CHAPTER 5:

MULTİVARİATE METHODS (SECTIONS 5.1-5.5)

Multivariate Data

- Multiple measurements with varied type/scale
- ***d*** inputs/features/attributes are correlated: *d*-variate
  - Simplification: feature selection
  - Exploration: model data, predict one var given the others
- *N* instances/observations/examples

$$\begin{matrix}\text{Data} \\ \text{matrix}\end{matrix} = \begin{bmatrix} X_1^1 & X_2^1 & \cdots & X_d^1 \\ X_1^2 & X_2^2 & \cdots & X_d^2 \\ \vdots & & & \\ X_1^N & X_2^N & \cdots & X_d^N \end{bmatrix}$$

Multivariate Parameters

$$\text{Mean}: E[\mathbf{x}] = \boldsymbol{\mu} = [\mu_1, \ldots, \mu_d]^T$$

check sign

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

$$\text{Correlation}: \text{Corr}(X_i, X_j) \equiv \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \qquad (\sigma_i = \sqrt{\sigma_{ii}})$$

$$\Sigma \equiv \text{Cov}(\mathbf{X}) = E\left[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T\right] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & & & \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

When independent?

Parameter Estimation (MLE)

$$\text{Sample mean } \mathbf{m} : m_i = \frac{\sum_{t=1}^{N} x_i^t}{N}, i = 1, \ldots, d$$

$$\text{Covariance matrix } \mathbf{S} : s_{ij} = \frac{\sum_{t=1}^{N} \left(x_i^t - m_i\right)\left(x_j^t - m_j\right)}{N}$$

$$\text{Correlation matrix } \mathbf{R} : r_{ij} = \frac{s_{ij}}{s_i s_j}$$

$$\text{Mean} : E[\mathbf{x}] = \boldsymbol{\mu} = [\mu_1, \ldots, \mu_d]^T$$

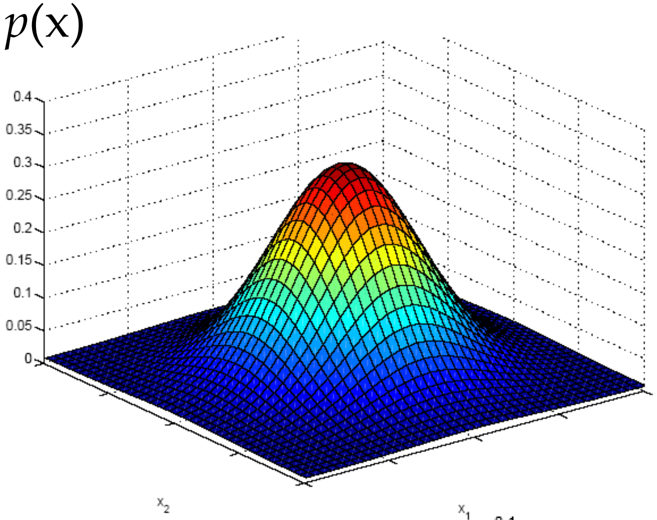
$$\text{Covariance} : \sigma_{ij} \equiv \text{Cov}\left(X_i, X_j\right) = E[(X_i - \mu_i)(X_j - \mu_j)]$$

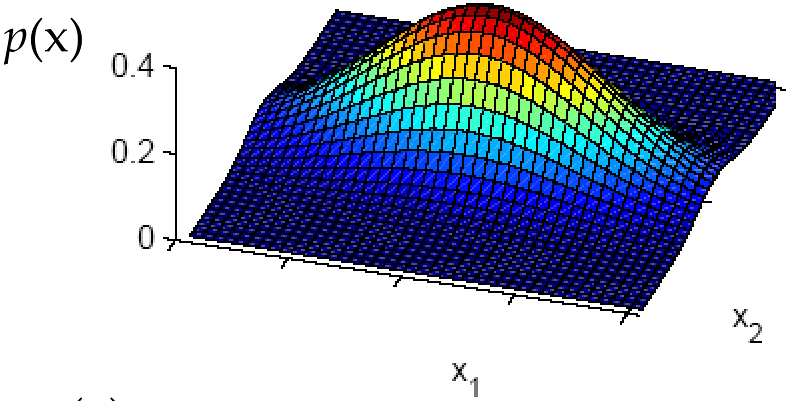
$$\text{Correlation} : \text{Corr}\left(X_i, X_j\right) \equiv \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

