Machine Learning

Bayesian Parameter Estimation
and Maximum Likelihood Estimation

Indian Institute of Information Technology
Sri City, Chittoor

Today's Agenda

- Classifiers, Discriminant Functions and Decision Surfaces
  - Multi category case
  - Two category case
- The Normal Density
  - Univariate
  - Multivariate
- Discriminant Functions for the Normal Density

Classifiers, Discriminant Functions and Decision Surfaces

1. The Multi-Category Case:

Consider representing a classifier as in terms of a set of *discriminant functions* $g_i(x)$, $i = 1, ..., c$.

The classifier is said to assign a feature vector x to class ω_i if:

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \qquad \text{for all } j \neq i.$$

Hence, the classifier can be viewed as a network or machine that computes c discriminant functions and selects the category corresponding to the largest discriminant.

Classifiers, Discriminant Functions and Decision Surfaces

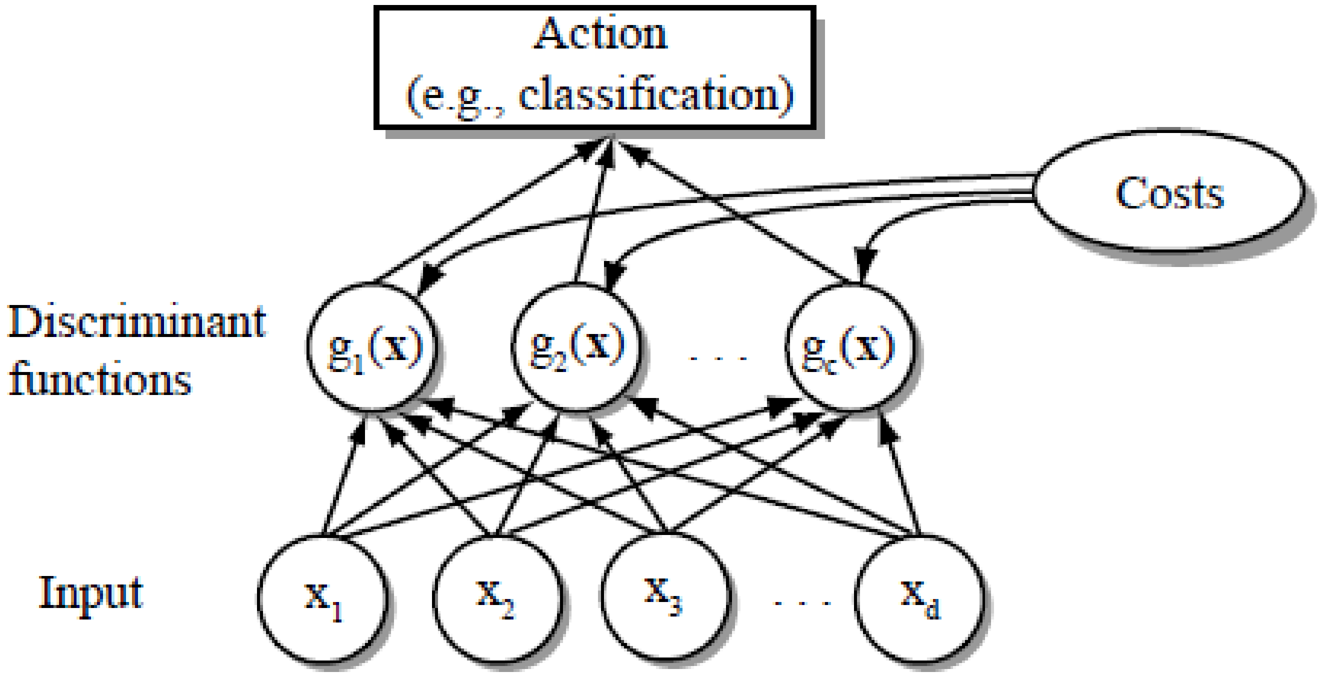

Figure: The functional structure of a general statistical pattern classifier which includes *d* inputs and *c* discriminant functions *gi*(**x**). A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly.

Classifiers, Discriminant Functions and Decision Surfaces

- A Bayes classifier is easily and naturally represented in this way. For the general case with risks, we can let $g_i(x) = -R(\alpha_i|x)$, since the maximum discriminant function will then correspond to the minimum conditional risk.

