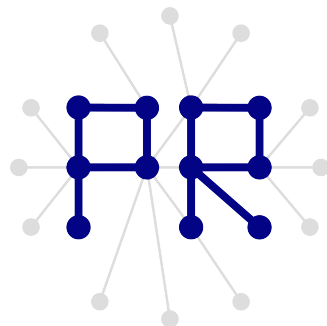
Pattern Recognition Lecture
"Feature Selection"

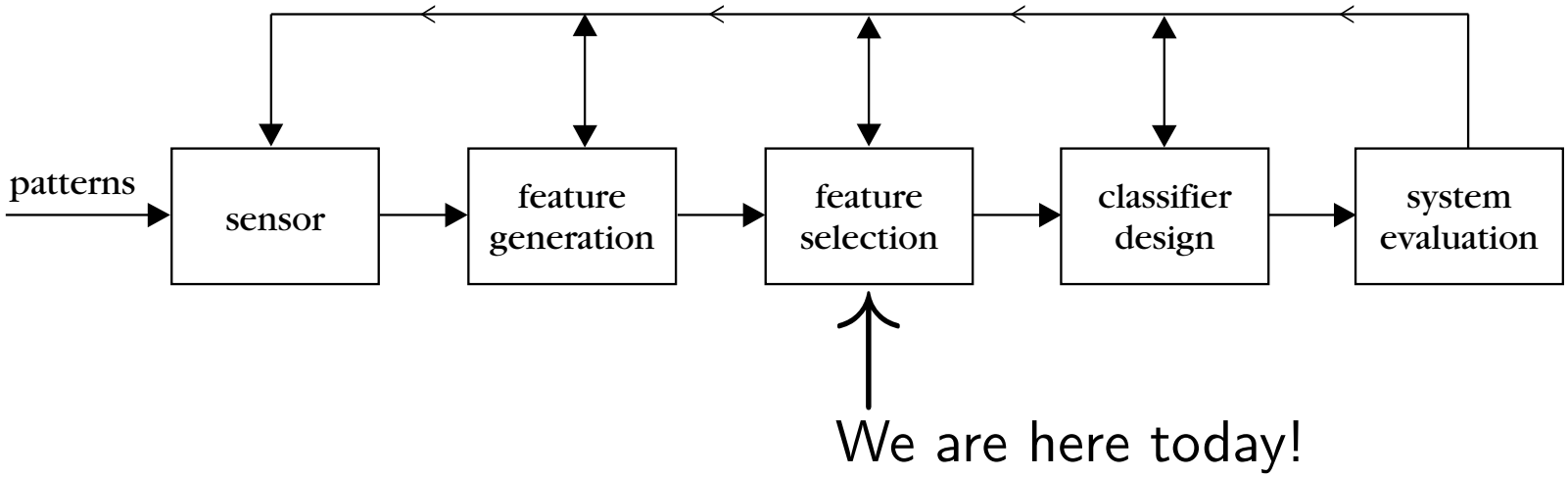
Prof. Dr. Marcin Grzegorzek

Research Group for Pattern Recognition
www.pr.informatik.uni-siegen.de

Institute for Vision and Graphics
University of Siegen, Germany



Pattern Recognition Chain



Overview

1 Introduction

2 Preprocessing

3 The Peaking Phenomenon

4 Feature Selection Based on Statistical Hypothesis Testing

5 The Receiver Operating Characteristics (ROC) Curve

Overview

Problem Statement

Feature Selection/Reduction

- Given a number of features, how can one select the most important of them so as to reduce their number and at the same time retain as much as possible of their class discriminatory information.

- An important measure is here the ratio N/l. Ratios as high as 10 to 20 are in some cases considered necessary.

Overview

Outlier Removal

- An outlier is a point that lies very far from the mean of the corresponding random variable.
- This distance is measured with respect to a given threshold.
- Outliers produce large errors during training.
- When the number of outliers is small, they are discarded.
- When the number of outliers is high, cost functions are applied.