Estimation of Missing Values

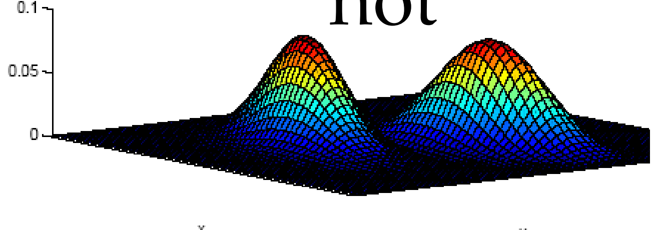
- What to do if certain instances have missing attributes?
- Ignore those instances: not a good idea if the sample is small
- Use ‘missing’ as an attribute: may give information
  - E.g. salary when applying for credit card
- Imputation: Fill in the missing value
  - Mean imputation: Use the most likely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

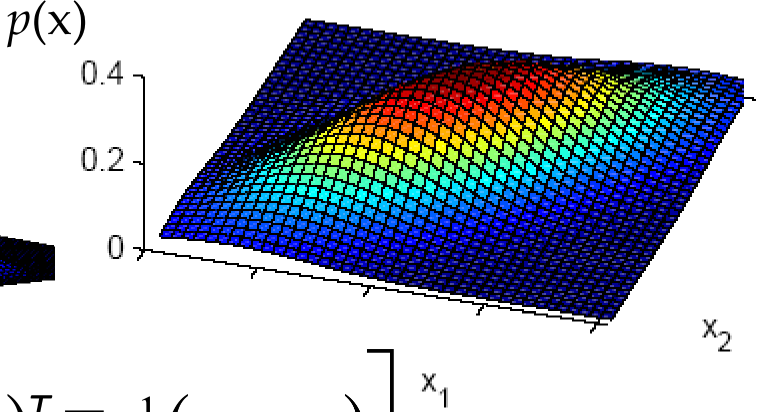
Multivariate Normal Distribution





not





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

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

1-d as special case, single mode, credit card application

Multivariate Normal Distribution

- □ Mahalanobis distance: $(\mathbf{x}-\boldsymbol{\mu})^T \sum^{-1} (\mathbf{x}-\boldsymbol{\mu})$

  measures the distance from $\mathbf{x}$ to $\boldsymbol{\mu}$ in terms of $\sum$ (normalizes for difference in variances and correlations)

$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

- □ Bivariate: $d = 2$

(nice property of Gaussian, hence called covariance matrix)

