Chapter 2 (Part 2): Bayesian Decision Theory (Sections 2.4-2.5)

- Multi-category classification
- Classifiers, Discriminant Functions and Decision Surfaces
- The Normal Density

Classifiers, Discriminant Functions and Decision Surfaces

- The multi-category case
  - Set of discriminant functions $g_i(x), i = 1,\ldots, c$
  - The classifier assigns a feature vector x to class ω_i if:

$$g_i(x) > g_j(x) \; \forall j \neq i$$

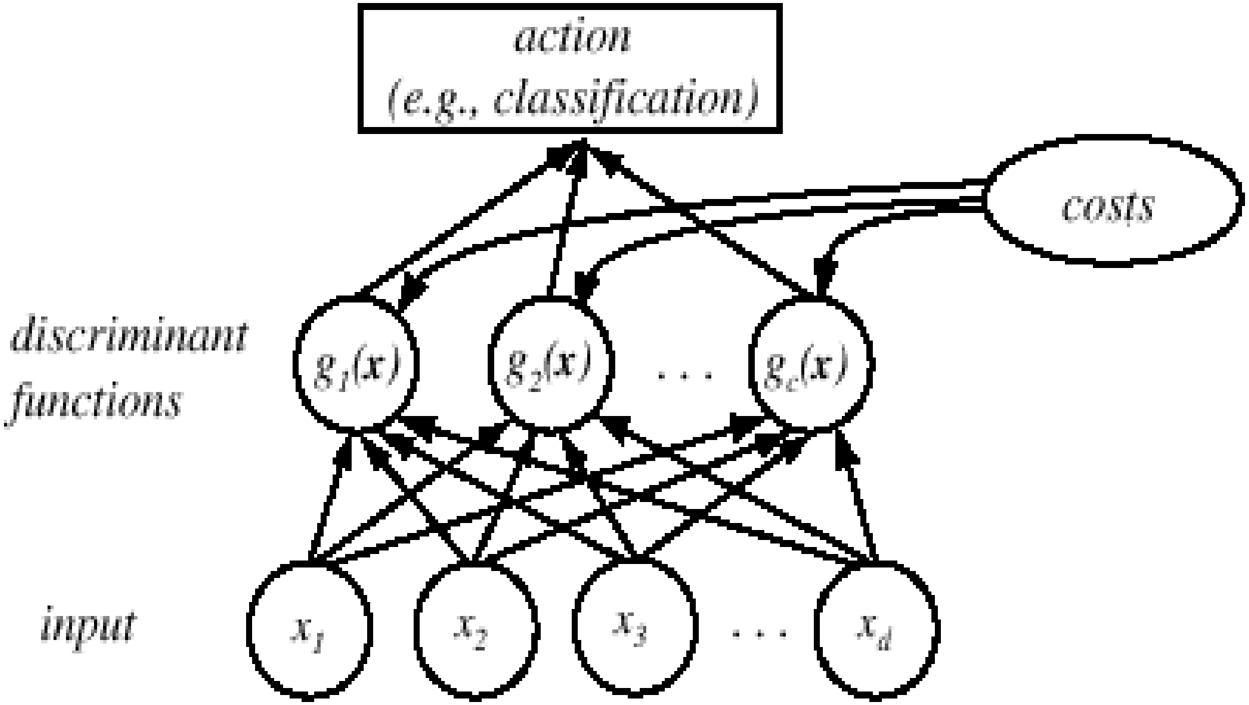


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- *For minimum Risk action, use $g_i(x) = - R(\alpha_i \mid x)$* (max. discriminant corresponds to min. risk!)

- For the minimum error rate classification, we take

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

(max. discrimination corresponds to max. posterior!)

$$g_i(x) \equiv P(x \mid \omega_i)\, P(\omega_i)$$

We get the same classifier even when logarithm of the discriminants

$$g_i(x) = \ln P(x \mid \omega_i) + \ln P(\omega_i)$$

(ln: natural logarithm!)