Data Normalisation - a Linear Method (1)

- Features with large values may have a larger influence in the cost function.
- The problem is overcome by normalising the features so that their values lie within similar ranges.
- One of the techniques here is the normalisation via the respective estimates of the mean and variance.

Data Normalisation - a Linear Method (2)

- For N available data of the feature no. k we have

$$\overline{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik}$$

$$\sigma_k^2 = \frac{1}{N}\sum_{i=1}^{N}(x_{ik} - \overline{x}_k)^2$$

$$\widehat{x}_{ik} = \frac{x_{ik} - \overline{x}_k}{\sigma_k}$$

- The resulting normalised features will now have zero mean and unit variance.

Data Normalisation - a Nonlinear Method

- Nonlinear methods are applied, if the data is not evenly distributed around the mean.
- One of the possibilities is here the softmax scaling

$$y = \frac{x_{ik} - \overline{x}_k}{r\sigma_k} \qquad \widehat{x}_{ik} = \frac{1}{1 + \exp(-y)}$$

- This is a squashing function limiting data to the range $[0, 1]$.
- The factor r is user defined.

Missing Data

- In practice, certain features may be missing from some feature vectors (e. g., in social sciences due to partial response in surveys).
- The traditional techniques complete the missing data by
  - zeros,
  - the unconditional mean computed from the available values, or
  - the conditional mean, if one has an estimate of the density function
- Another approach is to discard feature vectors with missing values.

Imputing from a Conditional Distribution (1)

- Another method to deal with missing data is called imputing from a conditional distribution.
- The idea here is to complete the data by respecting the statistical nature of the missing values.
- Let us denote complete feature vectors by $\mathbf{x}_{\text{com}}$, observed by $\mathbf{x}_{\text{obs}}$, and missing by $\mathbf{x}_{\text{mis}}$

$$\mathbf{x}_{\text{com}} = \begin{bmatrix} \mathbf{x}_{\text{obs}} \\ \mathbf{x}_{\text{mis}} \end{bmatrix}$$

Imputing from a Conditional Distribution (2)

- Under the assumption that the probability of missing a value does not depend on the value itself, we have

$$p(\mathbf{x}_{\text{mis}}|\mathbf{x}_{\text{obs}};\boldsymbol{\theta}) = \frac{p(\mathbf{x}_{\text{obs}},\mathbf{x}_{\text{mis}};\boldsymbol{\theta})}{p(\mathbf{x}_{\text{obs}};\boldsymbol{\theta})}$$

  where

$$p(\mathbf{x}_{\text{obs}};\boldsymbol{\theta}) = \int p(\mathbf{x}_{\text{com}};\boldsymbol{\theta})d\mathbf{x}_{\text{mis}}$$

- $\boldsymbol{\theta}$ is an unknown set of parameters.

- In practice, an estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ must first be obtained from $\mathbf{x}_{\text{obs}}$.

Multiple Imputation Procedure (MI)

- Imputing from a conditional distribution is referred as single imputation (SI)
- In MI for each missing value, $m > 1$ samples are generated.
- The results are then combined so that certain statistical properties are fulfilled.
- Instead of drawing a single point from $p(\mathbf{x}_{\text{mis}}|\mathbf{x}_{\text{obs}}; \widehat{\boldsymbol{\theta}})$ one can use different parameters $\widehat{\boldsymbol{\theta}}_{i=1,\ldots,m}$ and draw the m samples from

$$p(\mathbf{x}_{\text{mis}}|\mathbf{x}_{\text{obs}}; \widehat{\boldsymbol{\theta}}_i) \qquad i = 1, 2, \ldots, m$$

Overview

The Peaking Phenomenon - Introduction

- In order to design a classifier with good generalisation performance, the number of training points N must be large enough with respect to the number of features l.
- If a linear classifier $\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$ is supposed to be designed, the number of data points must be larger than $l + 1$.
- The larger the N, the better the estimate.

