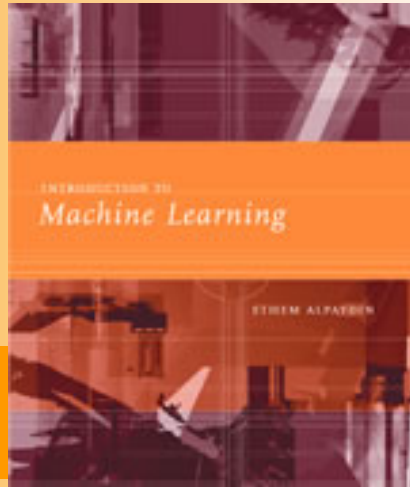




Lecture Slides for

INTRODUCTION TO

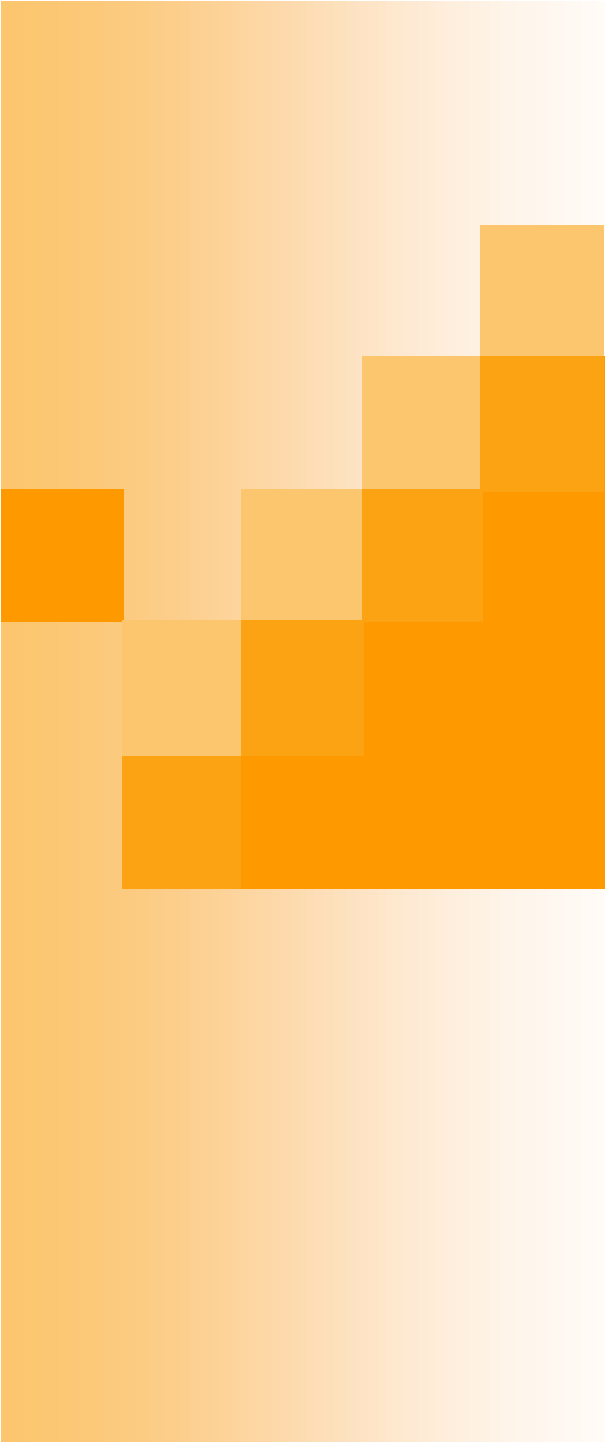
Machine Learning

ETHEM ALPAYDIN

© The MIT Press, 2004

alpaydin@boun.edu.tr

<http://www.cmpe.boun.edu.tr/~ethem/i2ml>



CHAPTER 5:

Multivariate Methods



Multivariate Data

- Multiple measurements (sensors)
- d inputs/features/attributes: d -variate
- N instances/observations/examples