- Feature space divided into c decision regions

$$\textit{if } g_i(x) > g_j(x)\ \forall j \neq i \textit{ then } x \textit{ is in } \mathcal{R}_i$$

($\mathcal{R}_i$ means assign x to ω_i)

- The two-category case
  - A classifier is a "dichotomizer" that has two discriminant functions g_1 and g_2

    Let $g(x) \equiv g_1(x) - g_2(x)$

    Decide ω_1 *if* $g(x) > 0$; Otherwise decide ω_2

Dichotomizer

- The computation of g(x)

$$
\begin{aligned}
g(x) &= P(\omega_1 \,|\, x) - P(\omega_2 \,|\, x) \\
&= \ln \frac{P(x \,|\, \omega_1)}{P(x \,|\, \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}
\end{aligned}
$$

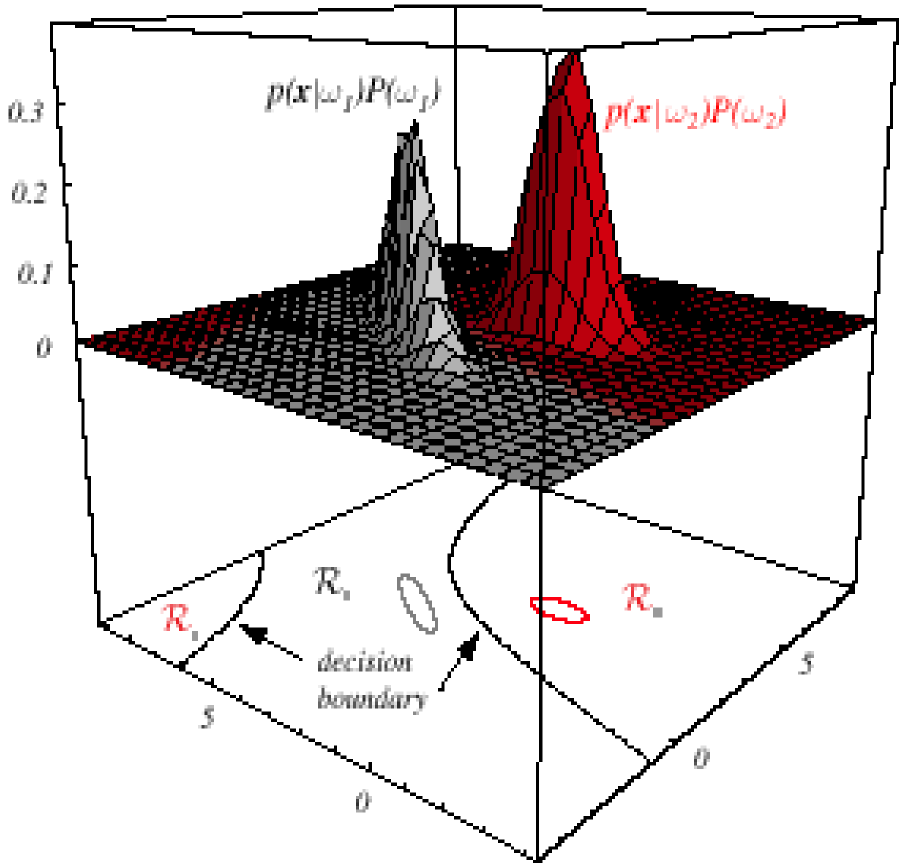


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

The Normal Density

- Univariate density

  - Density which is analytically tractable
  - Continuous density
  - A lot of processes are asymptotically Gaussian
  - Handwritten characters, speech sounds are ideally of this. Prototypes corrupted by random process (central limit theorem)

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right],$$

Where:

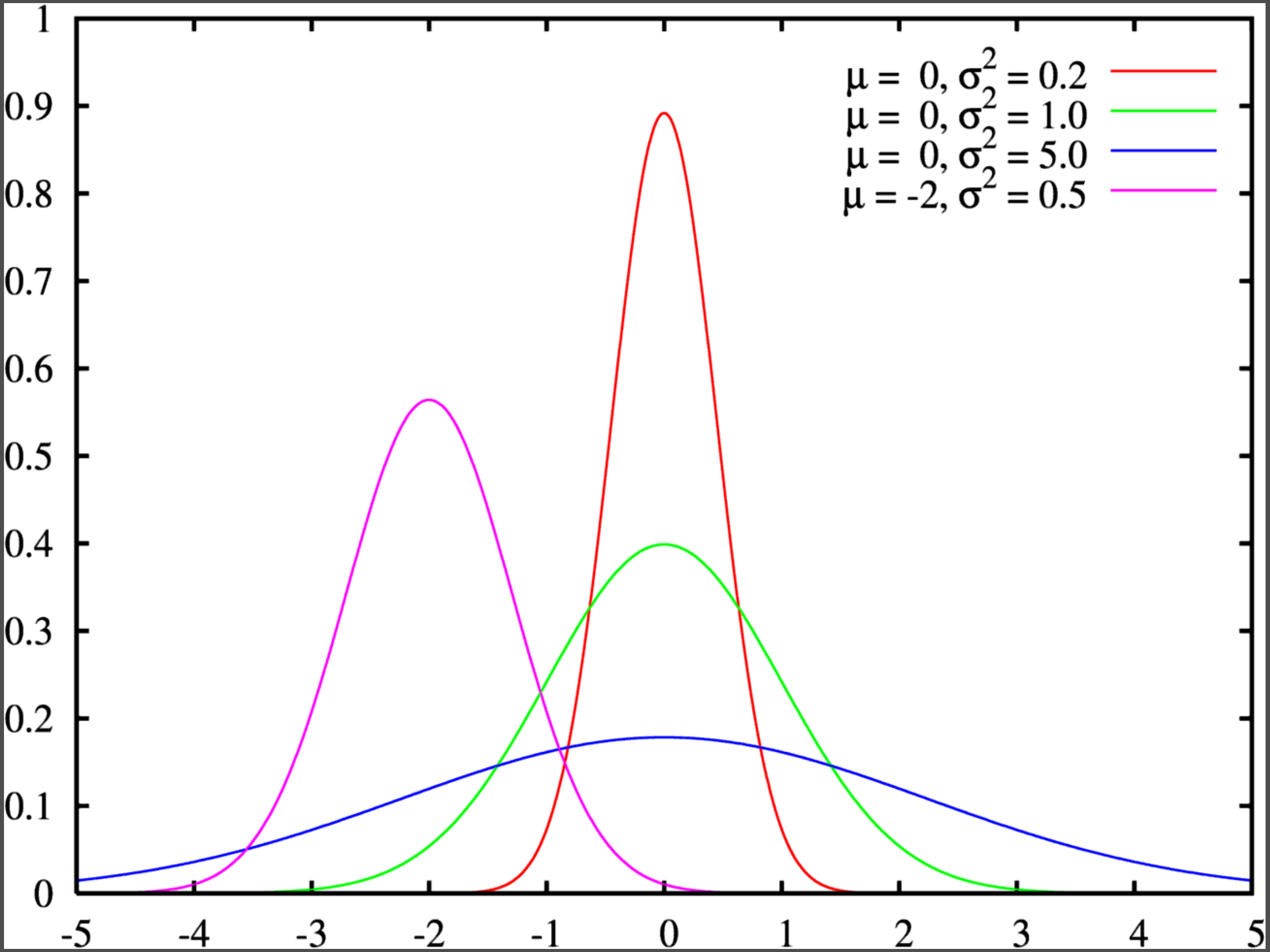
μ = mean (or expected value) of x = E[x]

σ^2 = expected squared deviation or variance = E[(x - μ)^2]

Normal Density

- $\mu = E[\,x\,] = \int_{-\infty}^{+\infty} x\, p(x)\, dx$
- $\sigma^2 = E[(x - \mu)^2] = \int (x - \mu)^2\, p(x)\, dx$
- Normal density has only two parameters, viz., μ and σ^2
- Normal density is often abbreviated as
  $p(x) \sim N(\mu, \sigma^2)$

Gaussian or Normal Distribution



Sample estimates

- Let D = $\{x^1, x^2, \ldots, x^n\}$ be a given training set of points, then from this sample we can estimate μ and σ^2 as follows:
  - $\mu \approx (1/n) \Sigma x^i$
  - $\sigma^2 \approx (1/n) \Sigma (x^i - \mu)^2$
- We will see later, how we got these !