The Peaking Phenomenon - Example (1)

- In this example, the interplay between N and l will be shown.
- Consider a two-class problem with equal a priori probabilities $P(\omega_1) = P(\omega_2) = \frac{1}{2}$.
- Both classes are represented by Gaussian distributions of the same covariance matrix $\boldsymbol{\Sigma} = \mathbf{I}$ and mean vectors $\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$ respectively, where

$$\boldsymbol{\mu} = \left[1, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{3}}, \ldots, \frac{1}{\sqrt{l}}\right]^{\mathrm{T}}$$

The Peaking Phenomenon - Example (2)

- Since the features are jointly Gaussian and $\boldsymbol{\Sigma} = \mathbf{I}$, the involved features are statistically independent.
- Moreover, the optimal Bayesian rule is equivalent to the minimum Euclidean distance classifier.
- An unknown feature vector $\mathbf{x}$ is classified to, say, ω_1 if

$$\|\mathbf{x} - \boldsymbol{\mu}\|^2 < \|\mathbf{x} + \boldsymbol{\mu}\|^2$$

$$\Downarrow$$

$$z \equiv \mathbf{x}^{\mathrm{T}}\boldsymbol{\mu} > 0$$

- If $z < 0$, we decide in favour of the class ω_2.

The Peaking Phenomenon - Example (3)

Case 1: Known Mean Value Vector μ

- The inner product z, being a linear combination of independent Gaussian variables, is also a Gaussian variable with

$$\|\boldsymbol{\mu}\|^2 = \sum_{i=1}^{l} \frac{1}{i} \qquad \text{and} \qquad \sigma_z^2 = \|\boldsymbol{\mu}\|^2$$

- The probability of committing an error turns out to be

$$P_e = \int_{b_l}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz \quad \text{where} \quad b_l = \sqrt{\sum_{i=1}^{l} \frac{1}{i}}$$

- Note that: $l \to \infty \Rightarrow b_l \to \infty \Rightarrow P_e \to 0$.

The Peaking Phenomenon - Example (3)

Case 2: Unknown Mean Value Vector $\boldsymbol{\mu}$

- The mean value vector $\boldsymbol{\mu}$ has to be estimated from the training data set, e. g., with the maximum likelihood estimation

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{k=1}^{N} s_k \mathbf{x}_k$$

  where $s_k = 1$ if $\mathbf{x}_k \in \omega_1$ and $s_k = -1$ if $\mathbf{x}_k \in \omega_2$.
- Decisions are taken depending on $z = \mathbf{x}^{\mathrm{T}} \widehat{\boldsymbol{\mu}}$.
- z can be considered as approximately Gaussian for large l

$$E[z] = \sum_{i=1}^{l} \frac{1}{i} \quad \text{and} \quad \sigma_z^2 = \left(1 + \frac{1}{N}\right) \sum_{i=1}^{l} \frac{1}{i} + \frac{l}{N} \quad .$$

The Peaking Phenomenon - Example (4)

Case 2: Unknown Mean Value Vector μ

- The probability error is then given by

$$P_e = \int_{b_l}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz = \frac{E[z]}{\sigma_z}$$

- It can be shown that $l \to \infty \Rightarrow b_l \to 0 \Rightarrow P_e \to \frac{1}{2}$.

The Peaking Phenomenon - Example (5)