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

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

$$\boxed{z_i = (x_i - \mu_i)/\sigma_i}$$

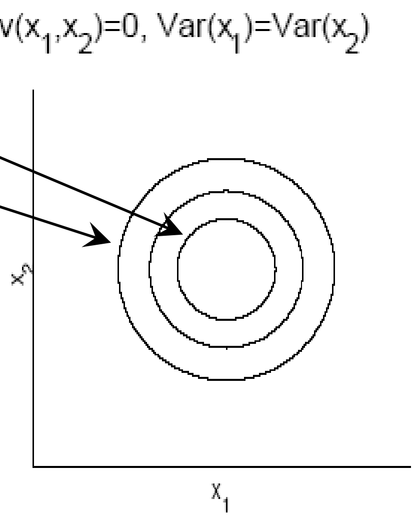
standardized variables

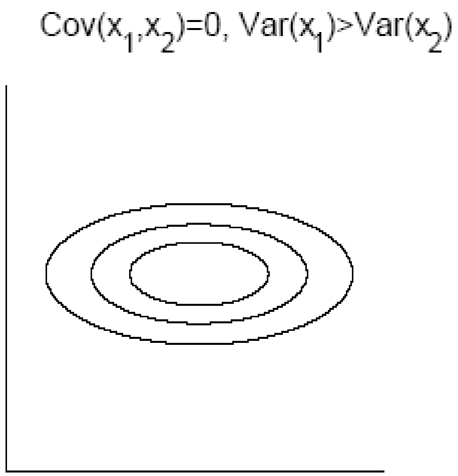
Bivariate Normal Distribution: isoprobable contours

e.g., $\{x : p(\mathrm{x}) = 0.7\}$

$\{x : p(\mathrm{x}) = 0.3\}$

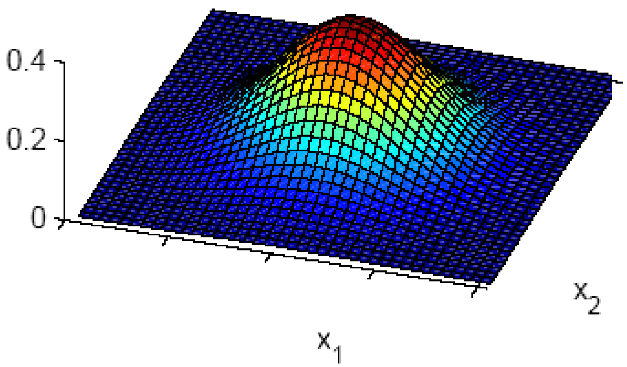
$$(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c^2$$