Normal (Gaussian) Distribution

- It is also called *"Gaussian Distribution"*
- ***Central Limit theorem (informally):*** The aggregate effect of a large number of small, independent random disturbances will lead to a Gaussian distribution.
- Because, in many cases, the patterns are corrupted versions of some ideal prototypes, so their distributions are indeed Gaussian.
- In addition, the normal distribution maximizes ***information entropy*** among all distributions with known mean and variance.

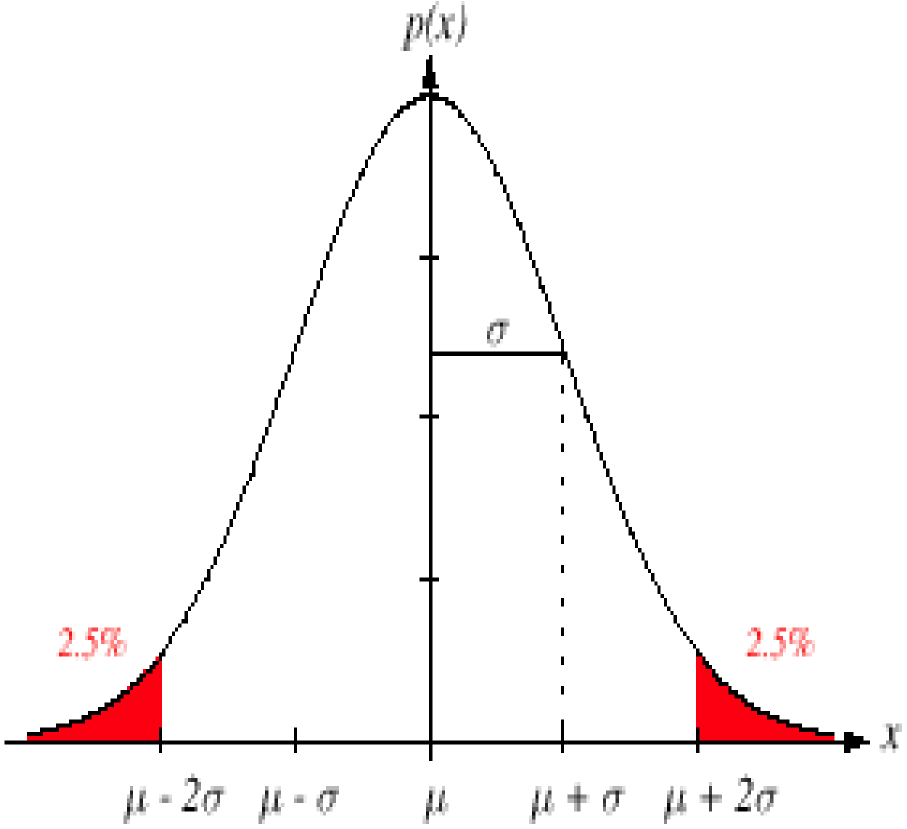


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Multivariate Normal Density

- Multivariate density
  - Multivariate normal density in d dimensions is:

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

where:

$\mathbf{x} = (x_1, x_2, \ldots, x_d)^t$ (t stands for the transpose vector form)

$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_d)^t$ mean vector

$\Sigma = d*d$ covariance matrix

$|\Sigma|$ and Σ^{-1} are determinant and inverse respectively

Covariance Matrix

$$\Sigma = E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^t]$$
$$= \int (\mathbf{x} - \mu)(\mathbf{x} - \mu)^t p(\mathbf{x})\, d\mathbf{x}$$

Let there are d features

A pattern $\mathbf{x} = (x_1, x_2, \ldots, x_d)^t$

Similarly $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_d)^t$

The ij th component of Σ is $\sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]$

This is covariance of the two features.

(how much one feature varies w.r.t to the other)

Sample covariance matrix

- From a given sample, an estimate of covariance matrix, called *sample covariance matrix* can be found as:

$$(1/n) \sum (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^t$$

- Let $\{ (-1,-2)^t, (0,1)^t, (2,0)^t, (3,1)^t \}$ be the sample. Find its covariance matrix.