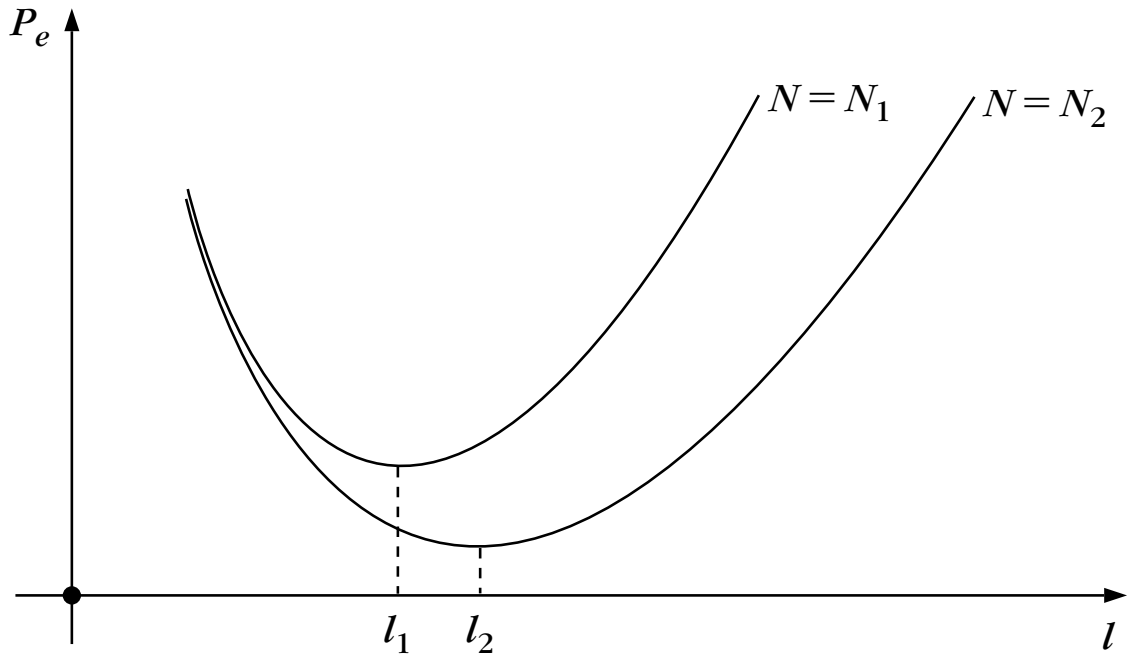
Conclusions

- If for any l the corresponding pdf is known, then we can perfectly discriminate the two classes by arbitrarily increasing the number of features.

- If the pdfs are not known and the associated parameters must be estimated using a finite training set, then the arbitrary increase of the number of features leads to the maximum possible value of the error rate $P_e = 0.5$.

The Peaking Phenomenon - Practice (1)

- Increasing the number of features for a finite N will initially bring an improvement in performance, but after a critical value the probability of error gets higher again.



- Which value is greater, N_1 or N_2?

The Peaking Phenomenon - Practice (2)

- The peaking phenomenon occurs for $l = l_1$ ($l = l_2$) for N_1 (N_2) training examples.
- The minimum in the curves occurs at some number $l = \frac{N}{\alpha}$, where α, usually, takes values in the range 2 and 10.
- Therefore, for small number of training data, a small number of features must be used.

Overview

Introduction

- Here, we look at each of the generated features independently and test its discriminatory capability for the problem at hand.
- Let x be the random variable representing a specific feature (element of a feature vector). We will investigate whether the values it takes for the different classes, say ω_1 and ω_2, differ significantly.
- The problem will be formulated in the context of statistical hypothesis testing.
- The hypotheses are:
  H_1: The values of the feature differ significantly.
  H_0: The values of the feature do not differ significantly.

Hypothesis Testing Basics (1)

- Let x be a random variable with pdf assumed to be known within an unknown parameter θ (e. g., μ or σ for the Gaussian pdf).

- The following hypothesis test is supposed to be performed:

$$H_1 : \theta \neq \theta_0 \qquad H_0 : \theta = \theta_0$$

- Let $x_{i=1,\ldots,N}$ be the experimental samples of the random variable x, and $q = f(x_1, \ldots, x_N)$ a problem-specific function.

- The function is selected so that the pdf of q can be easily parametrised in terms of the unknown θ: $p_q(q; \theta)$.