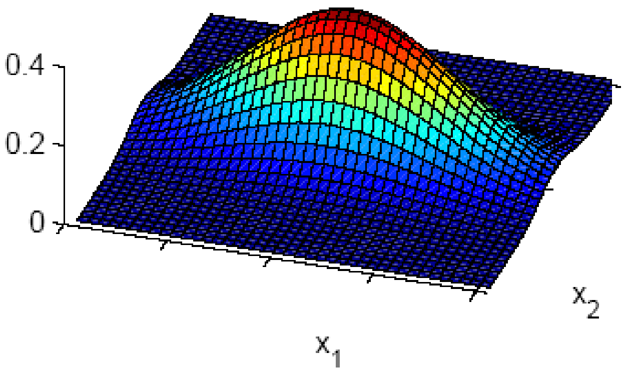




d-dimensional hyper-ellipsoid

center
shape
orientation

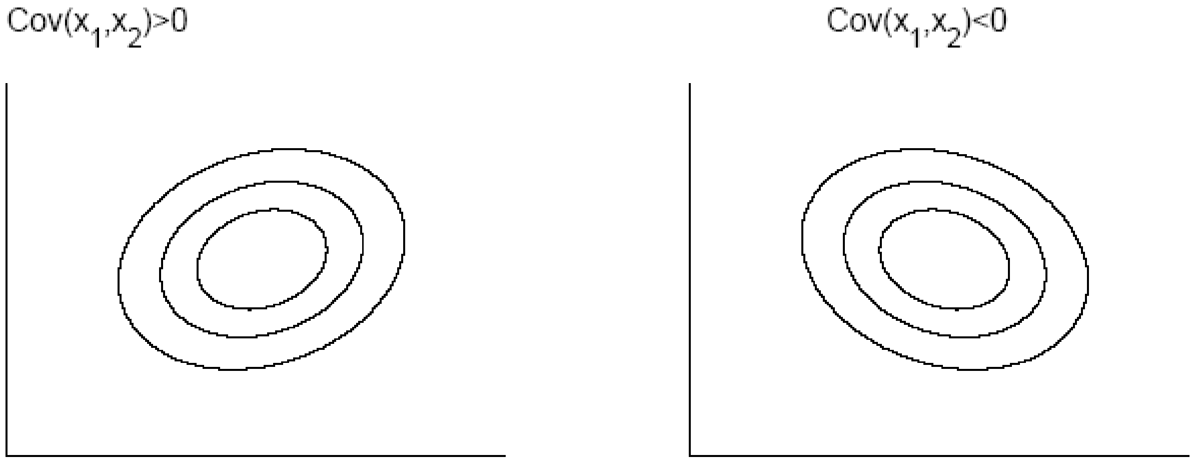




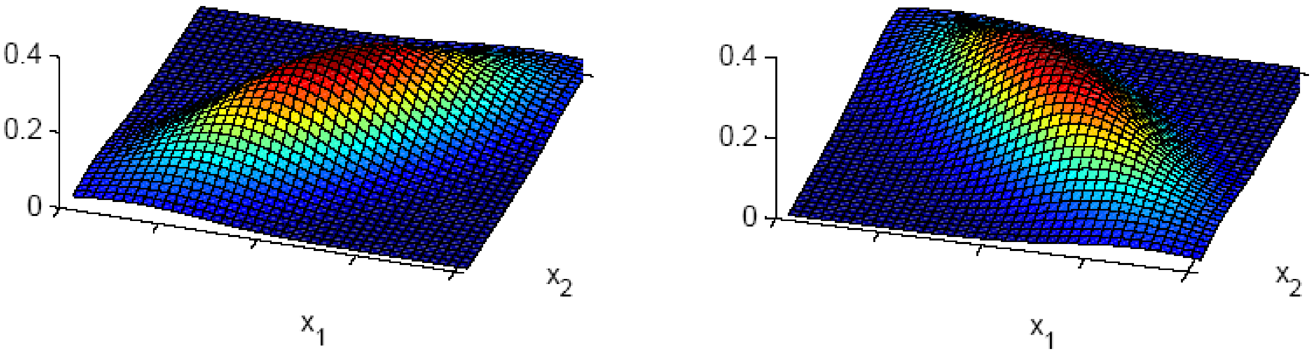
Bivariate Normal Distribution: isoprobable contours

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

d-dimensional hyper-ellipsoid



center
shape
orientation



Independent Inputs: Naive Bayes

- If x_i are independent, offdiagonals of $\sum$ are 0

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

  - Mahalanobis distance reduces to weighted (by $1/\sigma_i$) Euclidean distance:

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

- A normal distribution

standardization

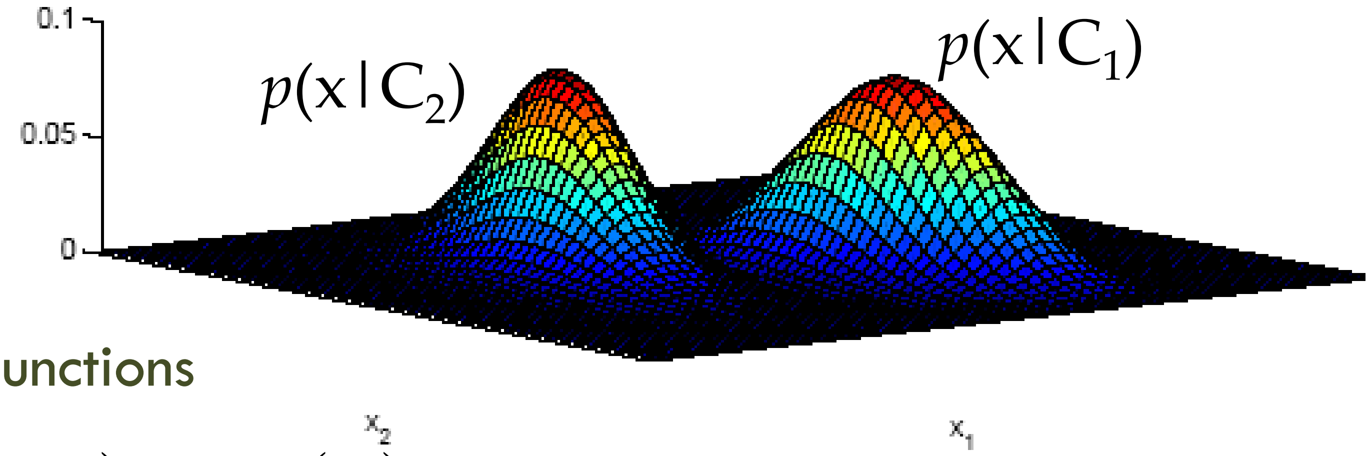
- If variances are also equal, reduces to Euclidean distance

