Neuronale Netze
Classification

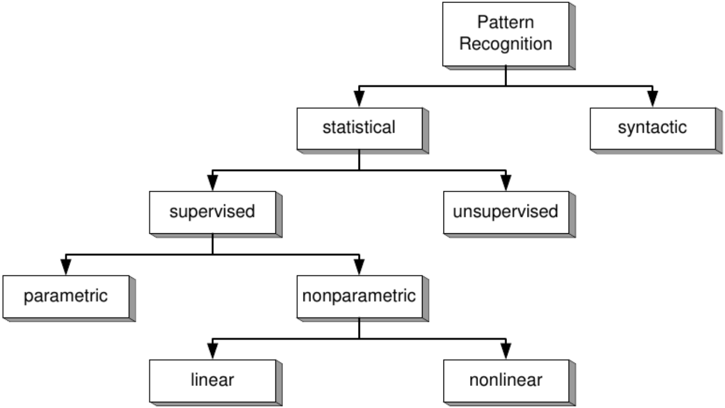
Kevin Kilgour
Institute of Anthropomatics - KIT

October 25, 2011

Pattern Recognition

- Static Patterns, no dependence on Time or Sequential Order
- Important Notions
  - Supervised - Unsupervised Classifiers
  - Parametric - Non-Parametric Classifiers
  - Linear - Non-linear Classifiers
- Classical Methods
  - Bayes Classifier
  - K-Nearest Neighbour
- Connectionist Methods
  - Perceptron
  - Multilayer Perceptrons

Pattern Recognition



Karlsruhe Institute of Technology

Supervised vs Unsupervised

- Supervised training:
  - Class to be recognized is known for each sample in training data.
  - Requires a priori knowledge of useful features and
  - knowledge/labeling of each training token (cost!).
- Unsupervised training:
  - Class is not known and structure is to be discovered automatically.
  - Feature-space-reduction
  - example: clustering, autoassociative nets

Unsupervised Classification

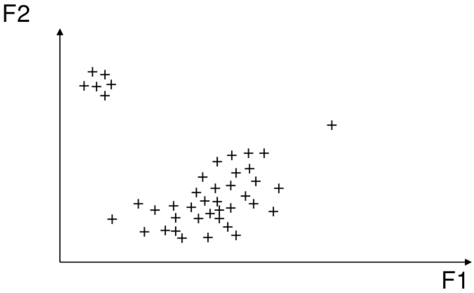


Figure: Classes Unknown: Find Structure

Unsupervised Classification

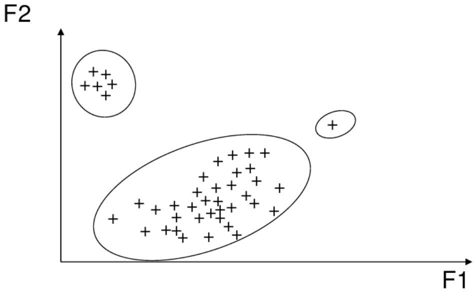
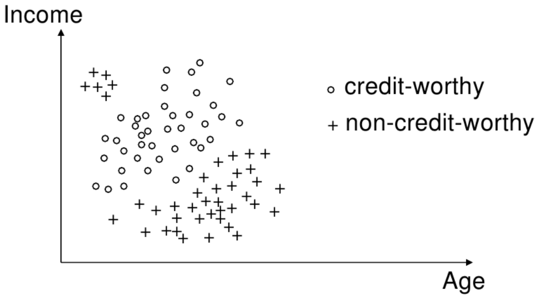


Figure: Classes Unknown: Find Structure.

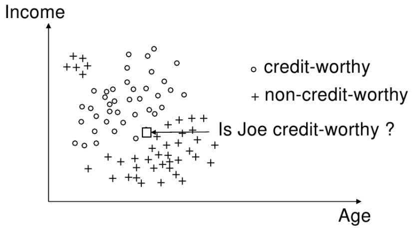
- How? How many?

