Bayes Decision Theory

Minimum-Error-Rate Classification

Classifiers, Discriminant Functions and Decision Surfaces

The Normal Density

Minimum-Error-Rate Classification

Actions are decisions on classes

If action α_i is taken and the true state of nature is ω_j then: decision is correct if i = j and in error if $i \neq j$

Seek a decision rule that minimizes the probability of error which is the error rate

Minimum Error Rate Classifier Derivation

zero-one loss function:

$$\lambda(\alpha_i, \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \quad i, j = 1, ..., c$$

Therefore, the conditional risk is:

$$R(\alpha_i/x) = \sum_{j=1}^{j=c} \lambda(\alpha_i/\omega_j) P(\omega_j/x)$$
$$= \sum_{j\neq 1} P(\omega_j/x) = 1 - P(\omega_i/x)$$

The risk corresponding to this loss function is the average probability error"

• Minimize the risk requires maximize $P(\omega_i / x)$

(since
$$R(\alpha_i / x) = 1 - P(\omega_i / x)$$
)
For Minimum error rate
Decide ω_i if $P(\omega_i / x) > P(\omega_i / x) \ \forall j \neq i$

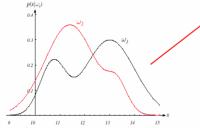
Likelihood Ratio Classification

Regions of decision and zero-one loss function,

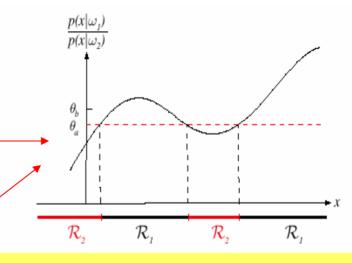
Let
$$\frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)} = \theta_{\lambda}$$
 then decide ω_1 if : $\frac{P(x/\omega_1)}{P(x/\omega_2)} > \theta_{\lambda}$

• If λ is the zero-one loss

$$\lambda = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$
then $\theta_{\lambda} = \frac{P(\omega_{2})}{P(\omega_{1})} = \theta_{a}$
if $\lambda = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}$ then $\theta_{\lambda} = \frac{2P(\omega_{2})}{P(\omega_{1})} = \theta_{b}$



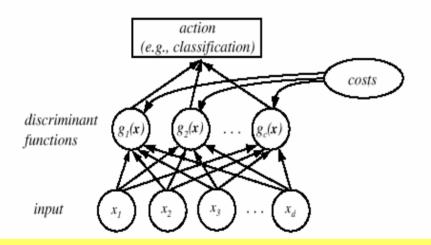
Class-conditional pdfs



- Likelihood Ratio $p(x/\omega_1)/p(x/\omega_2)$.
- If we use a zero-one loss function decision boundaries are determined by threshold θ_a .
- If loss function penalizes miscategorizing ω_2 as ω_1 more than converse we get larger threshold θ_b and hence \textbf{R}_1 becomes smaller

Classifiers, Discriminant Functions and Decision Surfaces

• Many methods of representing pattern classifiers Set of discriminant functions $g_i(x)$, i = 1,..., cClassifier assigns feature x to class ω_i if $g_i(x) > g_i(x) \ \forall j \neq i$



Classifier is a machine that computes c discriminant functions

Functional structure of a general statistical pattern Classifier with d inputs and c discriminant functions $g_i(x)$

Forms of Discriminant Functions

- Let $g_i(x) = -R(\alpha_i / x)$ (max. discriminant corresponds to min. risk!)
- For the minimum error rate, we take

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

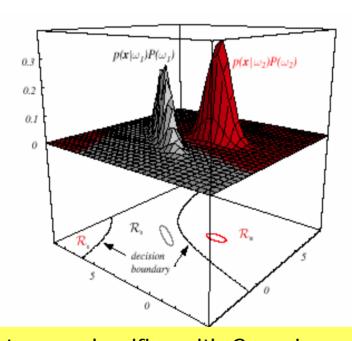
(max. discrimination corresponds to max. posterior!)
 $g_i(x) \equiv P(x \mid \omega_i) P(\omega_i)$
 $g_i(x) = \ln P(x \mid \omega_i) + \ln P(\omega_i)$

