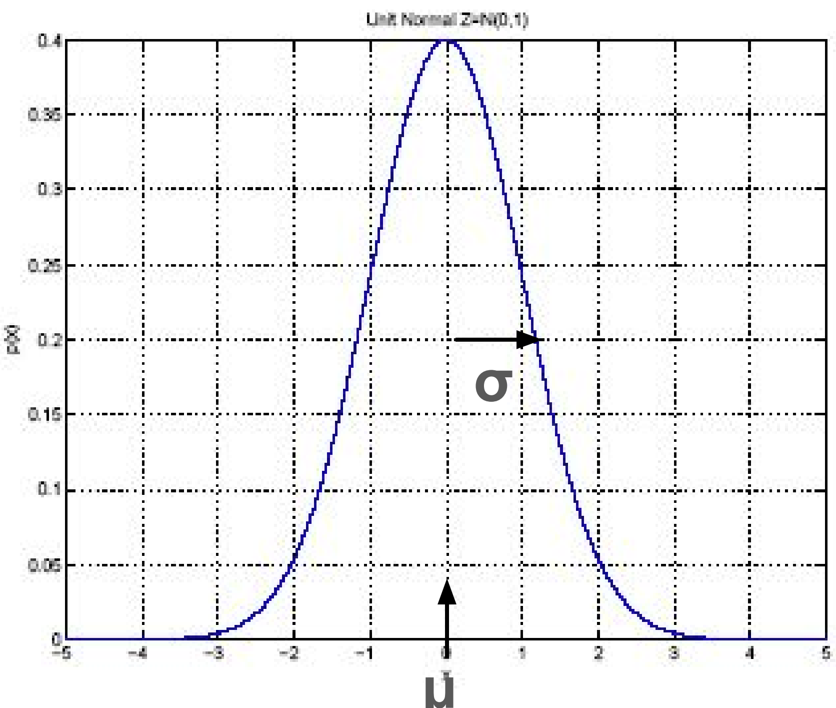
CENG 463
Machine Learning

Lecture 03 - Maximum Likelihood Estimation and Discriminants

Gaussian (Normal) Distribution

- μ: Mean
- σ: Standard deviation: average absolute difference from the mean
- σ2: Variance: average squared difference from the mean



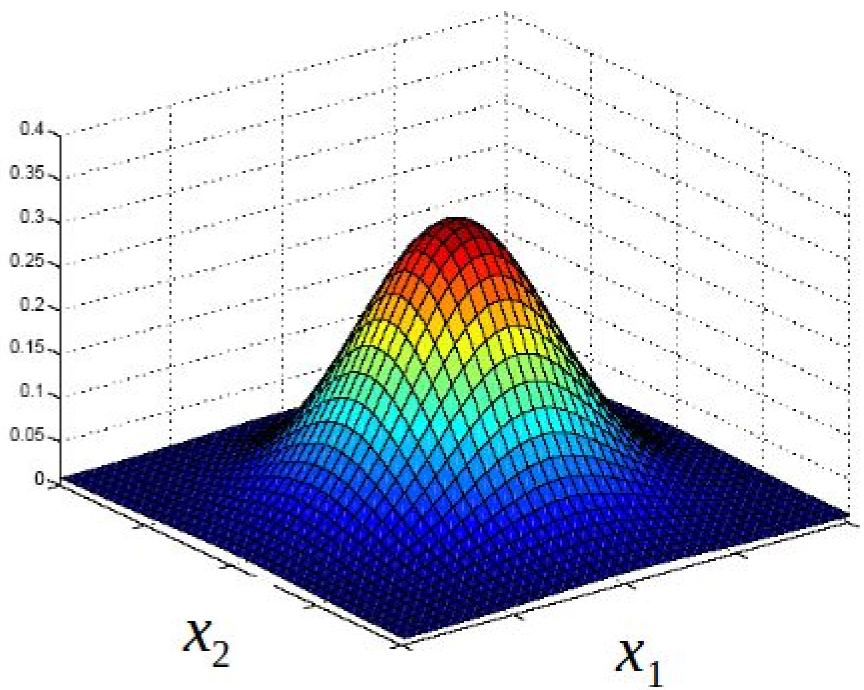
$$p(x) = N(\mu, \sigma^2)$$

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

Figure from Introduction to Machine Learning 2ed., E Alpaydın, 2010.

d-Dimensional Gaussian

Assume a d-dimensional sample set, X, (with N samples):



