Lecture 4

Spring 2018

Faglærer: Magne Hallstein Johnsen,

Institutt for elektronikk og telekommunikasjon, NTNU

NTNU

Lecture content

- Principle for a reference/template based classifier
- Which distance to choose
- How to find (how many...) references
- References found by clustering
- NN (Nearest Neighbour) versus KNN (K Nearest Neighbour) based classification

Principle for a reference/template based classifier

- Assume N_i, references $X_i = \{x_{1i}, x_{2i}, \ldots, x_{N_i i}\}$ from each class ω_i $i = 1, \ldots, C$
- Define a distance measure $d(x, x_{ik})$ between an input x and an arbitrary reference x_{ik}
- NN decision rule :

$$x \in \omega_j \;\Leftrightarrow\; j = argmin_i \; [min_k \; d(x, x_{ik})] \tag{1}$$

- Different versions depend on how to :
  - choose distance measure
  - choose/find references
  - choose classification rule (NN or KNN)
- Principle idea of clustering

Which distance to choose

- The most general distance measure is the Mahanalobis distance :

$$d(x, x_{ik}) = (x - x_{ik})^T \Sigma_{ik}^{-1} \ (x - x_{ik}) \tag{2}$$

  where each reference x_{ik} has a unique covariance matrix Σ_{ik}^{-1}.

- A suboptimal alternative is to replace the above reference specific covariance matrixes by class specific matrixes :

$$d(x, x_{ik}) = (x - x_{ik})^T \Sigma_{i}^{-1} \ (x - x_{ik}) \tag{3}$$

- Or even use a common matrix for all the references/classes

$$d(x, x_{ik}) = (x - x_{ik})^T \Sigma^{-1} \ (x - x_{ik}) \tag{4}$$

- Another type of simplification is to assume diagonal matrixes

- Finally, the simplest distance is the Euclidian distance (i.e. using a unit matrix)

$$d(x, x_{ik}) = (x - x_{ik})^T (x - x_{ik}) \tag{5}$$

Options for finding the references

- We need access to a training set of M_i vectors x_{im} from each class ω_i $i = 1, \ldots, C$

- An obvious solution is to use all training vectors as references, i.e. $N_i = M_i$.
  - Simple, no algorithm needed to find references (++)
  - Many distances, $N = M = \sum_i M_i$, must be calculated (- -)
  - Reference specific matrixes can not be found; i.e. distance alternative given by equation (2) can not be used (-)

- In practice nearly all systems use a relatively small set $N_i << M_i$ of references. They can be found by :
  - random drawing from training set (- -)
  - clustering of training set (++)

Finding references by clustering

- Separate clustering for each class ω_i.
- The below algorithm is used iteratively for respectively $N_i = 2, 3, \ldots,$ references until no reduction is seen in total accumulated distance by increasing N_i
- Pick $N_i = 2$ references from the training set

  A) Classify all M_i training vectors into one of N_i clusters using equation (1)

  B) Find accumulated distance and calculate/update mean (and eventually covariance) for the cluster

  C) Loop to step A until no reduction in accumulated distance for N_i references

  D) Increase N_i, i.e. pick a new initial reference

- Alternative clustering methods exist which find references in a discriminative way

NN versus KNN classification

- NN classification is given by equation (1)
- KNN classification (with $K > 2$) :
  - Find K references which have lowest distance to input.
  - Decide on the class which has **most** references among the K.
  - If more than one class fulfils this, choose the one with lowest distance
- The KNN decison rule can be modified if large spread in number of references between classes
- It can be shown that the NN classifier in mean has twice the error rate of the TOC classifier.
- The KNN classifier is better than NN in nearly all cases
- How to choose K?