CSE 555: Srihari

5

Decision Region

• Feature space divided into c decision regions if $g_i(x) > g_i(x) \ \forall j \neq i \ then \ x \ is \ in \ R_i$



2-D, two-category classifier with Gaussian pdfs
Decision Boundary = two hyperbolas
Hence decision region R2 is not simply connected
Ellipses mark where density is 1/e times that of peak distribution

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

The computation of g(x)

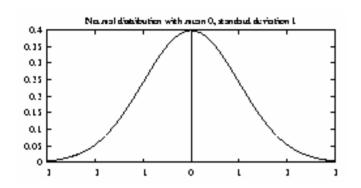
$$g(x) = P(\omega_1/x) - P(\omega_2/x)$$

$$= \ln \frac{P(x/\omega_1)}{P(x/\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

7

The Normal Distribution

A bell-shaped distribution defined by the probability density function



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

If the random variable X follows a normal distribution, then

- The probability that X will fall into the interval (a,b) is given by $\int_a^b p(x)dx$
- Expected, or mean, value of X is $E[X] = \int_{-\infty}^{\infty} x p(x) dx = \mu$
- Variance of X is $Var(x) = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx = \sigma^2$
- ullet Standard deviation of X, $oldsymbol{\sigma}^2$, is

$$\sigma_{x} = \sigma$$

Relationship between Entropy and Normal Density

Entropy of a distribution

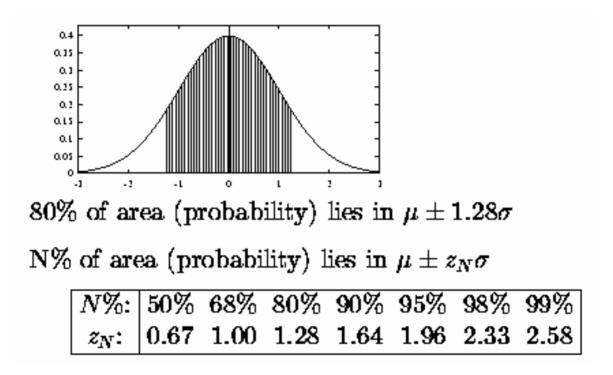
$$H(p(x)) = \int_{-\infty}^{\infty} p(x) \ln p(x) dx$$

Measured in nats. If log₂ is uses the unit is bits

Entropy measures uncertainty in the values of points selected randomly from a distribution

Normal distribution has maximum entropy over all distributions having a given mean and variance

Normal Distribution, Mean 0, Standard Deviation 1



With 80% confidence the r.v. will lie in the two-sided interval[-1.28,1.28]

The Normal Density in Pattern Recognition

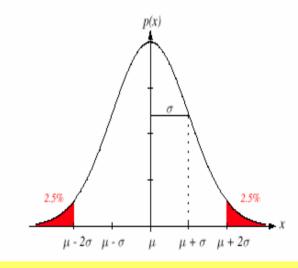
Univariate density

- Analytically tractable, continuous
- A lot of processes are asymptotically Gaussian
- Central Limit Theorem: aggregate effect of a sum of a large number of small, independent random disturbances will lead to a Gaussian distribution
- Handwritten characters, speech sounds are ideal or prototype corrupted by random process

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

Where:

 μ = mean (or expected value) of x σ^2 = expected squared deviation
or variance



Univariate normal distribution has roughly 95% of its area in the range $|x-\mu| < 2\sigma$.