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

Covariance:

$$\sigma_{ij} = \frac{\sum_{t=1}^{N} (x_i^t - \mu_i)(x_j^t - \mu_j)}{N}$$

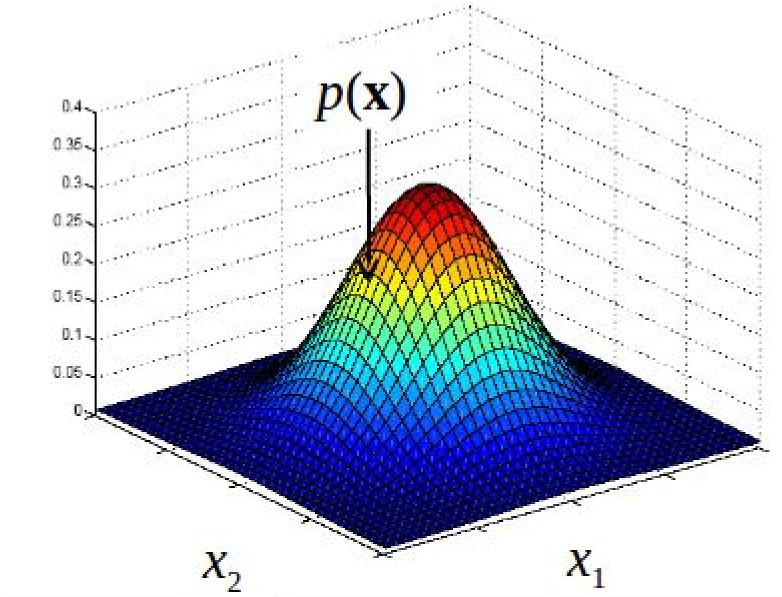
Covariance matrix:

$$\Sigma \equiv \mathrm{Cov}(\mathbf{x}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & & & \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

Figure from Introduction to Machine Learning 2ed., E Alpaydın, 2010.

d-Dimensional Gaussian

The probability of a new sample/location, x=(x1, x2,..., xd), in this d-dimensional space is computed using μ and Σ.

