$\tilde{g}(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$$
$$= \left(\boldsymbol{w}_1^T - \boldsymbol{w}_2^T\right)\boldsymbol{x} + \left(w_{10} - w_{20}\right)$$
$$= \boldsymbol{w}'^T \boldsymbol{x} + w_0'$$

covariances (off-diagonal terms of covariance matrix) are not equal to zero



kelihood

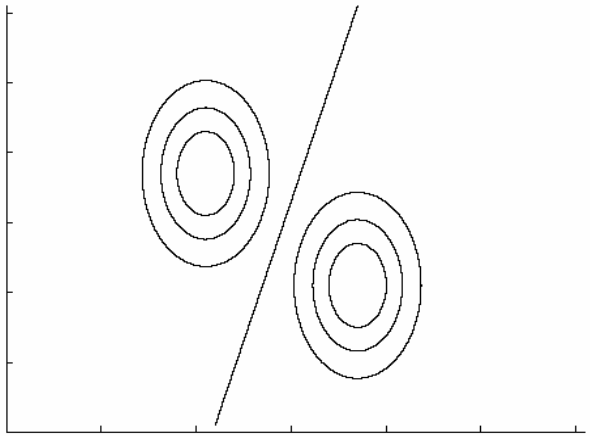
Figure 5.4: Covariances may be arbitary but shared by both classes.

Naïve Bayes' Classifier (1/2)

- The discriminant function with a normal class-conditional density sharing the same diagonal covariance matrix can be expressed as a naïve Bayes' classifier

$$g_i(\boldsymbol{x}) = -\frac{1}{2}\sum_{j=1}^{d}\left(\frac{x_j - m_{ij}}{s_j}\right)^2 + \log P(C_i)$$

variances (diagonal terms of covariance matrix) are not equal



likelihood

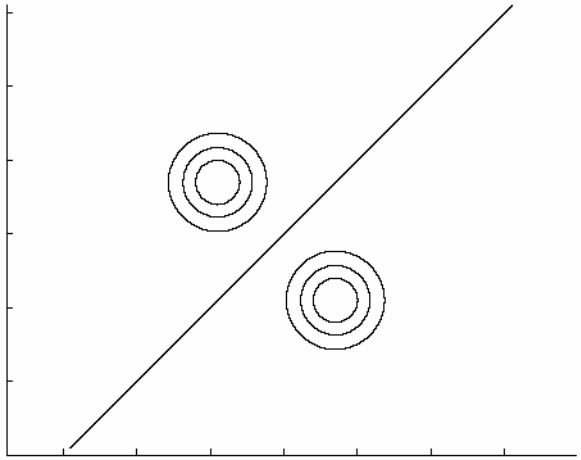
Figure 5.5: All classes have equal, diagonal covariance matrices but variances are not equal.

Naïve Bayes' Classifier (2/2)

- If the variances (diagonal terms) of the naïve Bayes' classifier are further set to equal for all dimensions

$$g_i(\boldsymbol{x}) = -\frac{1}{2s^2}\sum_{j=1}^{d}\left(x_j - m_{ij}\right)^2 + \log P(C_i)$$

The Mahalanobis distance is reduced to Euclidean distance. (all variables/features have the same variance and are independent)



likelihood

Figure 5.6: All classes have equal, diagonal covariance matrices of equal variances on both dimensions.

Nearest Mean Classifier

- Assign the data sample to the class of the nearest mean
  - If the priors $P(C_i)$ are further set to equal

$$g_i(\boldsymbol{x}) = -\sum_{j=1}^{d}\left(x_j - m_{ij}\right)^2 = -\left\|\boldsymbol{x} - \boldsymbol{m}_i\right\|^2$$

$$= -\left(\boldsymbol{x} - \boldsymbol{m}_i\right)^T\left(\boldsymbol{x} - \boldsymbol{m}_i\right)$$

$$= -\left(\boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{m}_i^T\boldsymbol{x} + \boldsymbol{m}_i^T\boldsymbol{m}_i\right)$$

$$\Rightarrow \bar{g}_i(\boldsymbol{x}) = \boldsymbol{m}_i^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{m}_i^T\boldsymbol{m}_i$$

$$= \boldsymbol{w}_i^T\boldsymbol{x} + w_{i0}$$

where $\boldsymbol{w}_i = \boldsymbol{m}_i$

$$w_{i0} = -\frac{1}{2}\boldsymbol{m}_i^T\boldsymbol{m}_i = -\frac{1}{2}\left\|\boldsymbol{m}_i\right\|^2$$

Tuning Complexity

- Tradeoff between the bias and variance of an estimator

Table 5.1 Reducing variance through simplifying assumptions.