Parametric Classification

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

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

likelihoods



- Discriminant functions

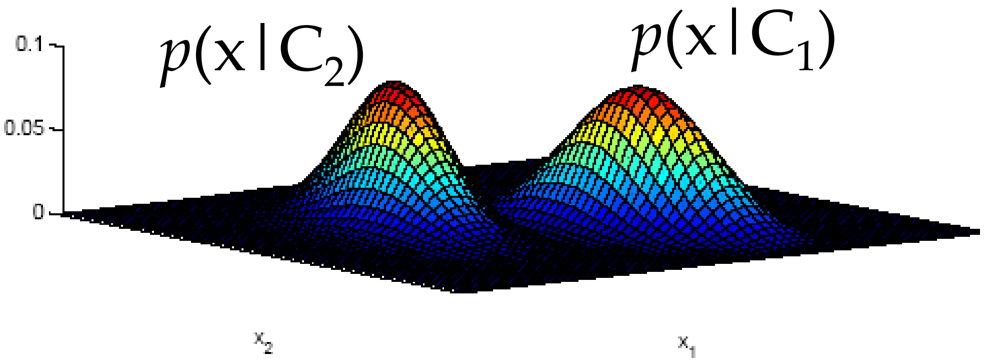
$$g_i(\mathbf{x}) = \log p(\mathbf{x}\,|\,C_i) + \log P(C_i)$$

$$= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i) + \log P(C_i)$$

Estimation of Parameters

$$g_i(\mathbf{x}) = -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i) + \log P(C_i)$$

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



$$\mu_i \approx \mathbf{m}_i = \frac{\sum_t r_i^t \mathbf{x}^t}{\sum_t r_i^t}$$

$$\Sigma_i \approx \mathbf{S}_i = \frac{\sum_t r_i^t (\mathbf{x}^t - \mathbf{m}_i)(\mathbf{x}^t - \mathbf{m}_i)^T}{\sum_t r_i^t}$$

Plugging in,

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

Case 1: Different $\mathbf{S}_i$

- Quadratic discriminant

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

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

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

where

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

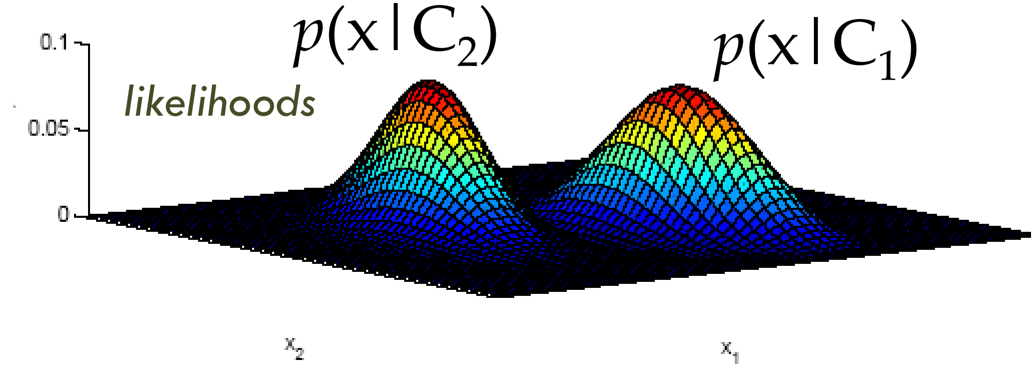
How many parameters?