Example

- Sample = { $(-1,-2)^t$, $(0,1)^t$, $(2,0)^t$, $(3,1)^t$ }

  $\boldsymbol{\mu} = (1,0)^t$

  zero mean normalized sample = { $(-2,-2)^t$, $(-1,1)^t$, $(1,0)^t$, $(2,1)^t$ }

  Sample covariance matrix =

$$(1/4)\,[\,(-2,-2)^t(-2,-2) + (-1,1)^t(-1,1) + (1,0)^t(1,0) + (2,1)^t(2,1)\,] = \begin{bmatrix} 10/4 & 5/4 \\ 5/4 & 6/4 \end{bmatrix}$$

Covariance Matrix

- The covariance matrix Σ is always symmetric and positive semidefinite.
- Positive semidefinite means $\mathbf{x}^t \Sigma \mathbf{x} \geq 0$, for any pattern **x**
- If **a** is unit vector in some direction, then $\mathbf{a}^t \Sigma \mathbf{a}$ is the variance of the projected patterns onto **a**
  - Projection of **x** onto **a** is: $\mathbf{a}^t \mathbf{x}$
- Hence, knowledge of the covariance matrix allows us to calculate the dispersion of the data in any direction, or in any subspace.

Covariance matrix

- Let $p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \Sigma)$
- Let A be a d-by-k matrix and $\mathbf{y} = A^t \mathbf{x}$
- Then $p(y) \sim N(A^t \boldsymbol{\mu}, \ A^t \Sigma A)$
- It is possible to find a linear transformation A_w , so that in the new space, $\Sigma = I$. We will see about this when we discuss