The peak of the distribution has value $p(\mu)=1/sqrt(2\pi\sigma)$

Multivariate density

Multivariate normal density in d dimensions is:

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

$$abbreviated \quad as$$

$$p(x) \sim N(\mu, \Sigma)$$

where:

 $x = (x_1, x_2, ..., x_d)^t$ (t stands for the transpose vector form) $\mu = (\mu_1, \mu_2, ..., \mu_d)^t$ mean vector $\Sigma = d^*d$ covariance matrix $\Sigma = d^*d$ are determinant and inverse respectively

Mean and Covariance Matrix

Formal Definitions

$$\mu = E[x] = \int xp(x)dx$$

$$\sum = E[(x - \mu)(x - \mu)^t] = \int (x - \mu)(x - \mu)^t p(x)dx$$

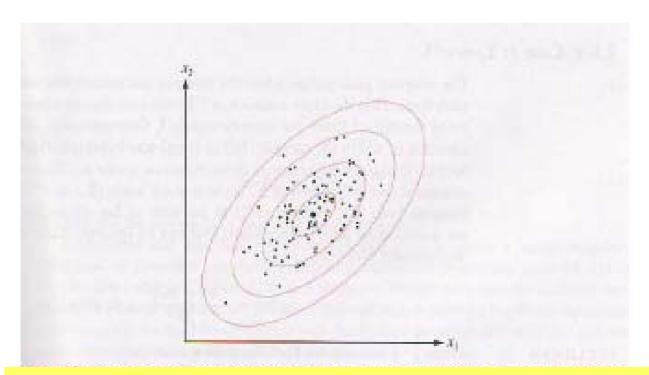
Mean vector has its components which are means of variables

Covariance:

Diagonal elements are variances of variables Cross-diagonal elements are covariances of pairs of variables Statistical independence means off-diagonal elements are zero

Multivariate Normal Density

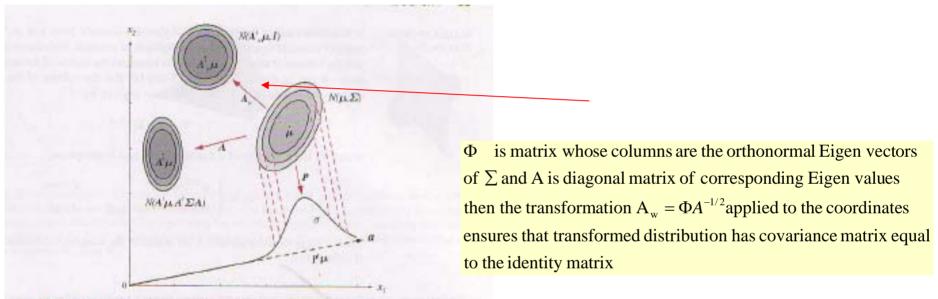
 Specified by d+d(d+1)/2 parameters: mean and independent elements of covariance matrix



Locii of points of constant density are hyperellipsoids

Samples drawn from a 2-D Gaussian lie in a cloud centered at the mean μ . Ellipses show lines of equal probability density of the Gaussian

Linear Combinations of Normally distributed variables are normally distributed



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(A^t μ ,A^t Σ A). Another linear transformation- a projection P onto a Line defined by vector a– leads to N(μ , σ^2) measured along that line. While the transforms yield distributions in a different space they are shown superimposed on the original x₁-x₂ space. A whitening transform A_w leads to a circularly symmetric Gaussian, here shown displaced.

CSE 555: Srihari

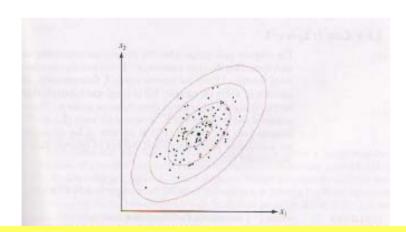
15

Mahanalobis Distance

$$r^2 = (x - \mu)^t \Sigma^{-1} (x - \mu)$$

is the Mahanalobis distance from x to μ

Contours of constant
Density are hyperellipsoids
of constant Mahanalobis
Distance



For a given dimensionality Scatter of samples varies directly with |E|1/2

Samples drawn from a 2-D Gaussian lie in a cloud centered at the mean μ . Ellipses show lines of equal probability density of the Gaussian