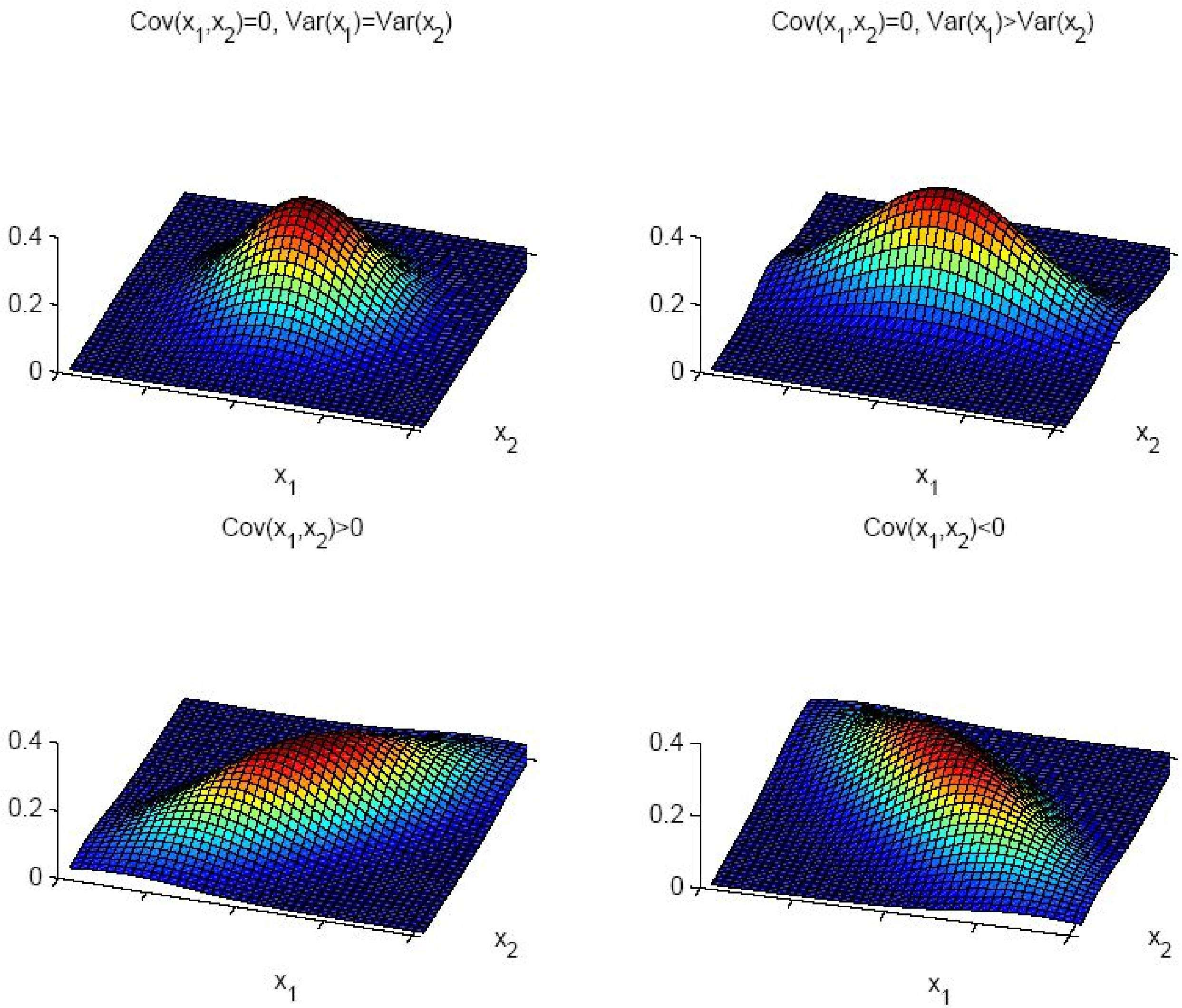


$$p(x_1, x_2, .. x_d) = p(x) = N_d(\mu, \Sigma)$$

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

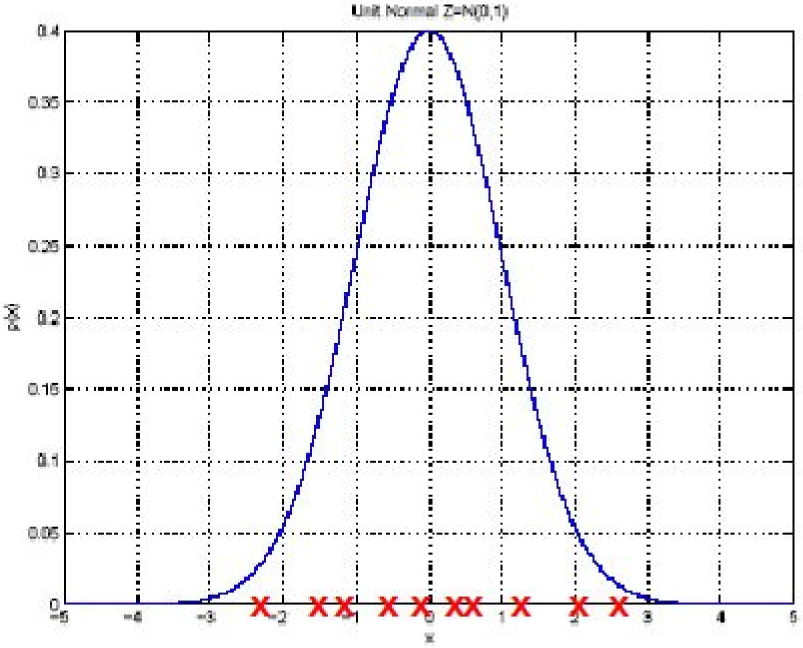
Figure from Introduction to Machine Learning 2ed., E Alpaydın, 2010.

2D Gaussian Examples



Maximum Likelihood Estimation

- MLE is the way to find the unknown parameters of the distribution of given data.



If you are given a dataset and if you know its PDF for a certain class, p(X|C), is a Gaussian distribution, MLE estimates the parameters μ and σ^2.

Maximum Likelihood: 1D Gaussian

In brief: Use the given samples to estimate the unknown Gaussian parameters (μ, σ^2)

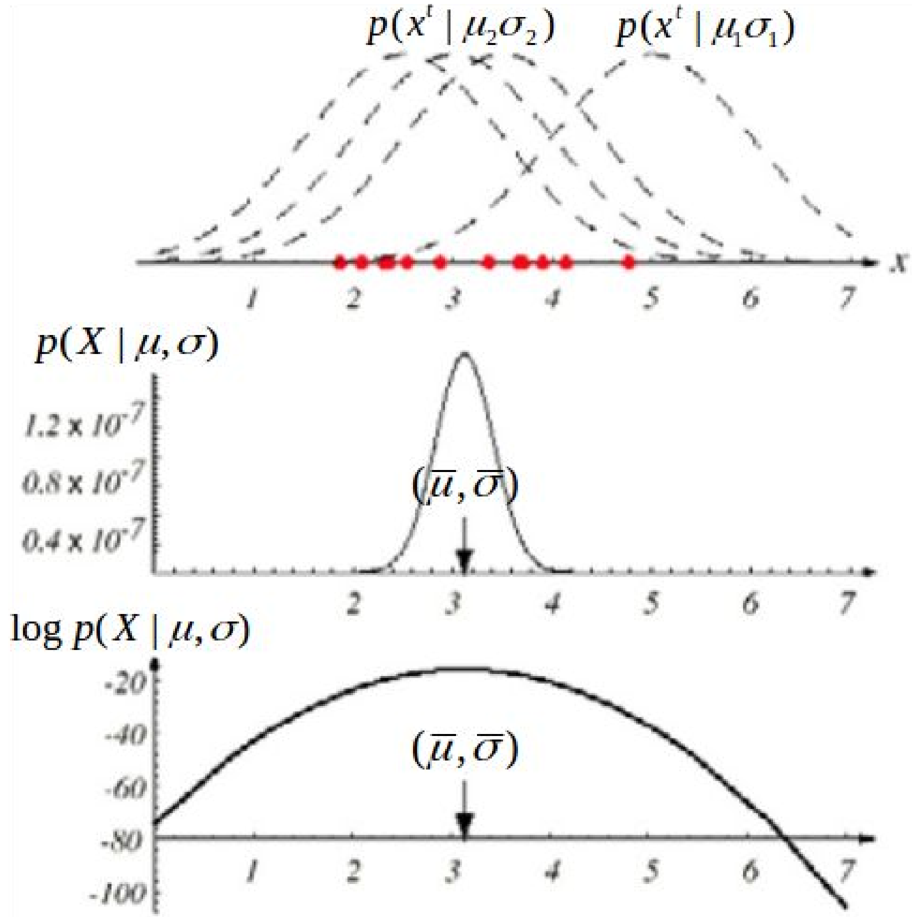
- Let a sample set, X, (with N samples), $X=\{x^1,\ldots,x^N\}$.
- Since the samples are independently chosen:

$$p(X \mid \mu, \sigma) = \prod_{t=1}^{N} p(x^t \mid \mu, \sigma)$$

- To find the parameters that maximize $p(X|\mu,\sigma)$, we differentiate it (take the derivative) and equate to zero.

Maximum Likelihood: 1D Gaussian

- For different (μ,σ), the observed samples give different $p(x^t|\mu,\sigma)$ values, resulting in different $p(X|\mu,\sigma)$.
- The argument for the maximum of such products is ML estimate.
- Using $\log p(X|\mu,\sigma)$ does not change the location of maxima.

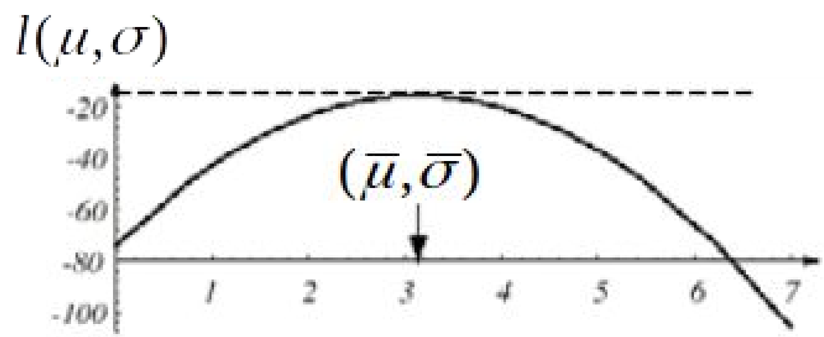


Maximum Likelihood: 1D Gaussian

- Better to work with logarithm for analytical purposes (as mentioned taking logarithm does not affect the maxima).
- Differentiate log likelihood, l(μ,σ) and equate it to zero to locate the parameters with maximum likelihood.

$$l(\mu,\sigma) = \log p(X \mid \mu,\sigma) = \sum_{t=1}^{N} \log p(x^t \mid \mu,\sigma)$$

$$\nabla l(\mu,\sigma) = \sum_{t=1}^{N} \nabla \log p(x^t \mid \mu,\sigma) = 0$$



Maximum Likelihood: 1D Gaussian

For 1D(univariate) Gaussian distribution:

$$\log p(x^t \mid \mu, \sigma) = -\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma}(x^t - \mu)^2$$

Differentiate:

$$\nabla_\mu l(\mu,\sigma) = 0 \longrightarrow \sum_{t=1}^{N}\frac{1}{\sigma}(x^t - \mu) = 0 \longrightarrow m = \frac{\sum_t x^t}{N}$$

$$\nabla_\sigma l(\mu,\sigma) = 0 \xrightarrow{\text{derivation is not shown*}} s^2 = \frac{\sum_t (x^t - m)^2}{N}$$

m, s^2 are the ML estimates for μ, σ^2 .