Supervised Classification



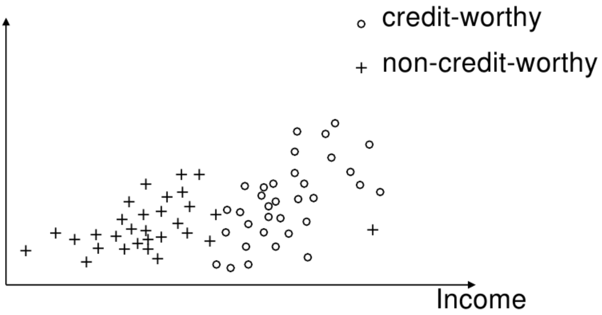
- Classes Known: Creditworthiness: Yes-No
- Features: Income, Age
- Classifiers

Classification Problem

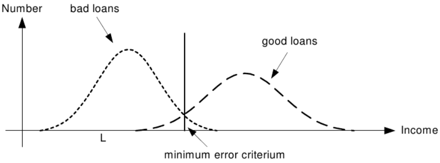


- Features: age, income
- Classes: creditworthy, non-creditworthy
- Problem: Given Joe's income and age, should a loan be made?
- Other Classification Problems: Fraud Detection, Customer Selection...

Classification Problem



Parametric - Non-parametric



- Parametric:
  - assume underlying probability distribution;
  - estimate the parameters of this distribution.
  - Example: "Gaussian Classifier"
- Non-parametric:
  - Don't assume distribution.
  - Estimate probability of error or error citerion directly from training data.
  - Examples: Parzen Window, k-nearest neighbour, perceptron...

Karlsruhe Institute of Technology

Bayes Decision Theory

- Bayes Rule: $P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$
- where: $p(x) = \sum_j p(x|\omega_j)P(\omega_j)$
- A priori probability $P(w_j)$
- A posteriori probability $P(\omega_j|x)$ (after observing x)
- Class-conditional Probability Density $p(x|\omega_j)$

Karlsruhe Institute of Technology

Example

...the use of repeatedly reactive enzyme immunoassay followed by confirmatory Western blot or immunofluorescent assay remains the standard method for diagnosing HIV-1 infection. A large study of HIV testing in 752 U.S. laboratories reported a sensitivity of 99.7% and specificity of 98.5% for enzyme immunoassay...

- $P(aids_{de}) = 0.001$ $\quad P(aids_{trans}) = 0.05$
- $P(aids) = 0.00005$ $\quad P(\neg aids) = 0.0005$
- $P(\oplus|aids) = 0.997$ $\quad P(\ominus|aids) = 0.003$
- $P(\oplus|\neg aids) = 0.015$ $\quad P(\ominus|aids) = 0.985$
- $P(aids|\oplus) = \frac{P(\oplus|aids)P(aids)}{P(\oplus)} = 0.016$
- $P(\oplus) = P(\oplus|aids)P(aids) + P(\oplus|\neg aids)P(\neg aids) = 0.0030$

Maximum a Posteriori

- Often: set of observations O
- Goal: best hypothesis h given O
- assume: best h = most probable h (called h_{MAP})

$$
\begin{aligned}
h_{MAP} &\equiv \operatorname*{argmax}_h P(h|O) \\
&= \operatorname*{argmax}_h \frac{p(O|h)P(h)}{p(O)} \\
&= \operatorname*{argmax}_h p(O|h)P(h)
\end{aligned}
$$

2 Classes Example

- $P(error|x) = \begin{cases} P(\omega_1|x), & \text{if we choose } \omega_1 \\ P(\omega_2|x), & \text{otherwise} \end{cases}$
- Goal: Minimum Error
- choose ω_1 if:
  $P(\omega_2|x) > P(\omega_1|x)$ $\quad p(x|\omega_2)P(\omega_2) > p(x|\omega_1)P(\omega_1)$
- and ω_2 if
  $P(\omega_2|x) < P(\omega_1|x)$ $\quad p(x|\omega_2)P(\omega_2) < p(x|\omega_1)P(\omega_1)$

Example

...the use of repeatedly reactive enzyme immunoassay followed by confirmatory Western blot or immunofluorescent assay remains the standard method for diagnosing HIV-1 infection. A large study of HIV testing in 752 U.S. laboratories reported a sensitivity of 99.7% and specificity of 98.5% for enzyme immunoassay...