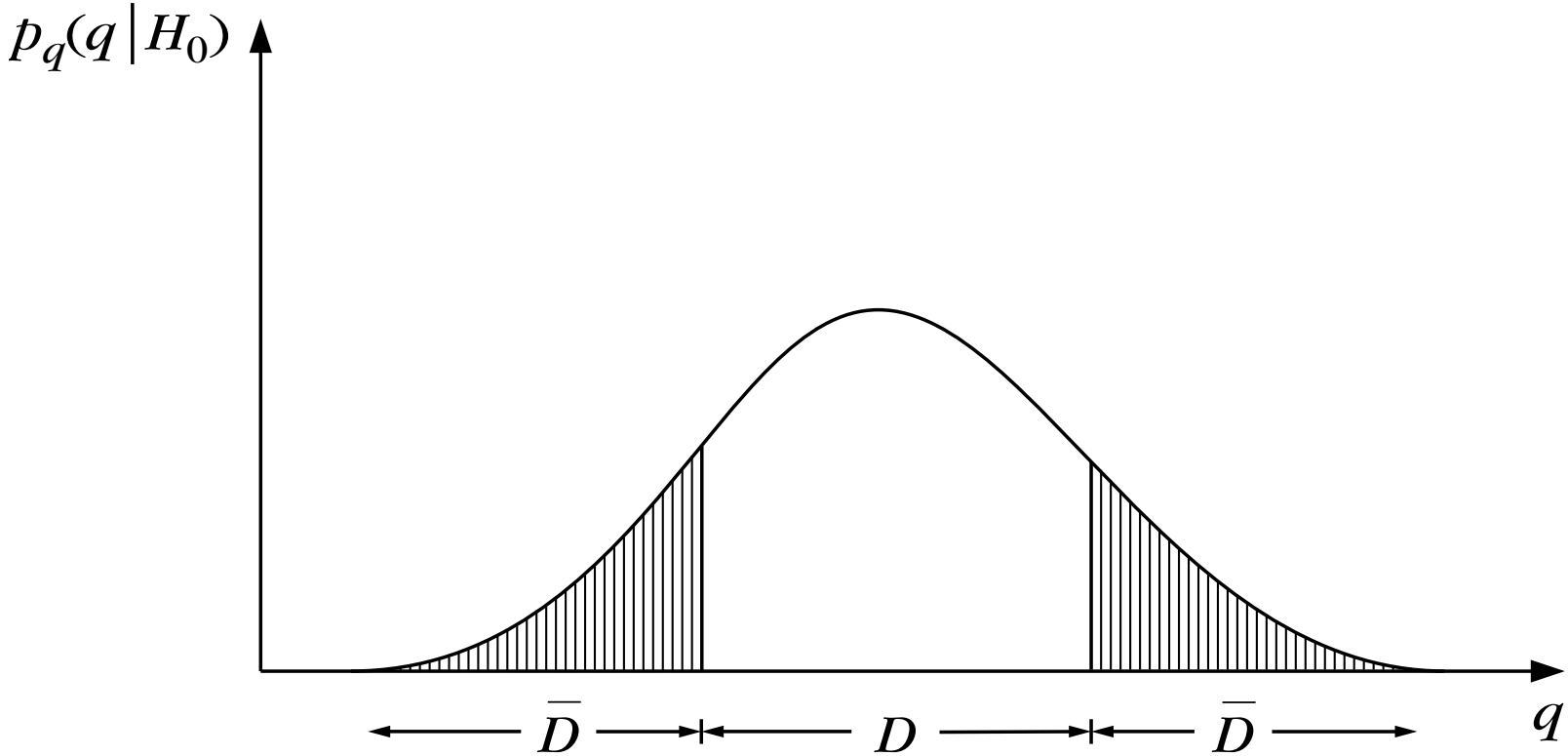
Hypothesis Testing Basics (2)

- Let D be the interval of q in which it has a high probability of lying under the hypothesis H_0, let $\overline{D}$ be its complement.
- D is known as the acceptance interval, and $\overline{D}$ is the critical interval, variable q is known as test statistic.
- The questions now refers to the probability of reaching a wrong decision. Let H_0 be true:

$$P(q \in \overline{D}|H_0) \equiv \rho$$

- This probability is the integral presented graphically on the next slide.

