Status of Mva based PID.

M. Babai

KVI/University of Groningen
The Netherlands.

August 31, 2010

M. Babai, J. Messchendorp, V. Suyam Jothi

Outline

Reminder on PR-systems

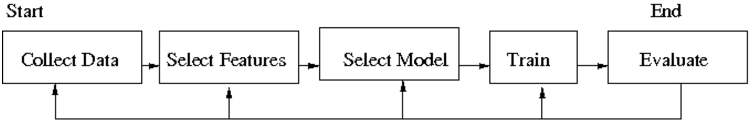
Available Algorithms

Panda-root Tasks

Current activities

Questions

Design of a Pattern recognition system



- Pre-processing: Select relevant information from data.
- Invariant: measurements do not have to change when the object appears in different context.
- Error-rate: Percentage of mis-assigned new patterns.
- Risk: Costs incorporation for each classification decision.
- (Cross-) Validation. Test set method, Leave one out, n-fold cross-validation, ...

Classification challenges

- The results depend on the variability of features.
- The variability can be affected by noise.
- How to cope with variability.

Available Algorithms.

- KNN (Density Estimator). Kd-tree based, standard (linear search) and KNN using projections.
  Pro: Easy to understand and use. Cons: Needs large data-set, relatively slow.
- Learning Vector Quantization (LVQ). LVQ1 and LVQ2.1 algorithms.
  Pros: Fast, small and easy to understand. Cons: Outputs are distances, difficult to find the optimal parameter set, time consuming training phase.

Available Algorithms (*Pre*-processing).

- Principal Component Analysis (PCA) based parameter transformation.
- K-Means Clustering.
  Proto-type initialization. *"Un-supervised"* class mean based clustering.

K- Nearest Neighbors

$$p_n(x) = \frac{K_n/n}{V_n}$$

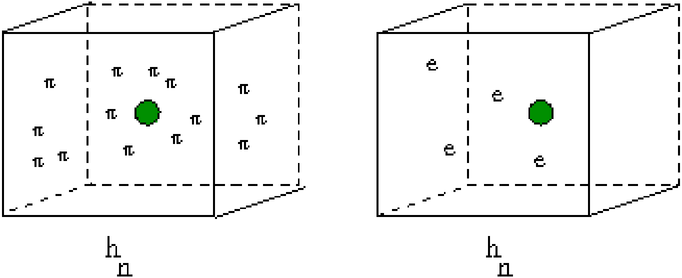
The cell is expanded until it encompasses K_n samples.

a posteriori probability is merely the fraction of samples within a cell with the label, k_i/k.

The Bayes decision rule becomes:

$$P(\omega_j|x) = \underset{i}{max} P(\omega_i|x).$$

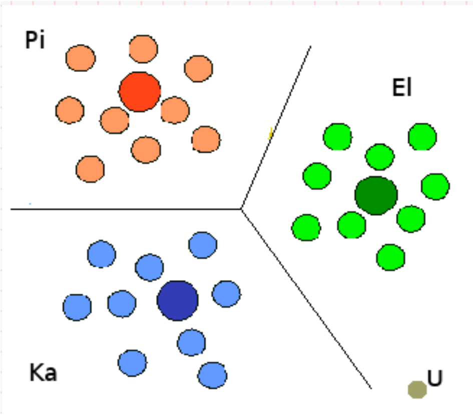
Probability densities (KNN)



$$\pi \; and \; e \; \in \{\text{Training data}\}$$

$$p_n(x) \propto \left(\frac{\#training \; elements}{h_n} \right)$$

Learning Vector Quantization



Available implementations

- These algorithms are available as library functions (libMva).
- "**PndPidMvaAssociatorTask**". Can be used with any set of parameters. User can control everything.
- "**PndPidEmcAssociatorTask**". Based on EMC only parameters. Only the classifier parameters can be set (the features are pre-defined).

The output is generated conform Panda Pid task ("PndPidProbability").

Using Mva's

- There are macro's and example programs available in "pandaroot/PndTools/MVA/".
- Weight files are available for KNN. Can be fetched from `kvit13.kvi.nl` (One can use the fetch script.).
- Documentation.
- Included parameters [p, E/p, lat, z20, z53].
- Labels $\{e,\ \pi,\ \mu$, K, p$\}$

Current activities:

- Pre-processing. Initialization scheme, parameter transformation, etc.
- Parameter normalization.
- Optimization of LVQ learning parameters and schemes.
- Performance (recognition quality) analysis and optimization.
- Combination of the results of different classifiers (Boosting, Bootstrap aggregating (bagging), etc.).
- Application to BESIII data.

Questions?