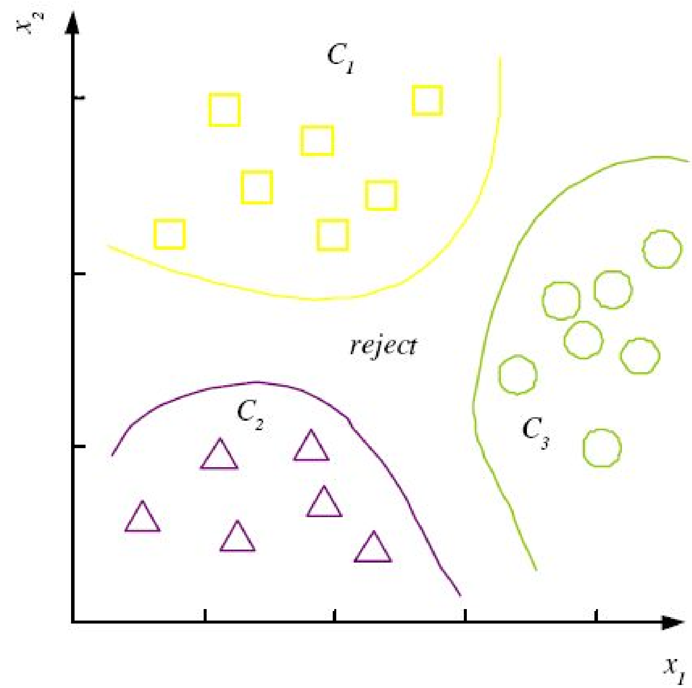
We could also use $(\bar{\mu}, \bar{\sigma}^2)$ to indicate that they are estimates.

Discriminant Functions

Remember our discriminant function using the maximum posterior or minimum risk:

$$\text{choose } C_i \text{ if } g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})$$

$$g_i(\mathbf{x}) = \begin{cases} -R(\alpha_i \mid \mathbf{x}) & \leftarrow \text{minimum risk} \\ P(C_i \mid \mathbf{x}) & \leftarrow \text{maximum posterior} \\ p(\mathbf{x} \mid C_i)P(C_i) & \leftarrow \text{unnormalized posterior} \end{cases}$$



K decision regions $R_1,\ldots,R_K$ $\longrightarrow$ $R_i = \{\mathbf{x} \mid g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})\}$

Figure from Introduction to Machine Learning 2ed., E Alpaydın, 2010.

Discriminant Function for 1D Gaussian

Remember our discriminant function using the posterior

$$g_i(x) = P(x \mid C_i) P(C_i)$$

or

$$g_i(x) = \log P(x \mid C_i) + \log P(C_i)$$

Assuming samples are coming from a Gaussian distribution

$$P(x \mid C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[- \frac{(x - \mu_i)^2}{2\sigma_i^2} \right]$$

Since Gaussian is exponential, we prefer log version:

$$g_i(x) = - \frac{1}{2} \log 2\pi - \log \sigma_i - \frac{(x - \mu_i)^2}{2\sigma_i^2} + \log P(C_i)$$

Discriminant Function for Given Data

Given the sample data where **r** is the label:

$$\mathcal{X} = \{x^t, r^t\}_{t=1}^{N} \quad r_i^t = \begin{cases} 1 \text{ if } x^t \in C_i \\ 0 \text{ if } x^t \in C_j,\ j \neq i \end{cases}$$

Prior and parameter estimates:

$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N} \quad m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t} \quad s_i^2 = \frac{\sum_t (x^t - m_i)^2 r_i^t}{\sum_t r_i^t}$$

Discriminant becomes:

$$g_i(x) = -\frac{1}{2}\log 2\pi - \log s_i - \frac{(x - m_i)^2}{2s_i^2} + \log \hat{P}(C_i)$$

Discriminant Function for Given Data

Simplifying discriminant function:

$$g_i(x) = \cancel{-\frac{1}{2}\log 2\pi} \cancel{- \log s_i} - \frac{(x - m_i)^2}{2s_i^2} \cancel{+ \log \hat{P}(C_i)}$$

constant in all g_i

if priors are equal

If also variances are equal, discriminant becomes:

$$g_i(x) = -(x - m_i)^2$$

which means a new sample is labeled to the class with the closest mean.

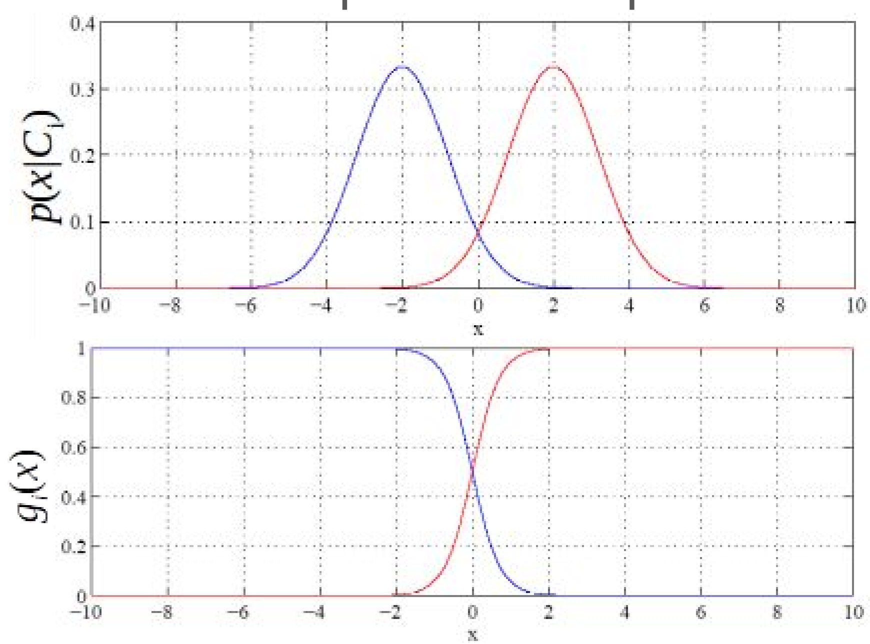


Figure from Introduction to Machine Learning 2ed., E Alpaydın, 2010.