Hypothesis Testing Basics (3)



$$P(q \in \overline{D}|H_0) \equiv \rho$$

Hypothesis Testing - The Known Variance Case (1)

- Let denote $E[x] = \mu$ and $E[(x - \mu)^2] = \sigma^2$
- A popular estimate of μ based on known samples is

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- Using a different set of N samples, a different estimate will result. So, $\overline{x}$ is also a random variable and can be described by a pdf $p_{\overline{x}}(\overline{x})$.

Hypothesis Testing - The Known Variance Case (2)

- The mean of the random variable $\overline{x}$ is

$$E[\overline{x}] = E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N} E[x_i] = \mu$$

- The variance of the random variable $\overline{x}$ is

$$\sigma_{\overline{x}}^2 = E[(\overline{x}-\mu)^2] = E\left[\left(\frac{1}{N}\sum_{i=1}^{N} x_i - \mu\right)^2\right]$$

$$\Downarrow$$

$$\sigma_{\overline{x}}^2 = \frac{1}{N^2}\sum_{i=1}^{N} E[(x_i-\mu)^2] + \frac{1}{N^2}\sum_{i}\sum_{j\neq i} E[(x_i-\mu)(x_j-\mu)]$$

Hypothesis Testing - The Known Variance Case (3)

- The statistical independence of the samples dictates

$$E[(x_i - \mu)(x_j - \mu)] = E[x_i - \mu]E[x_j - \mu] = 0$$

- So, the variance of the random variable $\overline{x}$ from the last slide can be written as

$$\sigma_{\overline{x}}^2 = \frac{1}{N^2} \sum_{i=1}^{N} E[(x_i - \mu)^2] = \frac{1}{N}\sigma^2$$

- In words, the larger the number of measurement samples, the smaller the variance of $\overline{x}$ around the true mean μ.

Hypothesis Testing - The Known Variance Case (4)

- Let us assume that we are given a the value $\widehat{\mu}$ and we have to decide upon

$$H_1 : E[x] \neq \widehat{\mu} \qquad H_0 : E[x] = \widehat{\mu}$$

- We define the test statistic as

$$q = f(x_1, x_2, \ldots, x_n) = \frac{\overline{x} - \widehat{\mu}}{\sigma/\sqrt{N}}$$

- Recalling the central limit theorem, the pdf of $\overline{x}$ under H_0 given $\widehat{\mu}$ is the Gaussian $\mathcal{N}(\widehat{\mu}, \frac{\sigma^2}{N})$

$$p_{\overline{x}}(\overline{x}) = \frac{\sqrt{N}}{\sigma\sqrt{2\pi}} \exp\left(-\frac{N(\overline{x} - \widehat{\mu})^2}{\sigma^2}\right)$$

Hypothesis Testing - The Known Variance Case (5)

- Hence, the probability density function of q under H_0 is approximately $\mathcal{N}(0,1)$. **Can you prove it?**

- For a significance level ρ the acceptance interval $D \equiv [-x_\rho, x_\rho]$ is chosen as the interval in which the random variable q lies with probability $1-\rho$. Recall

$$P(q \in \overline{D}|H_0) \equiv \rho$$

- Acceptance intervals $[-x_\rho, x_\rho]$ corresponding to various probabilities for an $\mathcal{N}(0,1)$ normal distribution are

ρ	0.2	0.15	0.1	0.05	0.02	0.01	0.002	0.001
$1-\rho$	0.8	0.85	0.9	0.95	0.98	0.99	0.998	0.999
x_ρ	1.282	1.44	1.645	1.967	2.326	2.576	3.09	3.291

Hypothesis Testing - The Known Variance Case (6)

The decision on the test hypothesis can now be reached by the following steps:

1. Given the N experimental samples of x, compute $\overline{x}$ and then q.
2. Choose the significance level ρ.
3. Compute from the tables for $\mathcal{N}(0,1)$ the acceptance interval $D = [-x_\rho, x_\rho]$ corresponding to probability $1-\rho$.
4. If $q \in D$ decide H_0, if not decide H_1

Hypothesis Testing - The Known Variance Case (6)

The decision on the test hypothesis can now be reached by the following steps:

1. Given the N experimental samples of x, compute $\overline{x}$ and then q.
2. Choose the significance level ρ.
3. Compute from the tables for $\mathcal{N}(0,1)$ the acceptance interval $D = [-x_\rho, x_\rho]$ corresponding to probability $1-\rho$.
4. If $q \in D$ decide H_0, if not decide H_1

Hypothesis Testing for Unknown Variance (1)

- If the variance of x is not known, it must be estimated $\widehat{\sigma}^2$.
- It can be shown that $E[\widehat{\sigma}^2] = \sigma^2$.
- The test statistic is now defined as

$$q = f(x_1, \ldots, x_n) = \frac{\overline{x} - \mu}{\widehat{\sigma}/\sqrt{N}}$$

- However, q is no longer a Gaussian variable. Assuming that x is a Gaussian random variable, then q is described by the so-called t-distribution with $N - 1$ degrees of freedom (see next slide).

Hypothesis Testing for Unknown Variance (2)

Degrees of Freedom	$1-\rho$ 0.9	$1-\rho$ 0.95	$1-\rho$ 0.975	$1-\rho$ 0.99	$1-\rho$ 0.995
10	1.81	2.23	2.63	3.17	3.58
11	1.79	2.20	2.59	3.10	3.50
12	1.78	2.18	2.56	3.05	3.43
13	1.77	2.16	2.53	3.01	3.37
14	1.76	2.15	2.51	2.98	3.33
15	1.75	2.13	2.49	2.95	3.29
16	1.75	2.12	2.47	2.92	3.25
17	1.74	2.11	2.46	2.90	3.22
18	1.73	2.10	2.44	2.88	3.20
19	1.73	2.09	2.43	2.86	3.17
20	1.72	2.09	2.42	2.84	3.15

Interval Values of Various Significance Levels and Degrees of Freedom for a t-Distribution

Feature Selection by Hypothesis Testing (1)

- Now the theory to hypothesis testing will be applied for feature selection in a classification problem.
- We will test the difference $\mu_1 - \mu_2$ between the means of the values taken by a feature in two classes.
- Let $x_{i=1,\ldots,N}$ be the samples values of the feature in class ω_1 with μ_1, and $y_{i=1,\ldots,N}$ in ω_2 with μ_2 respectively.
- Let us assume that the variance in both classes is the same $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Feature Selection by Hypothesis Testing (2)

- To decide about the closeness of the two mean values, we will test for the hypotheses:

$$H_1 : \Delta\mu = \mu_1 - \mu_2 \neq 0 \qquad H_0 : \Delta\mu = \mu_1 - \mu_2 = 0$$

- Assuming statistical independence between the random variables x and y let denote their difference by $z = x - y$.

- Obviously $E[z] = \mu_1 - \mu_2$ and due to the independence assumption $\sigma_z^2 = 2\sigma_2$.

Feature Selection by Hypothesis Testing (3)

- Now, we can compute the sample mean $\overline{z}$ using the sample means $\overline{x}$ and $\overline{y}$

$$\overline{z} = \sum_{i=1}^{N} (x_i - y_i) = \overline{x} - \overline{y}$$

- For the known variance case $\overline{z}$ follows the normal distribution $\mathcal{N}(\mu_1 - \mu_2, \frac{2\sigma^2}{N})$

Feature Selection by Hypothesis Testing (3)