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

Multivariate Parameters

- Mean: $E[\mathbf{x}] = \boldsymbol{\mu} = [\mu_1, \dots, \mu_d]^T$
- Covariance: $\sigma_{ij} \equiv \text{Cov}(X_i, X_j)$
- Correlation: $\text{Corr}(X_i, X_j) \equiv \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$

$$\boldsymbol{\Sigma} \equiv \text{Cov}(\mathbf{X}) = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T]$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & & & \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$



Parameter Estimation

- Sample mean ***m***

$$m_i = \frac{\sum_{t=1}^N x_i^t}{N}, i = 1, \dots, d$$

- Sample covariance matrix **S**

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

- Sample correlation matrix **R**

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

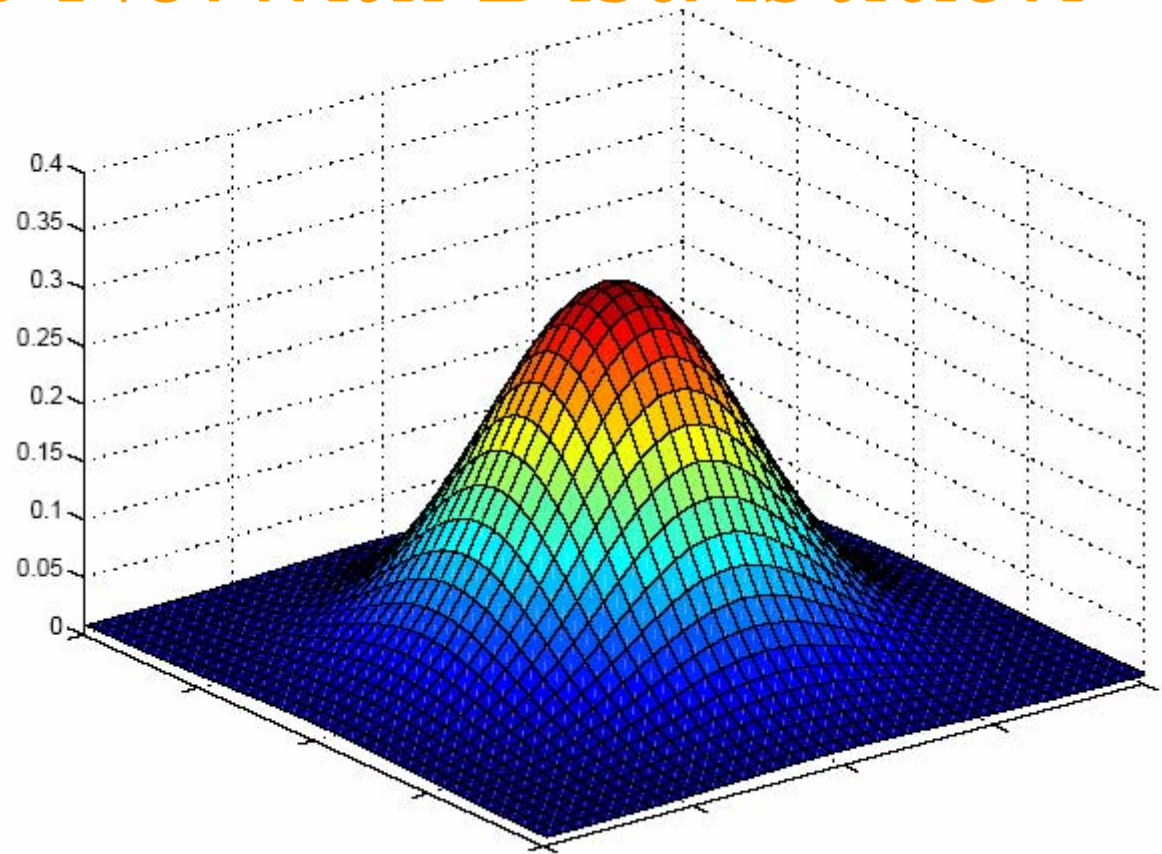


Estimation of Missing Values

- What to do if certain instances have missing attributes?
- Ignore those instances: not a good idea if the sample is small
- Use 'missing' as an attribute: may give information
- **Imputation:** Fill in the missing value
 - Mean imputation: Use the most likely value (e.g., mean)
 - Imputation by regression: Predict based on other attributes