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

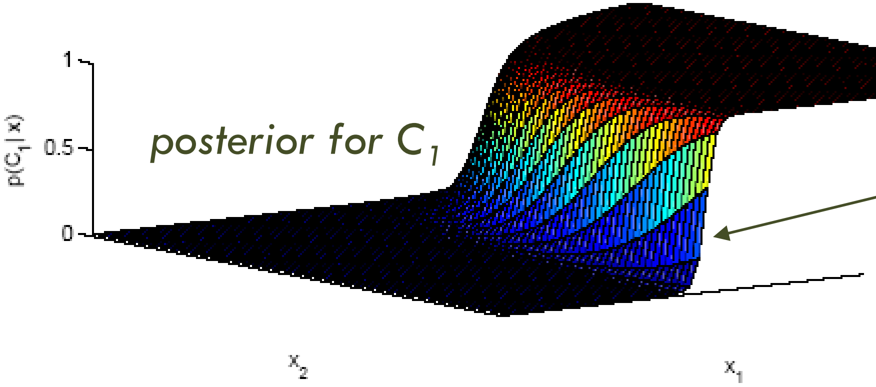
$$w_{i0} = -\frac{1}{2}\mathbf{m}_i^T\mathbf{S}_i^{-1}\mathbf{m}_i - \frac{1}{2}\log|\mathbf{S}_i| + \log\hat{P}(C_i)$$