- For the minimum error-rate case, we can simplify things further by taking $g_i(x) = P(\omega_i|x)$, so that the maximum discriminant function corresponds to the maximum posterior probability.

Classifiers, Discriminant Functions and Decision Surfaces

- In particular, for minimum-error-rate classification, any of the following choices gives identical classification results, but some can be much simpler to understand or to compute than others:

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j)} \tag{25}$$

$$g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i) \tag{26}$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i), \tag{27}$$

Where ln denotes natural logarithm.

Even though the discriminant functions can be written in a variety of forms, the decision rules are equivalent.

Decision Boundary

- The effect of any decision rule is to divide the feature space into *c decision regions*, *R*1,...,*R* decision *c*.
- If $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all $j = i$, then $\mathbf{x}$ is in region R_i, and the decision rule calls for us to assign $\mathbf{x}$ to ω_i.
- The regions are separated by decision boundaries, surfaces in feature space where ties occur among the largest discriminant functions.

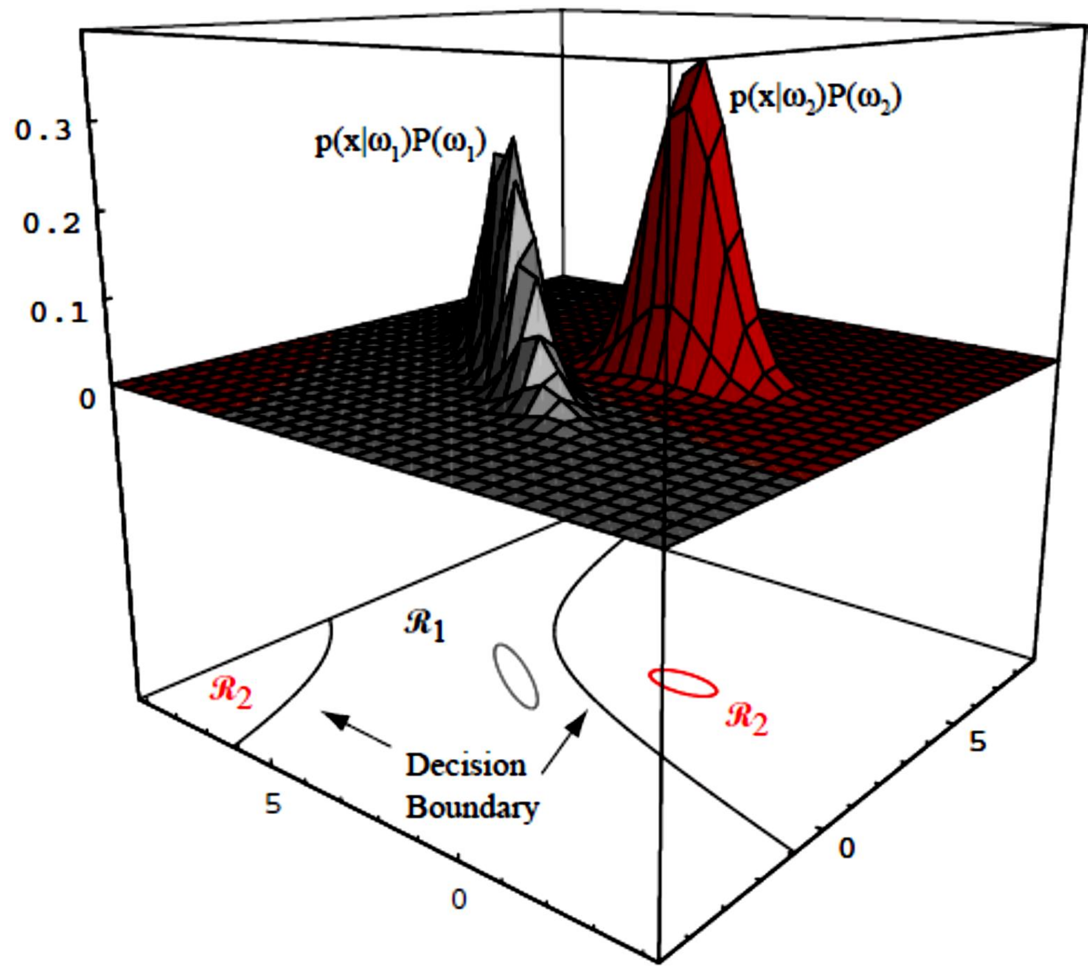

Figure 2.6: In this two-dimensional two-category classifier, the probability densities are Gaussian (with 1/*e* ellipses shown), the decision boundary consists of two hyperbolas, and thus the decision region *R*2 is not simply connected.