Numerical Example

$$X = [\mathbf{x}\ \ \mathbf{r}]$$

$$X = \begin{bmatrix} 50 & 1 \\ 40 & 1 \\ 30 & 1 \\ 15 & 1 \\ 15 & 1 \\ 30 & 2 \\ 20 & 2 \\ 10 & 2 \\ 10 & 2 \\ 5 & 2 \end{bmatrix}$$

$$m_i = \frac{\sum_t x_i^t}{N}$$

$$m_1 = \frac{50+40+30+15+15}{5} = 30$$

$$m_2 = \frac{30+20+10+10+5}{5} = 15$$

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

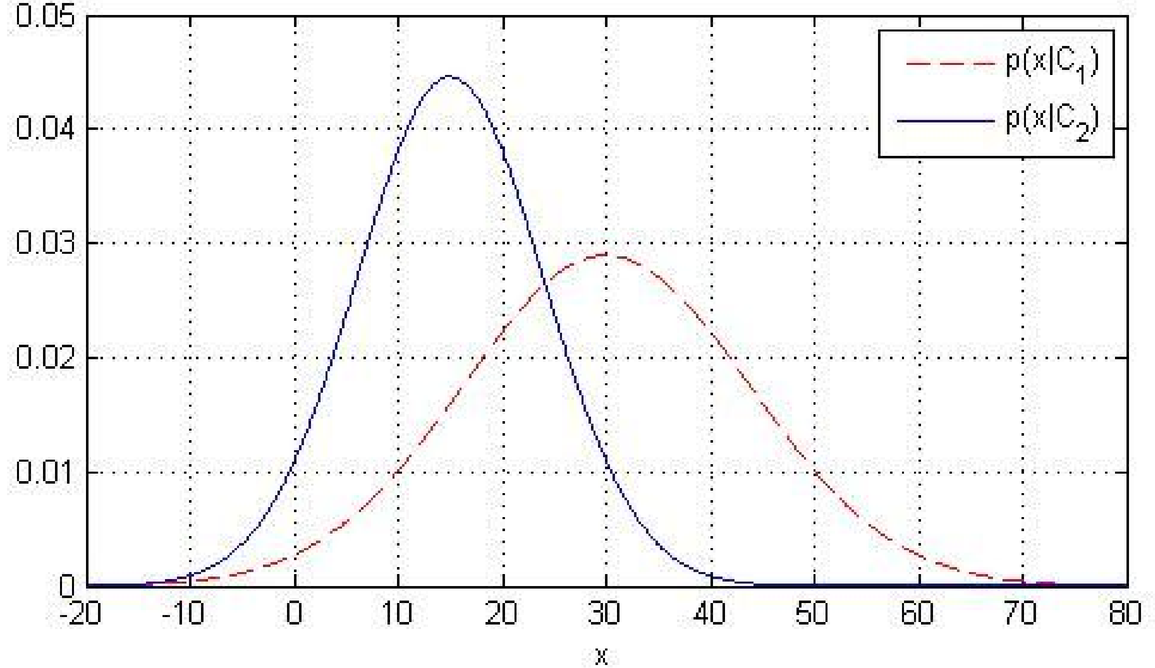
$$s_1^{\ 2} = \frac{20^2+10^2+0+15^2+15^2}{5} = 190$$

$$s_2^{\ 2} = \frac{15^2+5^2+5^2+5^2+10^2}{5} = 80$$

How do the likelihoods (Gaussians) look like?

Numerical Example

Gaussians look like:



$m_1 = 30 \qquad m_2 = 15 \qquad s_1^{\ 2} = 190 \qquad s_2^{\ 2} = 80$

Numerical Example

Priors are equal, $\hat{P}(C_i) = \frac{\sum_t r_i^t}{N} = \frac{5}{10}$ for each class.