Multivariate Normal Distribution

$$\mathbf{x} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$



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



Multivariate Normal Distribution

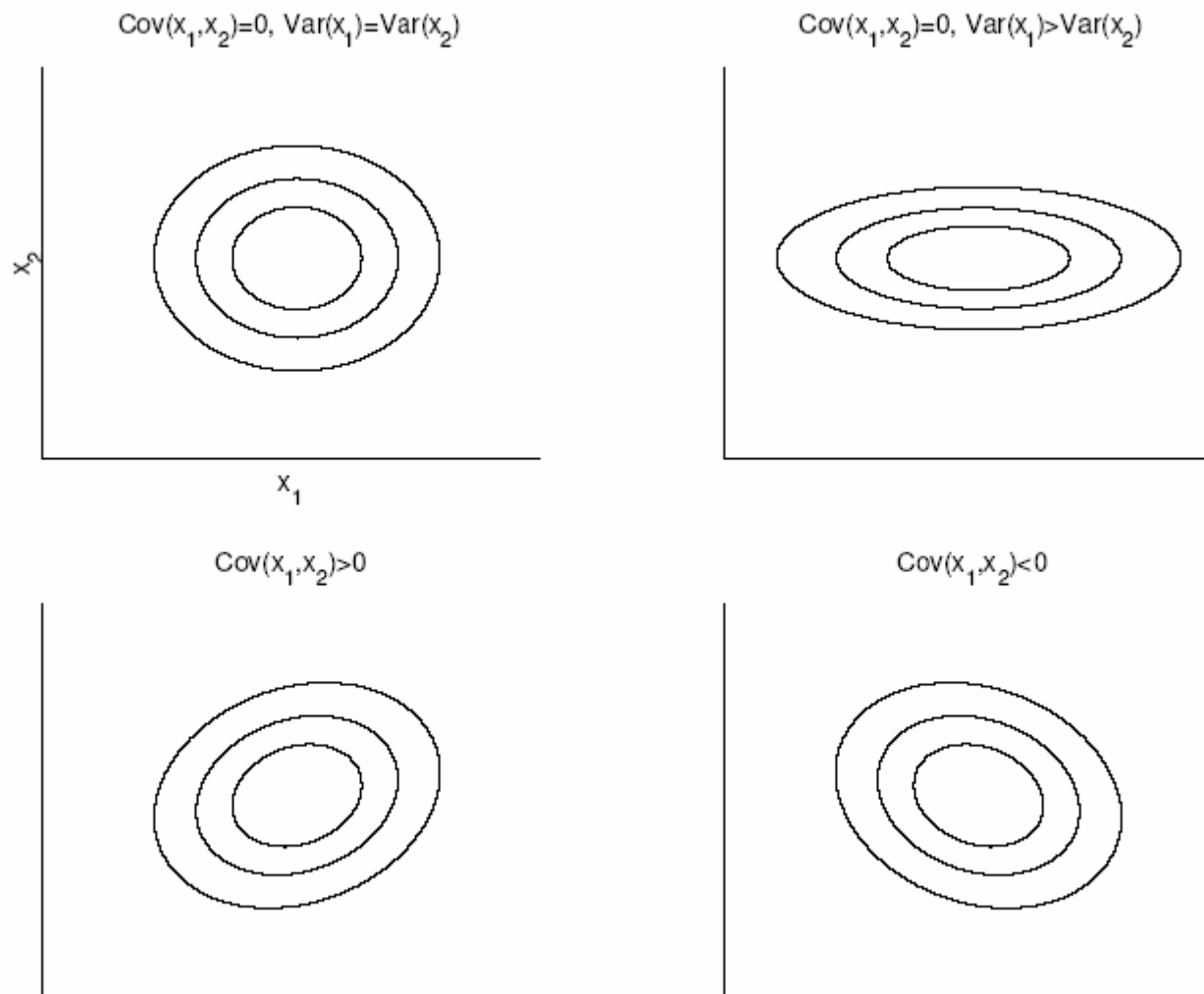
- Mahalanobis distance: $(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$
measures the distance from $\mathbf{x}$ to $\boldsymbol{\mu}$ in terms of $\boldsymbol{\Sigma}$
(normalizes for difference in variances and correlations)