  Principal Component Analysis

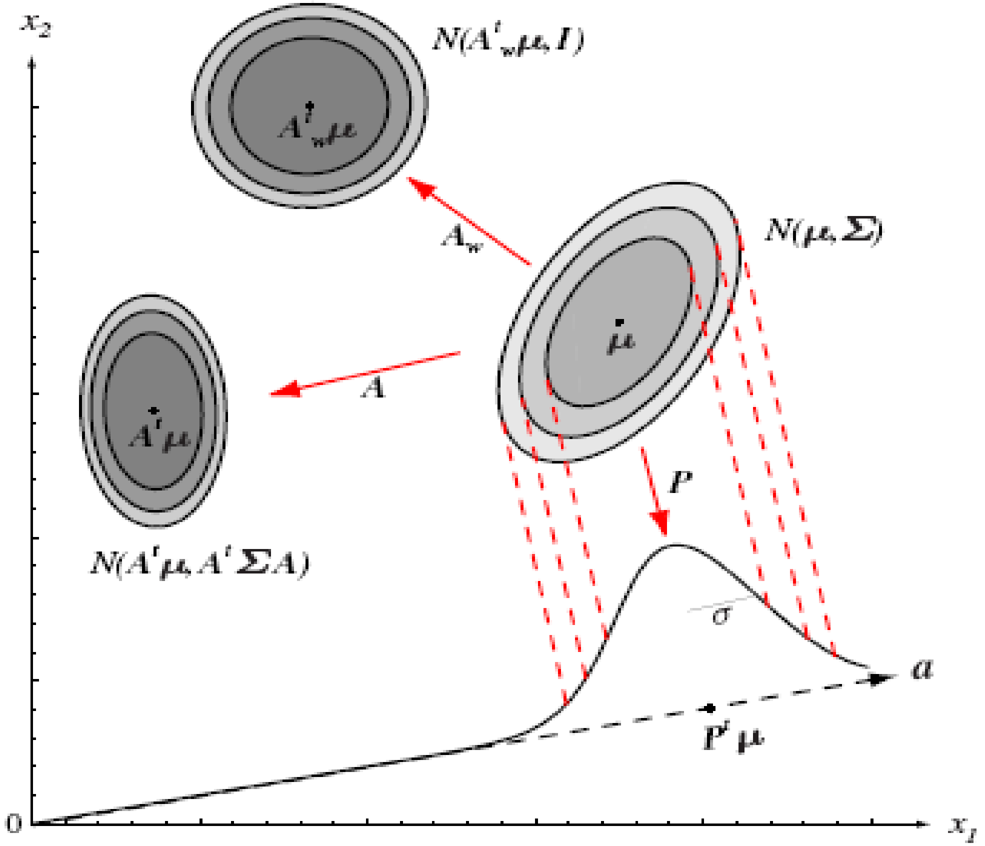


FIGURE 2.8. The action of a linear transformation on the feature space will convert an arbitrary normal distribution into another normal distribution. One transformation, **A**, takes the source distribution into distribution $N(\mathbf{A}^t\boldsymbol{\mu}, \mathbf{A}^t\Sigma\mathbf{A})$. Another linear transformation—a projection **P** onto a line defined by vector **a**—leads to $N(\mu, \sigma^2)$ measured along that line. While the transforms yield distributions in a different space, we show them superimposed on the original x_1x_2-space. A whitening transform, $\mathbf{A}_w$, leads to a circularly symmetric Gaussian, here shown displaced. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Multivariate Normal Density

- The multivariate normal density is completely specified by $d + d(d+1)/2$ parameters.
- Samples drawn from a normal population tend to fall in a single cloud or cluster.
- The center of the cluster is determined by the mean vector, and the shape of the cluster is determined by the covariance matrix.
- Loci of points of constant density are hyperellipsoids for which the quantity

  $(\mathbf{x}-\boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})$ is constant.

Mahalanobis Distance

- $r^2 = (\mathbf{x}-\boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})$ is called the squared *Mahalanobis* distance.
- Thus, the contours of constant density are hyperellipsoids of constant Mahalanobis distance.

1893 – 1972

Founder of Indian Statistical Institute

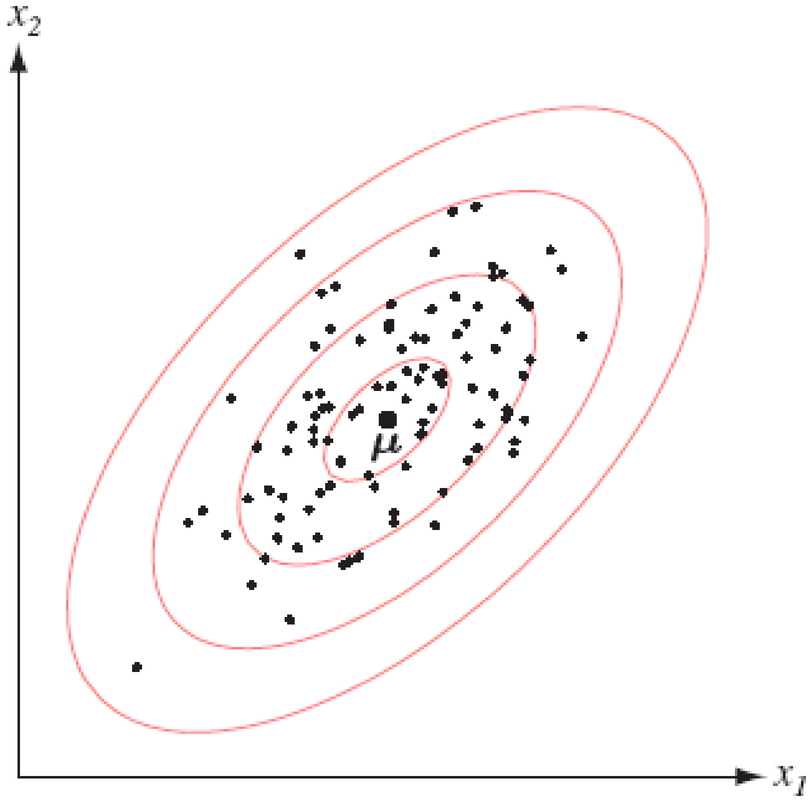


FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean $\boldsymbol{\mu}$. The ellipses show lines of equal probability density of the Gaussian. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Next …

- Discriminant functions for the Normal Density.
- Bayes decision theory with discrete features.