$p(\mathrm{x}|\mathrm{C}_2)$ $p(\mathrm{x}|\mathrm{C}_1)$

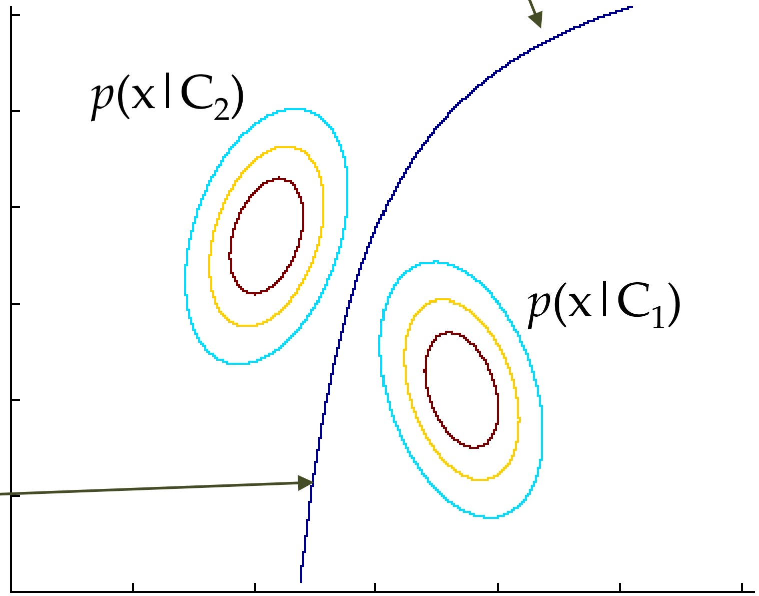
likelihoods



posterior for C_1



discriminant:
$P\ (\mathrm{C}_1 | \boldsymbol{x}) = 0.5$



$$g_1(x) = g_2(x)$$

$$\Updownarrow$$

$$x^T W_1 x + w_1 x + w_{1,0}$$
$$= x^T W_2 x + w_2 x + w_{2,0}$$

Case 2: Common/Shared Covariance Matrix **S**

- Shared common sample covariance **S**

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

- Discriminant reduces to

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

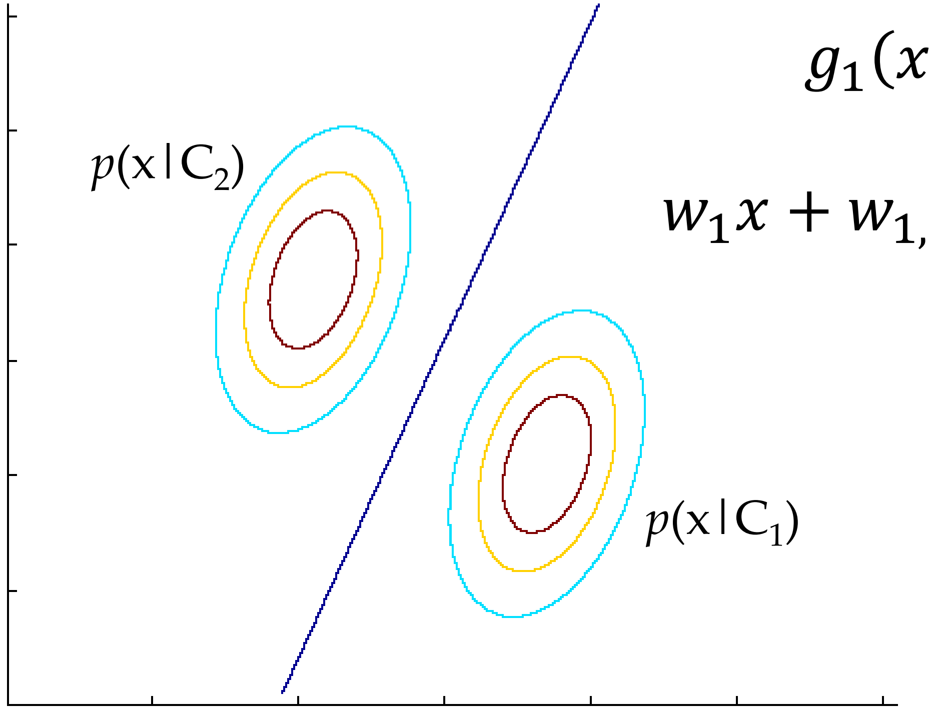
which is a linear discriminant (quadratic term cancels)

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

where

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

Common Covariance Matrix **S**



$$g_1(x) = g_2(x)$$

⇕

$$w_1 x + w_{1,0} = w_2 x + w_{2,0}$$

Case 3: Shared and diagonal S

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

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

Recall

$$p(x) = \prod_{j=1}^{d} p_j(x_j) = \frac{1}{(2\pi)^{d/2}\prod_j s_j}\exp\left(-\frac{1}{2}\sum_{j=1}^{d}\left(\frac{x_j - m_j}{s_j}\right)^2\right)$$