- Bivariate: $d = 2$ $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$

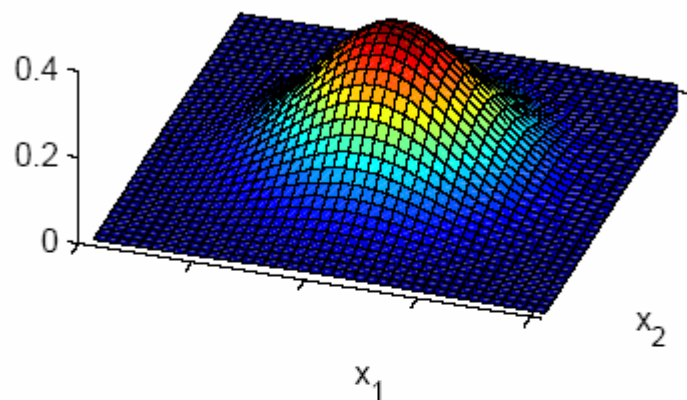
$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left[-\frac{1}{2(1-\rho^2)} (z_1^2 - 2\rho z_1 z_2 + z_2^2) \right]$$

$$z_i = (x_i - \mu_i) / \sigma_i$$

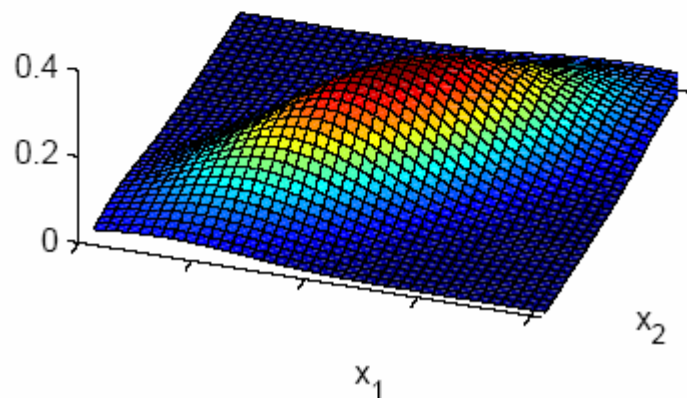
Bivariate Normal



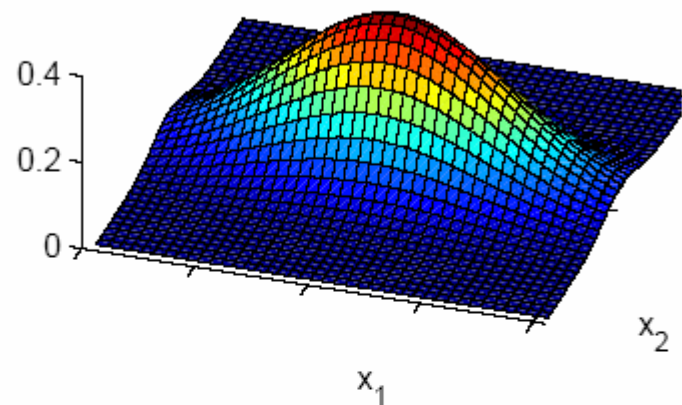
$\text{Cov}(x_1, x_2) = 0, \text{Var}(x_1) = \text{Var}(x_2)$



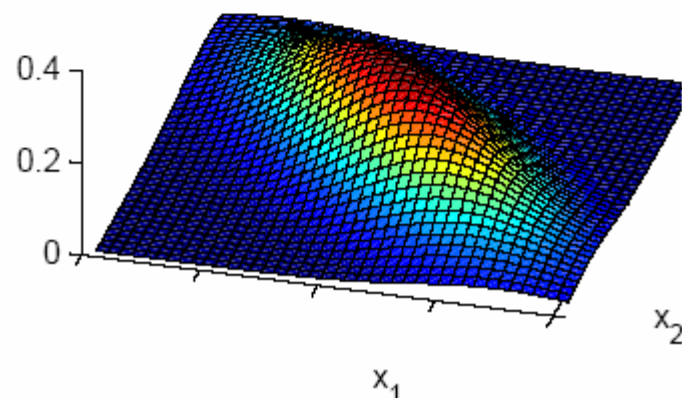
$\text{Cov}(x_1, x_2) > 0$



$\text{Cov}(x_1, x_2) = 0, \text{Var}(x_1) > \text{Var}(x_2)$



$\text{Cov}(x_1, x_2) < 0$





Independent Inputs: Naive Bayes

- If x_i are independent, offdiagonals of Σ are 0, Mahalanobis distance reduces to weighted (by $1/\sigma_i$) Euclidean distance:

$$p(\mathbf{x}) = \prod_{i=1}^d p_i(x_i) = \frac{1}{(2\pi)^{d/2} \prod_{i=1}^d \sigma_i} \exp \left[-\frac{1}{2} \sum_{i=1}^d \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \right]$$

