TTT4275 Lecture 1
Spring 2018

Faglærer: Magne Hallstein Johnsen,

Institutt for elektronikk og telekommunikasjon, NTNU

Lecture content

- Basic statistics
- True versus decided class regions and borders
- The theoretical optimal classifier (TOC)
  - Bayes decision rule (BDR) and TOC
  - From TOC to practical design
- Introducing parametric BDR classifiers
- Introducing non-BDR classifiers
  - The discriminant classifier
  - Linear classifiers
  - Nonlinear classifiers
  - Distance/template classifiers

Basic statistics

- Joint and conditional densities/probabilites and Bayes law
  - Definitions
    - Prior probability $P(\omega_i)$
    - A posteriori probability $P(\omega_i/x)$
    - Class-independent density $p(x)$
    - Class-dependent density $p(x/\omega_i)$
  - Bayes law :

    $p(\omega_i, x) = P(\omega_i/x)p(x) = p(x/\omega_i)P(\omega_i) \Rightarrow$

    $P(\omega_i/x) = p(x/\omega_i)P(\omega_i)/p(x)$
  - $p(x) = \sum_i p(\omega_i, x)$
  - $P(\omega_i) = \int_{-\infty}^{\infty} p(\omega_i, x)dx$

True class versus decision regions and borders

- The **true** class borders and regions in the input room are unknown and depend on the chosen features
- If the true regions do not overlap we have a separable problem $\Rightarrow$ zero error rate is theoretically possible
- In practice we always face nonseparable problems (overlapping class regions)!
- The **decision** borders and regions are given by the specific classifier used.
- The decision regions/borders are different from the **true** class regions/borders.
- The decision regions do not overlap (by definition)

The theoretical optimal classifier (TOC)

- TOC is given by the Bayes decision rule (BDR).
  - BDR : $x \in \omega_k \;\Leftrightarrow\; P(\omega_k/x) = \max_i P(\omega_i/x)$
  - Optimal with respect to minimum error rate.

- Using the below Bayes law and the fact that $p(x)$ is class independent.
  - $P(\omega_i/x) = p(x/\omega_i)P(\omega_i)/p(x)$

- We get the following version of the BDR rule :
  - $x \in \omega_k \;\Leftrightarrow\; p(x/\omega_k)P(\omega_k) = \max_i p(x/\omega_i)P(\omega_i)$

From TOC to practical design

- $P(\omega_i), P(\omega_i/x), p(x/\omega_i) \;\; i = 1, C$ are **never** known, in contrast to within detection!

- Thus the optimal BDR classifier does not exist!!

- This opens for two different strategies :
  - Using alternative classifier structures than BDR
  - If a BDR classifier is wanted, we must choose a (mathematically) manageable form for $p(x/\omega_i) \Rightarrow$
  - This "Plug in BDR/MAP" classifier is a **model/approximation** to the true/unknown optimal BDR classifier
  - Independent on strategy and structure; the classifiers have to be **trained** in order to do the correct decisions !!

Introducing parametric form in BDR classifiers

- Gauss is a mathematical attractive form for continuous x :
  - $p(x/\omega_i) = N(\mu_i, \Sigma_i)$
- In most cases a single Gaussian is a too coarse approximation $\Rightarrow$ large error rate
- A Gaussian mixture model (GMM) can approximate most densities :
  - $p(x/\omega_i) = \sum_{k=1}^{K} c_{ik} N(\mu_{ik}, \Sigma_{ik})$
- Other parametric densities (Laplace etc.) are application specific alternatives
- The parameters $\Theta = \{c_{ik}, \mu_{ik}, \Sigma_{ik} \quad i = 1, C \quad k = 1, K\}$ must be estimated by training !!

Introducing the discriminant classifiers : part 1

- The discriminant classifiers are a broad class
- One has to define a discriminant function $g_i(x)$ for each class
- Decision rule : $x \in \omega_k \;\Leftrightarrow\; g_k(x) = \max_i g_i(x)$
- Decision borders : $g_j(x) = g_i(x) \quad j \neq i$
- Linear discriminant classifier
  - Assume vector input x and vector output $g(x)$ ($C \geq 2$ classes).
  - $g = Wx + w_0$

Introducing the discriminant classifiers : part 2

- Examples of **non**linear discriminant classifiers
  - A Gaussian BDR classifier :
    * $g_i(x) = ln(p(x/\omega_i)) + ln(P(\omega_i) \Rightarrow$
    * $g_i(x) = -0.5(x-\mu_i)^T \Sigma_i (x-\mu_i) + C_i$
  - The distance/template based classifier are related to the Gaussian BDR classifier:
    * Assume $\Sigma_j = \Sigma$ and $P(\omega_i) = 1/C$
    * Mahalanobis distance : $MD_i = (x-\mu_i)^T \Sigma (x-\mu_i) \Rightarrow$
    * $g_i(x) = -0.5 MD_i$
  - The MultiLayer Perceptron (MLP)