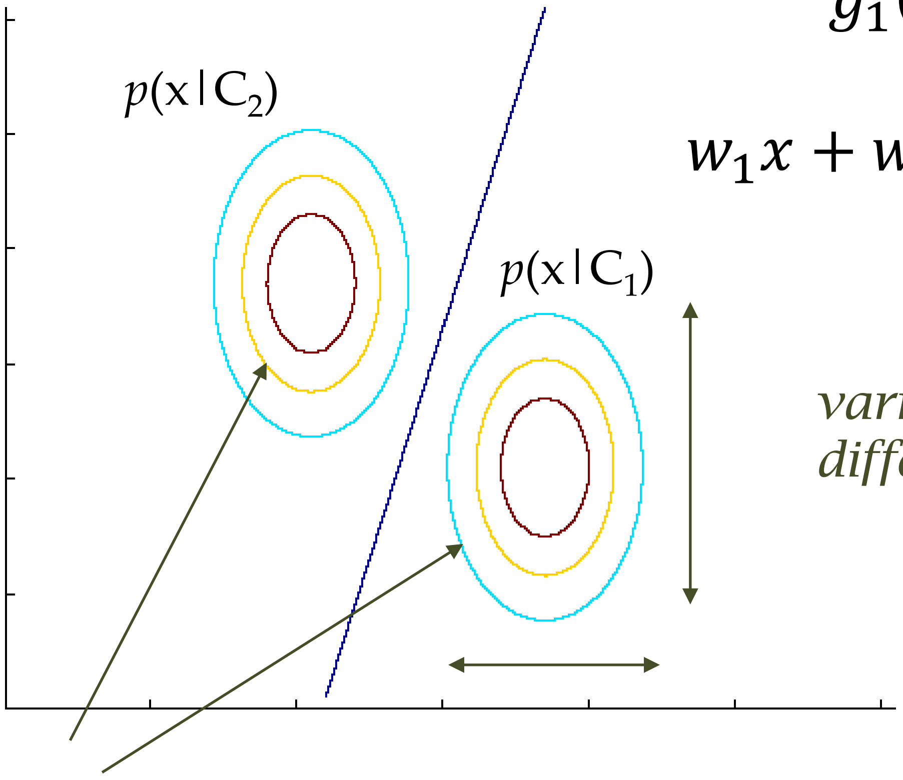
Classify based on weighted Euclidean distance (in s_j units) to the nearest mean

Diagonal S

$$g_1(x) = g_2(x)$$

$$\Updownarrow$$

$$w_1 x + w_{1,0} = w_2 x + w_{2,0}$$



variances may be different

must be axis-aligned (no oblique)

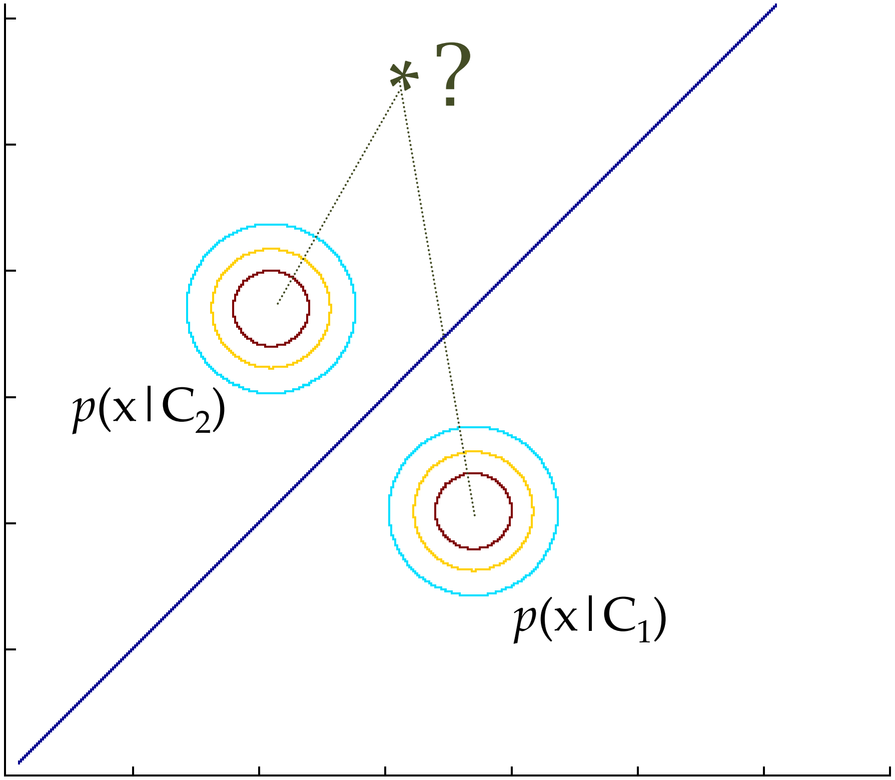
Case 4: Diagonal and shared **S**, and equal variances

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

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

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

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