Discriminant function becomes: $g_i(x) = -\log s_i - \frac{(x - m_i)^2}{2s_i^2}$

$$g_1(x) = -\log\sqrt{190} - \frac{(x-30)^2}{2 \cdot 190} = -2.62 - \frac{(x-30)^2}{2 \cdot 190}$$

$$g_2(x) = -\log\sqrt{80} - \frac{(x-15)^2}{2 \cdot 80} = -2.19 - \frac{(x-15)^2}{2 \cdot 80}$$

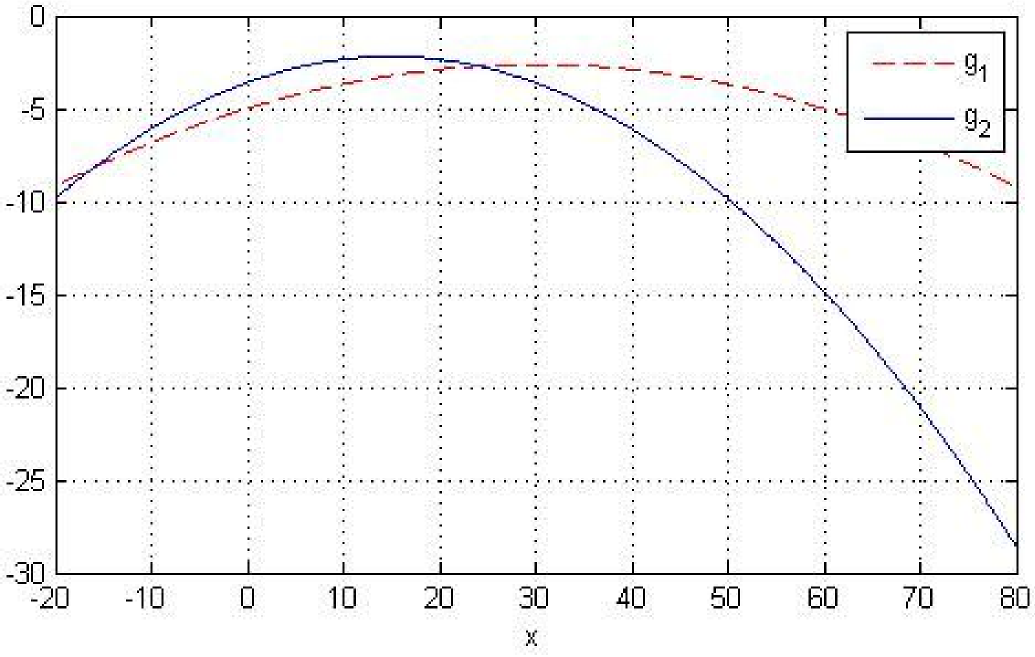
Now we can apply these discriminants to new samples:

$g_1(10) = -3.67$ $\quad g_1(20) = -2.88$ $\quad g_1(30) = -2.62$

$g_2(10) = -2.43$ $\quad g_2(20) = -2.34$ $\quad g_2(30) = -3.59$

Numerical Example

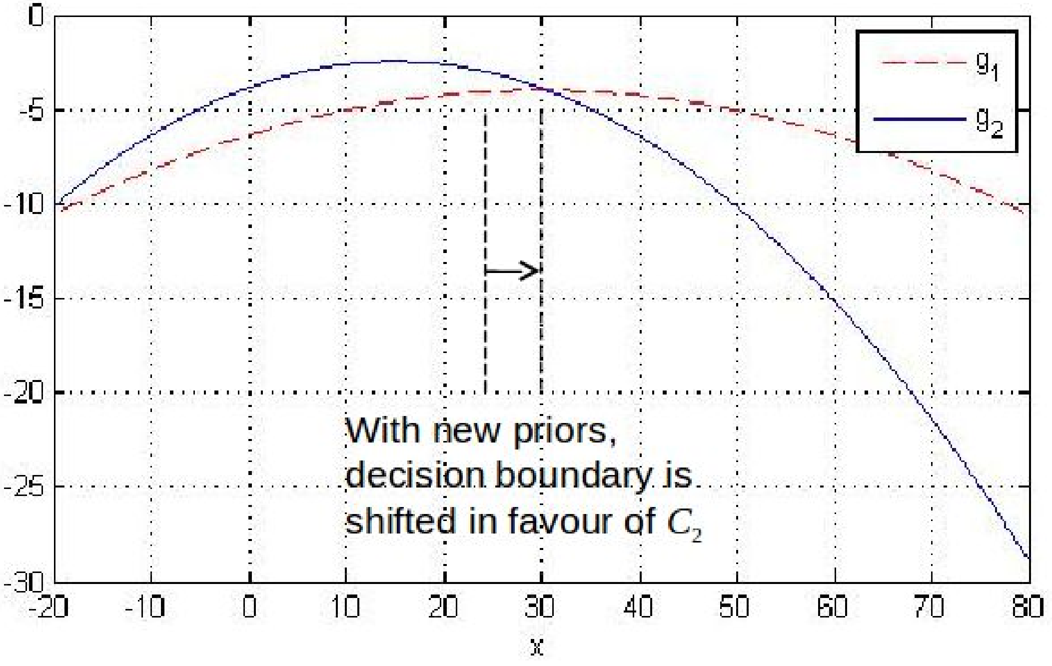
Discriminant functions look like:



Numerical Example

What if priors were not equal?