- If variances are also equal, reduces to Euclidean distance



Parametric Classification

- If $p(\mathbf{x} | C_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

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

- Discriminant functions are

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

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



Estimation of Parameters

$$\begin{aligned}\hat{P}(C_i) &= \frac{\sum_t r_i^t}{N} \\ \mathbf{m}_i &= \frac{\sum_t r_i^t \mathbf{x}^t}{\sum_t r_i^t} \\ \mathbf{S}_i &= \frac{\sum_t r_i^t (\mathbf{x}^t - \mathbf{m}_i)(\mathbf{x}^t - \mathbf{m}_i)^T}{\sum_t r_i^t}\end{aligned}$$

$$g_i(\mathbf{x}) = -\frac{1}{2} \log |\mathbf{S}_i| - \frac{1}{2} (\mathbf{x} - \mathbf{m}_i)^T \mathbf{S}_i^{-1} (\mathbf{x} - \mathbf{m}_i) + \log \hat{P}(C_i)$$

Different S_i

■ Quadratic discriminant

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

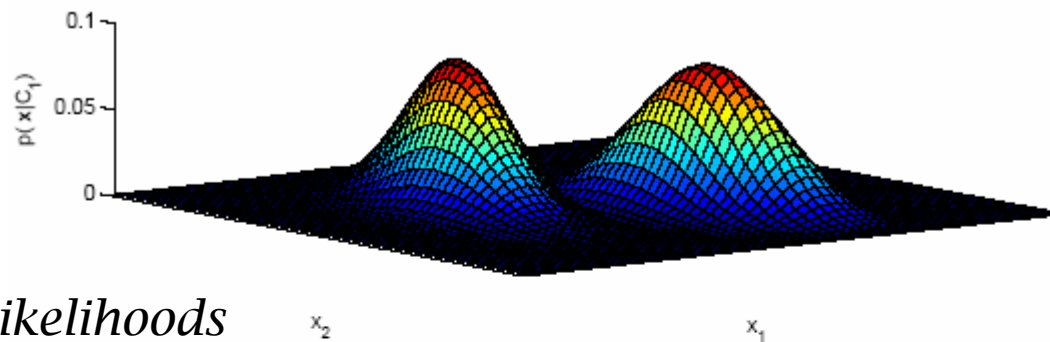
$$g_i(\mathbf{x}) = \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

where

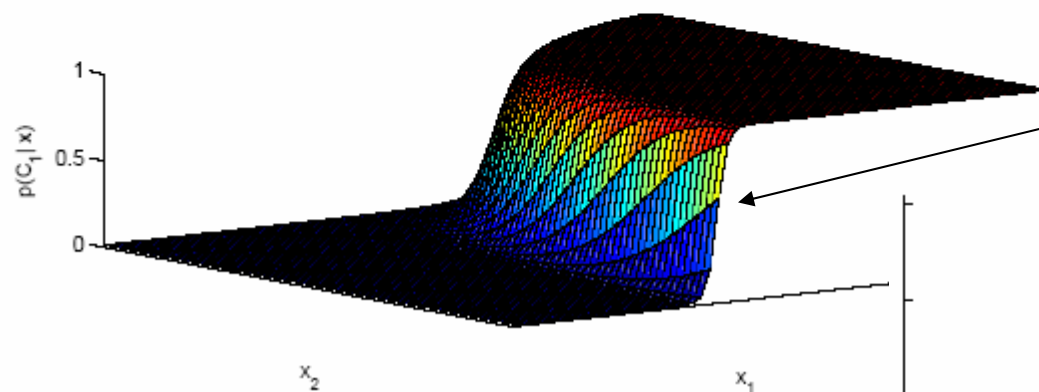
$$\mathbf{W}_i = -\frac{1}{2} \mathbf{S}_i^{-1}$$