- If the variance is not known, then we choose the test statistic

$$q = \frac{(\overline{x} - \overline{y}) - (\mu_1 - \mu_2)}{s_z \sqrt{\frac{2}{N}}}$$

where

$$s_z^2 = \frac{1}{2N-2} \left(\sum_{i=1}^{N} (x_i - \overline{x})^2 + \sum_{i=1}^{N} (y_i - \overline{y})^2 \right)$$

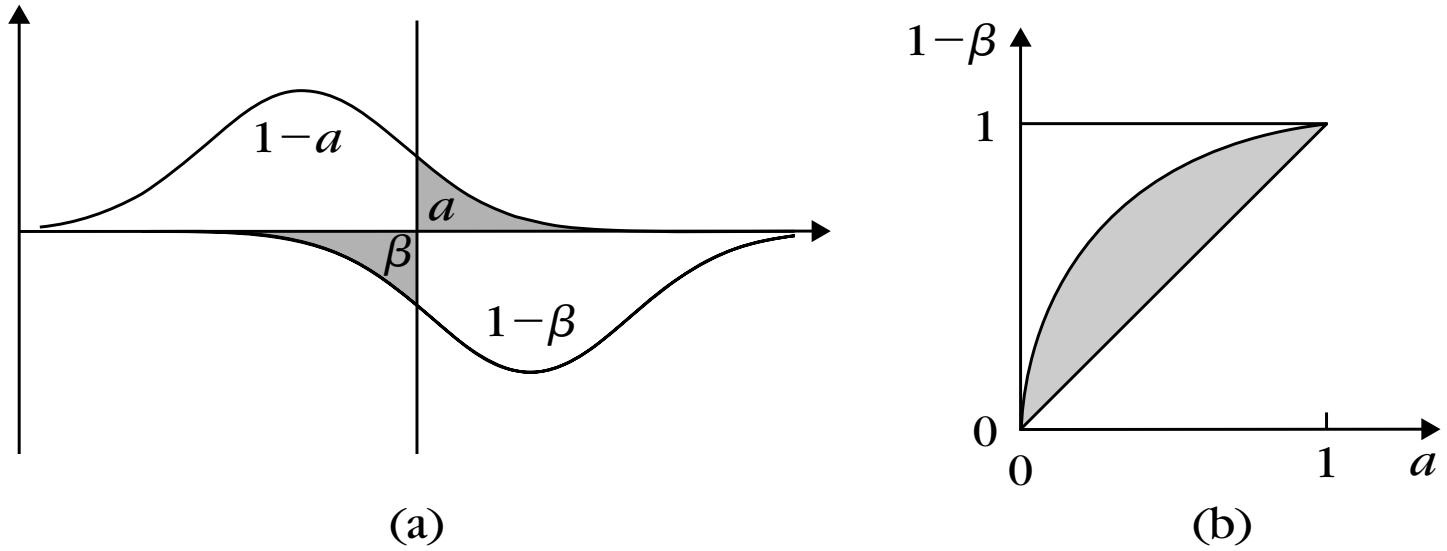
Overview

Introduction

- The hypothesis tests do not work well for some cases.
- The mean values μ_1 and μ_2 of ω_1 and ω_2 may differ significantly, but the spread around the means may be large enough to blur the class distinction.
- We will now focus on techniques providing information about the overlap between the classes.

Overlapping Probability Density Functions (1)

- Here we can see two overlapping pdfs describing the distribution of a feature in two classes, together with a threshold (one pdf has been inverted for illustration purposes)



- We decide class ω_1 for values on the left and ω_2 for values on the right on the threshold.

Overlapping Probability Density Functions (2)

- The probabilities for wrong decisions are denoted by a and β for the classes ω_1 and ω_2 respectively.
- If there is a total overlap, then for any position of the threshold we get $a = 1 - \beta$. This case corresponds to the straight line in Figure (b) on the previous slide.
- As the two pdfs move apart, the corresponding curve departs from the straight line.
- The area on the figure (b) (previous slide) varies between zero, for complete overlap, and $1/2$ for complete separation. This is a measure of the class discrimination capability of the specific feature.