Let $\hat{P}(C_1) = 0.25$ and $\hat{P}(C_2) = 0.75$



Maximum Likelihood: General Form

- The underlying distribution for data X does not have to be Gaussian.
- Maximum likelihood estimation can be performed for any parametric function.
- Let the parameters be Θ, then, the likelihood:

$$p(X \mid \Theta) = \prod_{t=1}^{N} p(x^t \mid \Theta)$$

- and log likelihood:

$$l(\Theta) = \log p(X \mid \Theta) = \sum_{t=1}^{N} \log p(x^t \mid \Theta)$$

Multivariate Example

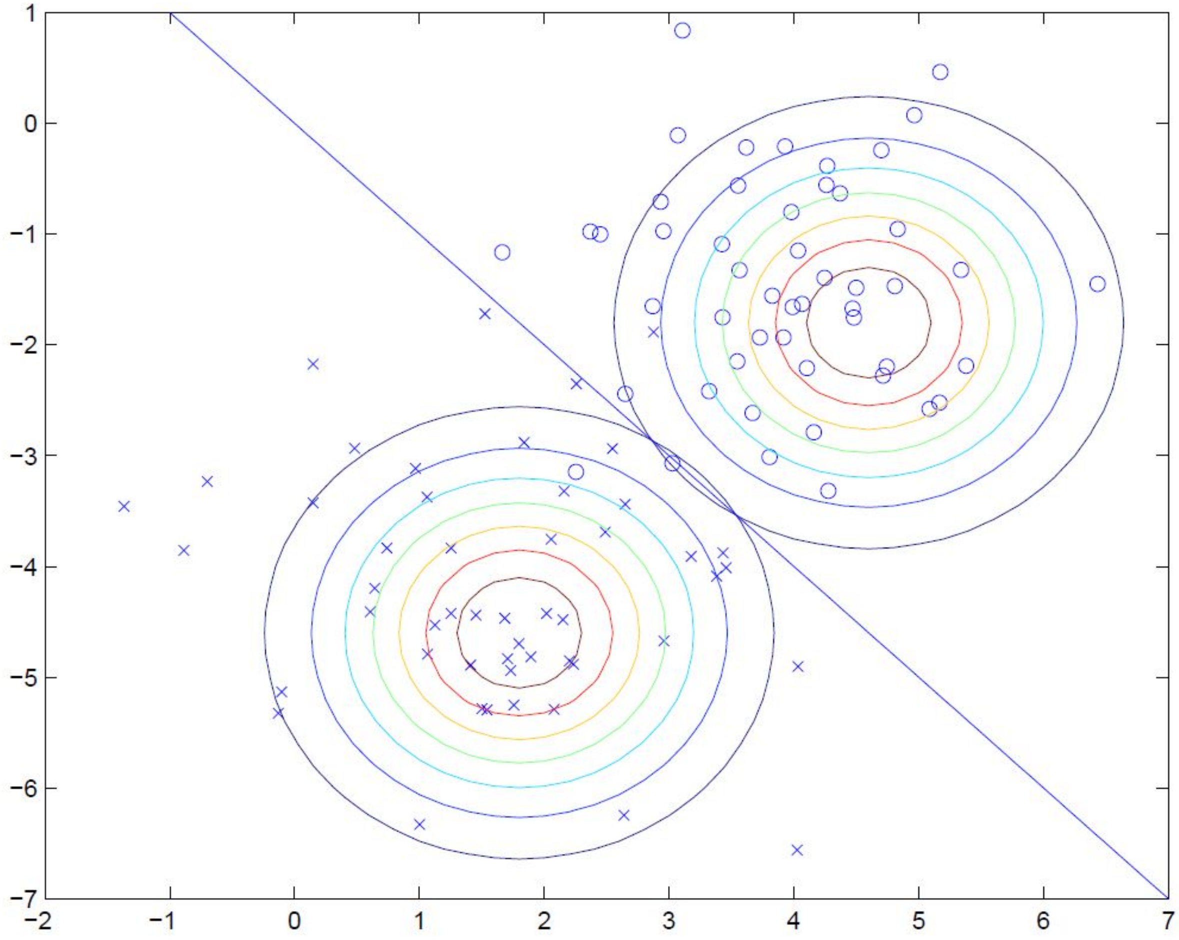


Figure from the Notes of Andrew NG.

Summary

We have learned about:

- Gaussian Distribution
- Maximum Likelihood Estimates for Gaussian Distribution
- Discriminant Functions for Data of 1D Gaussian