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

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

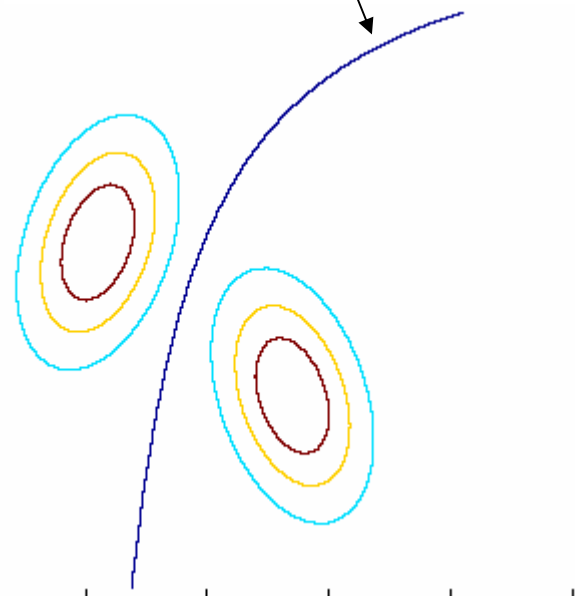


likelihoods



posterior for C_1

discriminant:
 $P(C_1|\mathbf{x}) = 0.5$





Common Covariance Matrix S

- Shared common sample covariance S

$$S = \sum_i \hat{P}(C_i) S_i$$

- Discriminant reduces to

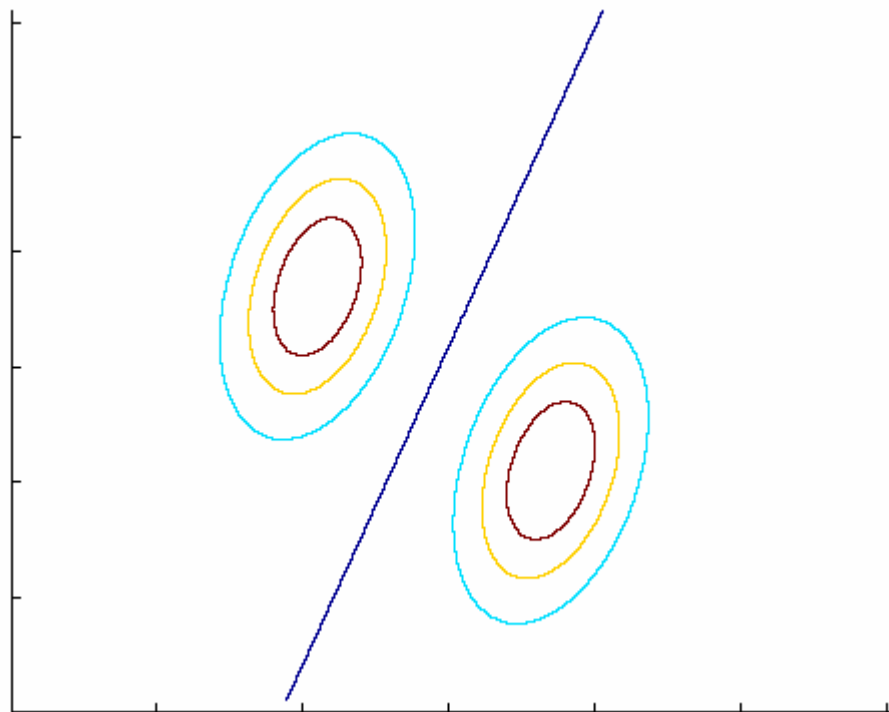
$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T S^{-1}(\mathbf{x} - \mathbf{m}_i) + \log \hat{P}(C_i)$$

which is a **linear discriminant** $g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$