Assumption	Covariance matrix	No. of parameters
Shared, Hyperspheric	$\mathbf{S}_i = \mathbf{S} = s^2\mathbf{I}$	1
Shared, Axis-aligned	$\mathbf{S}_i = \mathbf{S}$, with $s_{ij} = 0$	d
Shared, Hyperellipsoidal	$\mathbf{S}_i = \mathbf{S}$	$d(d+1)/2$
Different, Hyperellipsoidal	$\mathbf{S}_i$	$K \cdot (d(d+1)/2)$

  - Simplifying covariance matrix → decreasing number of parameters → introducing estimation bias
  - Arbitrary covariance matrix → much more data is needed → introducing estimation variance

- Regularized Discriminant Analysis (RDA, 1989)
  - Use a weighted average of three special cases of covariance matrix

$$\boldsymbol{S}' = \alpha \cdot \sigma^2 \boldsymbol{I} + \beta \cdot \boldsymbol{S} + (1 - \alpha - \beta) \cdot \boldsymbol{S}_i$$

a shared diagonal covariance matrix (with equal variance) for all classes

a shared covariance matrix for all classes

a specific covariance matrix for each class

Discrete Features (1/4)

- Features that take one of *n* different values
  - E.g. $\text{pixel} \in \{\text{on}, \text{off}\}; \text{color} \in \{\text{red}, \text{blue}, \text{green}, \text{black}\}$
- If each feature is a binary random variable $x_j^t \ \left(x_j^t = 1 \text{ or } x_j^t = 0\right)$
  - Bernoulli distribution

$$P_{ij} = P\left(x_j^t = 1 \middle| C_i\right) \quad 1 - P_{ij} = P\left(x_j^t = 0 \middle| C_i\right) \quad \Rightarrow P\left(x_j^t \middle| C_i\right) = \left(P_{ij}\right)^{x_j^t} \left(1 - P_{ij}\right)^{\left(1 - x_j^t\right)}$$

  - If features are further assumed to be independent
    - Assumption for Naïve Bayes' classifier

$$\boldsymbol{x}^t = \left\{x_1^t, \cdots, x_j^t, \cdots, x_d^t\right\}$$

$$P\left(\boldsymbol{x}^t \middle| C_i\right) = P\left(x_1^t, \cdots, x_d^t \middle| C_i\right) = \prod_{j=1}^{d} P\left(x_j^t \middle| C_i\right) = \prod_{j=1}^{d} \left(P_{ij}\right)^{x_j^t} \left(1 - P_{ij}\right)^{\left(1 - x_j^t\right)}$$

    - Discriimant function

$$g_i\left(\boldsymbol{x}^t\right) = \log P\left(\boldsymbol{x}^t \middle| C_i\right) + \log P\left(C_i\right) = \sum_{j=1}^{d} \left[x_j^t \log P_{ij} + \left(1 - x_j^t\right) \log\left(1 - P_{ij}\right)\right] + \log P\left(C_i\right)$$

      - Linear ?

Discrete Features (2/4)

- Appendix A

$$g_i(\boldsymbol{x}^t) = \sum_{j=1}^{d}\left[x_j^t \log P_{ij} + (1-x_j^t)\log(1-P_{ij})\right] + \log P(C_i)$$

$$= \left[\sum_{j=1}^{d} x_j^t \log P_{ij}\right] - \left[\sum_{j=1}^{d} x_j^t (1-\log P_{ij})\right] + \left[\sum_{j=1}^{d} (1-\log P_{ij})\right] + \log P(C_i)$$

$$= \left[\sum_{j=1}^{d} x_j^t (2\cdot(\log P_{ij})-1)\right] + \left[\sum_{j=1}^{d} (1-\log P_{ij})\right] + \log P(C_i)$$

$$= \boldsymbol{w}_i^T \boldsymbol{x}^t + D_i \quad \textbf{Linear}$$

where

$$\boldsymbol{w}_i = \begin{bmatrix} 2\cdot(\log P_{i1})-1 \\ 2\cdot(\log P_{i2})-1 \\ \vdots \\ 2\cdot(\log P_{id})-1 \end{bmatrix} \qquad D_i = \left[\sum_{j=1}^{d} (1-\log P_{ij})\right] + \log P(C_i)$$

Discrete Features (3/4)

- Maximum likelihood estimation for P_{ij} (for binary variable)

$$\hat{P}_{ij} = \frac{\sum_{t=1}^{N} x_j^t \cdot r_i^t}{\sum_{t=1}^{N} r_i^t}$$

- Extension: features are independent multinomial random variables

$$x_j^t \in \{v_{j1}, \cdots, v_{jk}, \cdots, v_{jn_j}\} \qquad z_{jk}^t = \begin{cases} 1 & \text{if } x_j^t = v_{jk} \\ 0 & \text{otherwise} \end{cases}$$

  - Define the probability that x_j^t belongs to C_i and take v_{jk} :

$$P_{ijk} = P\left(z_{jk}^t = 1 \middle| C_i\right) = P\left(x_j^t = v_{jk} \middle| C_i\right) \quad \Rightarrow \quad \prod_{j=1}^{d} P\left(x_j^t \middle| C_i\right) = \prod_{k=1}^{n_j} P_{ijk}^{z_{jk}^t}$$

Discrete Features (4/4)

– Assumption for Naïve Bayes' classifier

$$P\left(\boldsymbol{x}^t \middle| C_i\right) = P\left(x_1^t, \cdots, x_d^t \middle| C_i\right) = \prod_{j=1}^{d} P\left(x_j^t \middle| C_i\right) = \prod_{j=1}^{d} \prod_{k=1}^{n_j} P_{ijk}^{z_{jk}^t}$$

– Discriimant function

$$g_i\left(\boldsymbol{x}^t\right) = \log P\left(\boldsymbol{x}^t \middle| C_i\right) + \log P\left(C_i\right) = \sum_{j=1}^{d} \sum_{k=1}^{n_j} z_{jk}^t \log P_{ijk} + \log P\left(C_i\right)$$

- Linear ?

– Maximum likelihood estimation for P_{ijk}

$$\hat{P}_{ijk} = \frac{\sum_{t=1}^{N} z_{jk}^t \cdot r_i^t}{\sum_{t=1}^{N} r_i^t}$$