- $P(aids_{de}) = 0.001$ $P(aids_{trans}) = 0.05$
- $P(aids) = 0.00005$ $P(\neg aids) = 0.0005$
- $P(\oplus | aids) = 0.997$ $P(\ominus | aids) = 0.003$
- $P(\oplus | \neg aids) = 0.015$ $P(\ominus | aids) = 0.985$
- $P(\oplus | aids) P(aids) = 0.00004985$
- $P(\oplus | \neg aids) P(\neg aids) = 0.00299985$
- $h_{MAP} = \neg aids$

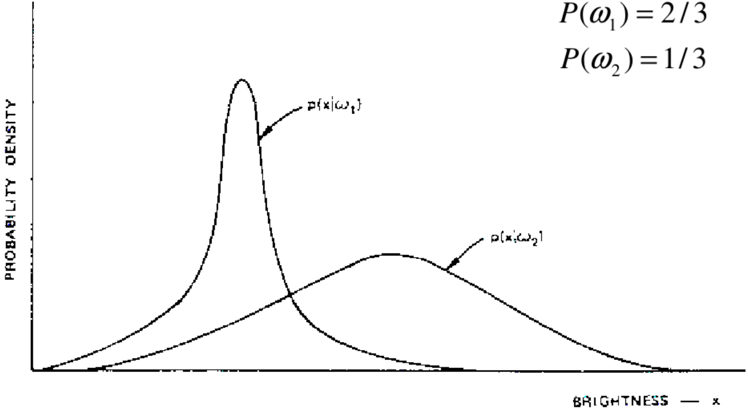
Multiclass Example

- choose ω_i if: $P(\omega_i|x) > P(\omega_j|x) \qquad \forall j$

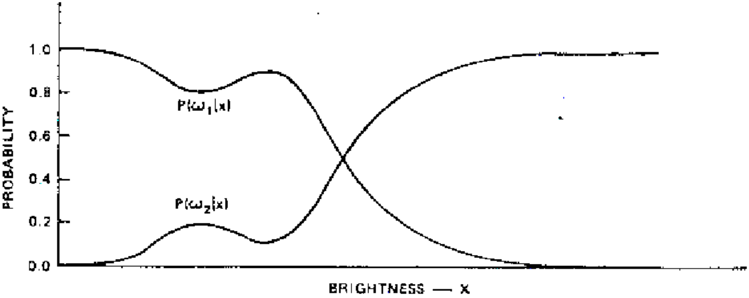
Alex Waibel, Kevin Kilgour

Example

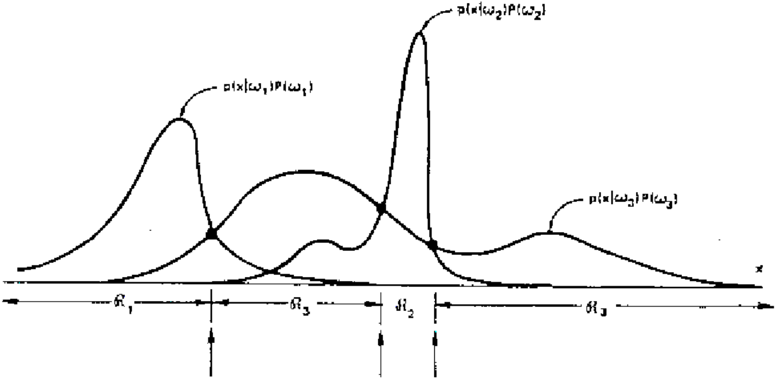
$$P(\omega_1) = 2/3$$
$$P(\omega_2) = 1/3$$



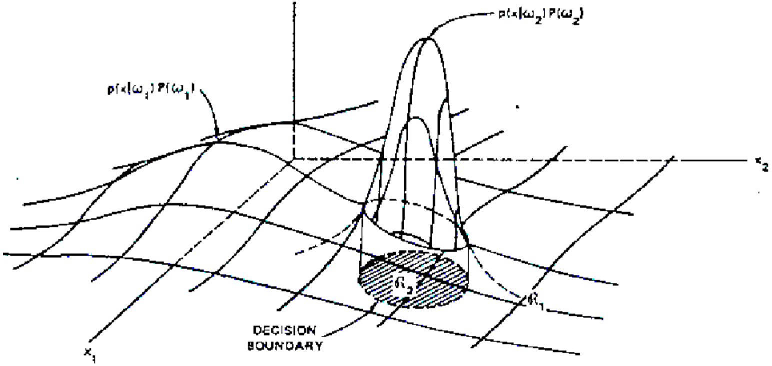
Example- A posteriori probabilities



Example - Decision Boundaries I



Example - Decision Boundaries II



Classifier Design in Practice

- Need a priori probability $P(\omega_i)$ (not too bad)
- Need class conditional PDF $p(x|\omega_i)$
- Problems:
  - limited training data
  - limited computation
  - class-labelling potentially costly and errorful
  - classes may not be known
  - good features not known
- Parametric Solution:
  - Assume that $p(x|\omega_i)$ has a particular prametric form
  - Most common representative: multivariate normal density