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

$$w_{i0} = -\frac{1}{2} \mathbf{m}_i^T S^{-1} \mathbf{m}_i + \log \hat{P}(C_i)$$

Common Covariance Matrix S





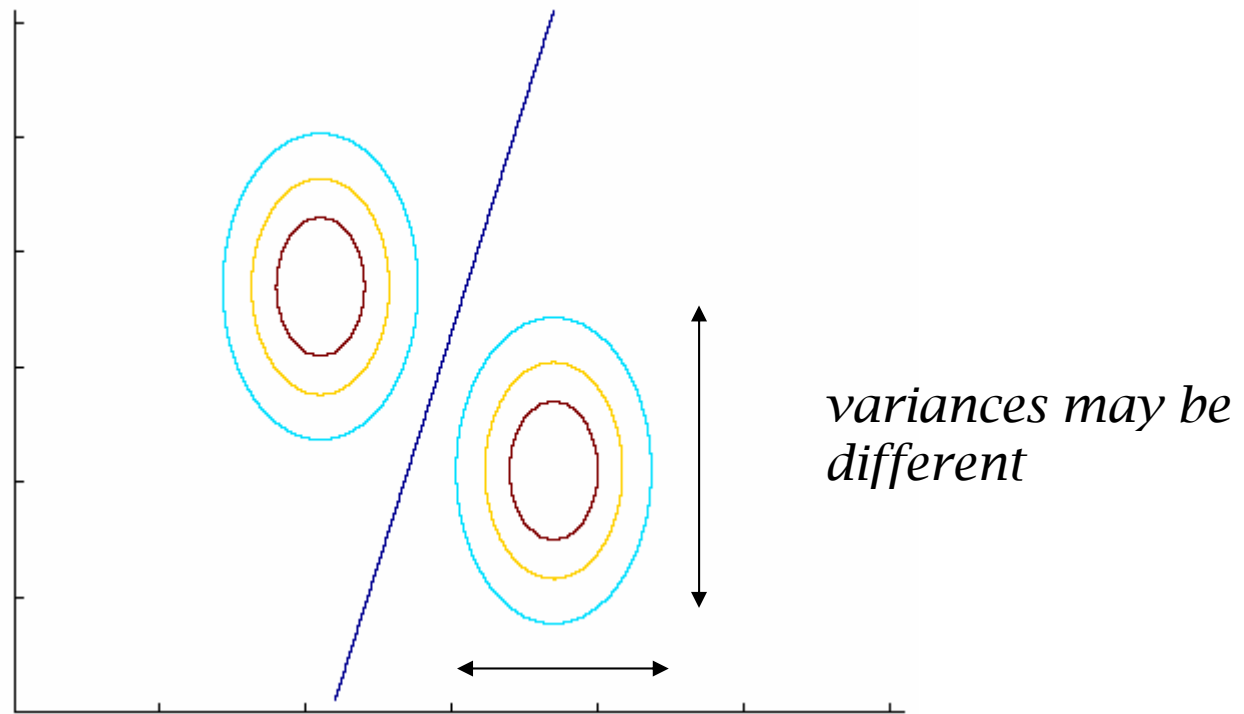
Diagonal Σ

- When $x_j, j = 1, \dots, d$, are independent, Σ is diagonal
 $p(\mathbf{x}|C_i) = \prod_j p(x_j|C_i)$ (Naive Bayes' assumption)