Classifiers, Discriminant Functions and Decision Surfaces

1. The two-Category Case:

- The two-category case is just a special instance of the multicategory case.
- Indeed, a classifier that places a pattern in one of only two categories has a special name — ***a dichotomizer.***
- Instead of using two discriminant functions g_1 and g_2 and assigning x to ω_1 if $g_1 > g_2$, it is more common to define a single discriminant function,

$$g(\mathbf{x}) \equiv g_1(\mathbf{x}) - g_2(\mathbf{x}),$$

Decision rule: Decide ω_1 if $g(x) > 0$; otherwise decide ω_2.

Classifiers, Discriminant Functions and Decision Surfaces

- Thus, a dichotomizer can be viewed as a machine that computes a single discriminant function g(x), and classifies x according to the algebraic sign of the result..
- minimum-error-rate discriminant function can be written as:

$$g(\mathbf{x}) = P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x}) \tag{29}$$

$$g(\mathbf{x}) = \ln\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \ln\frac{P(\omega_1)}{P(\omega_2)}. \tag{30}$$

- A classifier for more than two categories is called a ***polychotomizer***.

The Normal Density

- How the decision of a Bayes classifier is determined?

- Expectation (*E*): *Expected value* of a scalar function *f*(*x*), defined expectation for some density *p*(*x*):

$$\mathcal{E}[f(x)] \equiv \int_{-\infty}^{\infty} f(x)p(x)dx.$$

- If we have samples in a set *D* from a discrete distribution, we must sum over all samples as

$$\mathcal{E}[f(x)] = \sum_{x \in \mathcal{D}} f(x)P(x),$$

Where P(x) is the probability mass at x.

For recap:
H. Pishro-Nik, "Introduction to probability, statistics, and random processes", available at https://www.probabilitycourse.com, Kappa Research LLC, 2014.

The Normal Density

1. Univariate Density

- We begin with the continuous univariate normal or Gaussian density,

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

- The *expected value* of *x* (an average, here taken over the feature space) is,

$$\mu \equiv \mathcal{E}[x] = \int_{-\infty}^{\infty} x p(x)\, dx,$$

and where the expected squared deviation or variance is

$$\sigma^2 \equiv \mathcal{E}[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, dx.$$

- *The univariate normal density is completely specified by two parameters: its mean mean μ and variance $\sigma^{2.}$*
- p(x) ~N(μ, σ^2) to say that x is distributed normally with mean μ and variance σ^2

The Normal Density

Univariate Density:

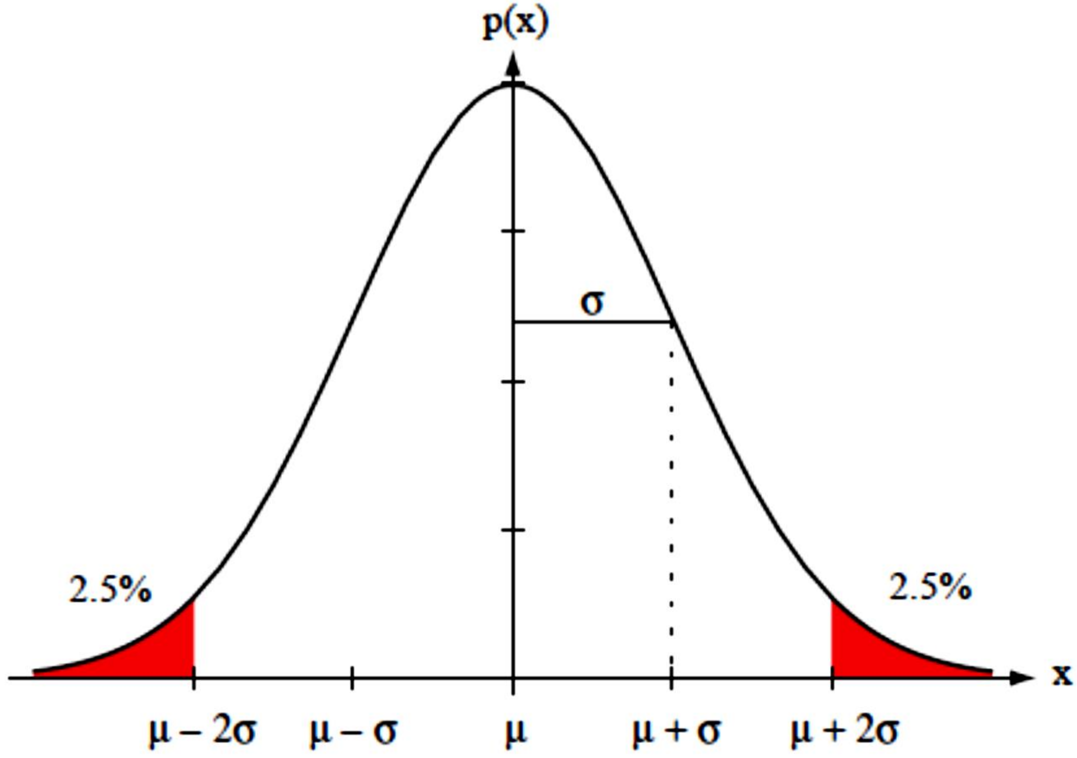

Figure 2.7: A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$.

Entropy and Central Limit Theorem

- ***Entropy:*** The entropy is a non-negative quantity that describes the fundamental uncertainty in the values of points selected randomly from a distribution.
- Entropy of a distribution is given by,

$$H(p(x)) = -\int p(x)\,\ln p(x)\,dx,$$

- Entropy is measured in nats.
- It can be shown that the normal distribution has the maximum entropy of all distributions having a given mean and variance.
- ***Central Limit Theorem:*** The aggregate effect of a large number of small, independent random disturbances will lead to a Gaussian distribution.

Entropy and Central Limit Theorem

2. Multivariate Density:

- Multivariate normal distribution or multivariate Gaussian distribution, is a generalization of the one-dimensional (univariate) normal distribution to higher dimensions.
- The general multivariate normal density in d dimensions is written as,

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right], \qquad (37)$$

where x is a d-component column vector, μ is the d-component mean vector, Σ is the d-by-d covariance matrix, |Σ| and Σ^{-1} are its determinant and inverse, respectively, covariance and $(x - \mu)^t$ is the transpose of x − μ.

- *The eqn. 37 can represented as p*(x) $\sim N(\mu,\Sigma)$.

Entropy and Central Limit Theorem

2. Multivariate Density:

- The *expected value* of *x* (an average, here taken over the feature space) is,

$$\boldsymbol{\mu} \equiv \mathcal{E}[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x})\, d\mathbf{x}$$

variance is,

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

where the expected value of a vector or a matrix is found by taking the expected values of its components.

- *In other words, if x^i is the i^{th} component of x, μ_i the i^{th} component of μ, and σ_{ij} the ij^{th} component of Σ, then*

$$\mu_i = \mathcal{E}[x_i]$$

$$\sigma_{ij} = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)].$$

Entropy and Central Limit Theorem

2. Multivariate Density:

- The *expected value* of *x* (an average, here taken over the feature space) is,

$$\boldsymbol{\mu} \equiv \mathcal{E}[\mathbf{x}] = \int \mathbf{x}p(\mathbf{x})\, d\mathbf{x}$$

variance is,

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

where the expected value of a vector or a matrix is found by taking the expected values of its components.

- *In other words, if x^i is the i^{th} component of x, μ_i the i^{th} component of μ, and σ_{ij} the ij^{th} component of Σ, then*

$$\mu_i = \mathcal{E}[x_i]$$

$$\sigma_{ij} = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)].$$

Discriminant Functions for the Normal Density

We saw that the minimum-error-rate classification can be achieved by use of the discriminant functions

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i).$$

This expression can be readily evaluated if the densities $p(x|\omega_i)$ are multivariate normal, i.e., if $p(x|\omega_i) \sim N(\mu i,\Sigma_i)$.

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i).$$