Gaussian Classifier

Univariate Normal Density:

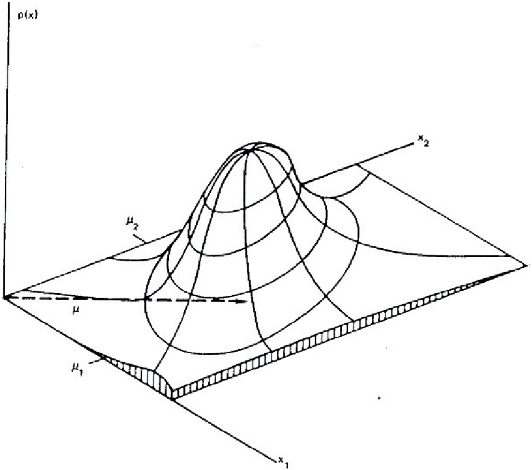
$$p(x) = \frac{e^{-\frac{1}{2}(\frac{\vec{x}-\vec{\mu}}{\sigma})^2}}{\sqrt{2\pi\sigma}}$$
$$\sim N(\vec{\mu}, \sigma^2)$$

Multivariate Density:

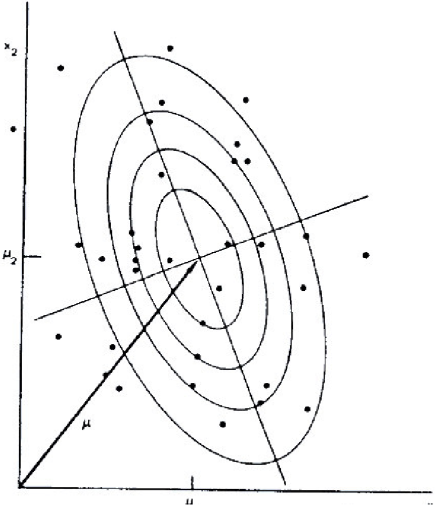
$$p(x) = \frac{e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})}}{(2\pi)^{\frac{1}{2}} |\Sigma|^{\frac{1}{2}}}$$
$$\sim N(\vec{\mu}, \Sigma)$$

Estimate using: MLE (Maximum Likelihood Estimation)

Example - Bivariate Normal Density



Example - Scatter Diagram



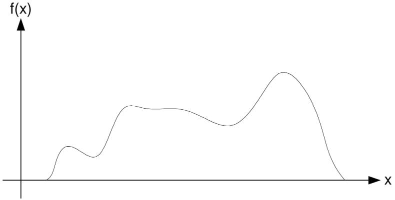
Problems of Classifier Design

- Features:
  - What and how many features should be selected?
  - Any features?
  - The more the better?
  - If additional features not useful, classifier will automatically ignore them?

Curse of Dimensionality

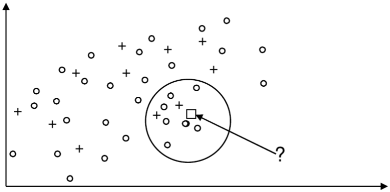
- Generally, adding more features indiscriminantly leads to worse performance!
- Reason:
  - Training Data vs. Number of Parameters
  - Limited training data.
- Solution:
  - select features carefully
  - Reduce dimensionality
  - Principle Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)

Problems



- Normal distribution does not model this situation well.
- Other densities may be mathematically intractable.
- $\Longrightarrow$ non-parametric techniques

K-Nearest Neighbours (KNN)



- To classify sample x:
  - Find k-nearest neighbours of x.
  - Determine the class most frequently represented among those k samples (take a vote)
  - Assign x to that class.
- Similar: Parzen Window

KNN-Classifier: Problem

- For finite number of samples n, we want k to be:
  - **large**: for reliable estimate
  - **small**: to guarantee that all k neighbours are reasonably close.
- Need training database to be larger.