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

Classify based on weighted Euclidean distance (in s_j units) to the nearest mean

Diagonal S





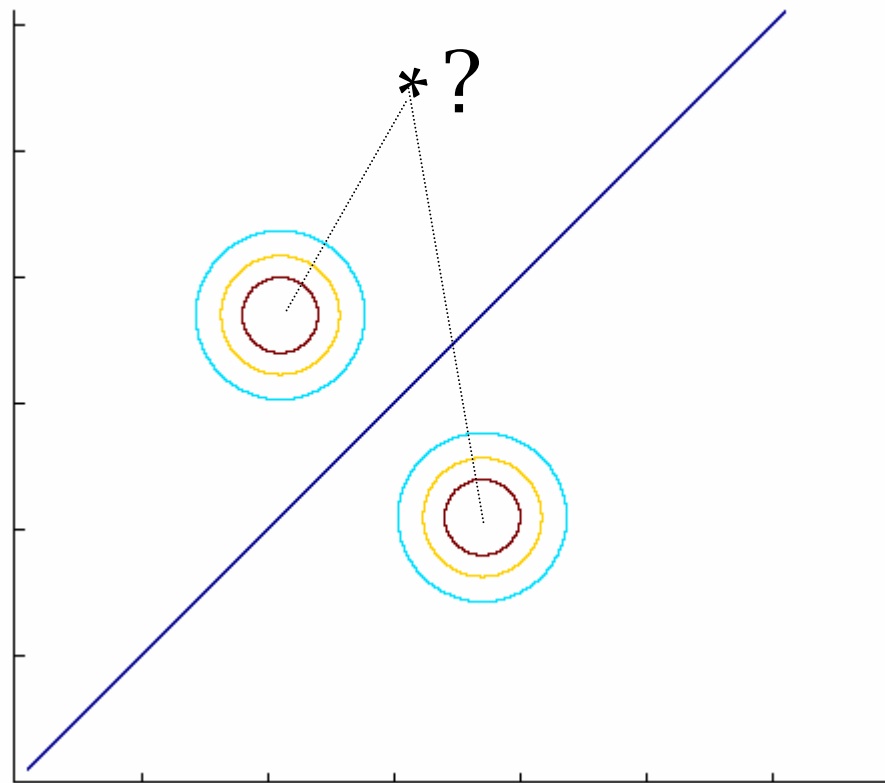
Diagonal S , equal variances

- **Nearest mean classifier:** Classify based on Euclidean distance to the nearest mean

$$\begin{aligned} g_i(\mathbf{x}) &= -\frac{\|\mathbf{x} - \mathbf{m}_i\|^2}{2s^2} + \log \hat{P}(C_i) \\ &= -\frac{1}{2s^2} \sum_{j=1}^d (x_j^t - m_{ij})^2 + \log \hat{P}(C_i) \end{aligned}$$

- Each mean can be considered a **prototype** or **template** and this is **template matching**

Diagonal S , equal variances





Model Selection

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	$\mathbf{S}_i$	$K \cdot (d(d+1)/2)$

- As we increase complexity (less restricted $\mathbf{S}$), bias decreases and variance increases
- Assume simple models (allow some bias) to control variance (regularization)

Discrete Features

- **Binary** features: $p_{ij} \equiv p(x_j = 1 | C_i)$
if x_j are **independent** (Naive Bayes')

$$p(\mathbf{x} | C_i) = \prod_{j=1}^d p_{ij}^{x_j} (1 - p_{ij})^{(1-x_j)}$$

the discriminant is **linear**

$$\begin{aligned} g_i(\mathbf{x}) &= \log p(\mathbf{x} | C_i) + \log P(C_i) \\ &= \sum_j \left[x_j \log p_{ij} + (1 - x_j) \log(1 - p_{ij}) \right] + \log P(C_i) \end{aligned}$$

Estimated parameters $\hat{p}_{ij} = \frac{\sum_t x_j^t r_i^t}{\sum_t r_i^t}$



Discrete Features

- **Multinomial** (1-of- n_j) features: $x_j \in \{v_1, v_2, \dots, v_{n_j}\}$

$$p_{ijk} \equiv p(z_{jk} = 1 | C_i) = p(x_j = v_k | C_i)$$

if x_j are **independent**

$$p(\mathbf{x} | C_i) = \prod_{j=1}^d \prod_{k=1}^{n_j} p_{ijk}^{z_{jk}}$$

$$g_i(\mathbf{x}) = \sum_j \sum_k z_{jk} \log p_{ijk} + \log P(C_i)$$

$$\hat{p}_{ijk} = \frac{\sum_t z_{jk}^t r_i^t}{\sum_t r_i^t}$$



Multivariate Regression

$$r^t = g(\mathbf{x}^t | w_0, w_1, \dots, w_d) + \epsilon$$

- Multivariate linear model

$$w_0 + w_1 x_1^t + w_2 x_2^t + \dots + w_d x_d^t$$

$$E(w_0, w_1, \dots, w_d | \mathcal{X}) = \frac{1}{2} \sum_t (r^t - w_0 - w_1 x_1^t - w_2 x_2^t - \dots - w_d x_d^t)^2$$

- Multivariate polynomial model:

Define new higher-order variables

$$z_1 = x_1, z_2 = x_2, z_3 = x_1^2, z_4 = x_2^2, z_5 = x_1 x_2$$

and use the linear model in this new $\mathbf{z}$ space

(basis functions, kernel trick, SVM: Chapter 10)