This document is also available in PDF[1] format.

Purpose: To introduce the Bayesian classifier while sticking to parametric PDFs.

Material: LECTURE NOTES[2] on Gaussians, Section 4.1 in notes. Also Devijver & Kittler.

General: The first classifier in this course. Useful for static (and dynamic - see HMM) patterns. Assumes the underlying densities are either known or can be estimated accurately. Easy to implement. Optimal if assumptions are valid.

Topics:

- Bayes' rule (Eq. 4.1) supply a way to calculate the posterior probability of a pattern class given a feature vector $P(\omega_i | \boldsymbol{x})$ with $1 \leq i \leq C$. A cost can be coupled to each classification decision. Minimising the expected cost leads to Eqs. 4.17 and 4.18. Note the respective roles played by the class-conditional densities $f(\boldsymbol{x}|\omega_i)$ and the prior probabilities $P(\omega_i)$.
- Using a sequence of feature vectors $\boldsymbol{x}_{1\ldots T}$, instead of only one, can enhance recognition accuracy. If the temporal relationship between the feature vectors is ignored (i.e. if it is assumed that the vectors are statistically independent), a PDF for the sequence can be calculated with Eq. 4.21. Substituting this PDF into the classifier results in Eqs. 4.22 and 4.23. We will see in a later lecture that another set of assumptions on the temporal relationship between the feature vectors results in the hidden Markov model (HMM).
- The Gaussian PDF has many good properties, in terms of modelling assumptions as well as computational tractability. Devijver&Kittler App A supplies expressions for estimating its mean and covariance matrix.
- The Bayes classifier is, however, not limited to the Gaussian PDF. The Gaussian mixture model (GMM) can model arbitrary functions and is related to the radial basis function (RBF) neural net. The HMM, which is a fairly sophisticated time-dependent model, ultimately also is only a parametric PDF. We will also see in a later lecture that a multi-layer perceptron (MLP) can be viewed as a posterior probability estimator.
- Note that both the above functions are examples of discriminant functions. Due to the multiplication of PDFs or probabilities, numerical under- or overflow problems are common. A monotonic increasing function of the original function will still result in a valid discriminant function. A very commonly used function for this purpose is the log function which changes the products to sums, resulting in Eqs. 4.24 and 4.25. Sometimes one encounters expressions such as $L = \log(e^{L_1} + e^{L_2} + \ldots + e^{L_M} + \ldots + e^{L_N})$, where none of the individual terms e^{L_n} are expressible in linear form. This is not as daunting as it might seem at first and can be calculated as $L = L_M + \log(e^{L_1 - L_M} + e^{L_2 - L_M} + \ldots + 1 + \ldots + e^{L_N - L_M})$ where $L_M = \max(L_1, L_2, \ldots L_N)$.

Project: (To be completed by the next lecture)

- The exponent of the Gaussian PDF contains the expression $(\boldsymbol{x} - \boldsymbol{a}_x)^T C_{XX}^{-1}(\boldsymbol{x} - \boldsymbol{a}_x)$. This closely resembles the squared Euclidean distance $(\boldsymbol{x} - \boldsymbol{a}_x)^T(\boldsymbol{x} - \boldsymbol{a}_x)$. Investigate and give a geometrical interpretation of the rôle of C_{XX}^{-1} in the Gaussian density. (Hint: use Choleski factorisation.)
- Using the `mean` and `cov` functions from Matlab implies two passes through the training feature vectors. How would you go about calculating both the mean and covariance while making only one pass through the data?
- In the following experiments, use both the simvowel set as well as either the timit set or the faces set as data sets. Follow the instructions in the data set document[3] on how to choose training and test sets. Represent each vowel/speaker/person with a multi-dimensional (full and diagonal covariance) Gaussian PDF and set up a Bayesian classifier. Experiment with different levels of the rejection option. Also use PCA and LDA to first reduce the feature vector dimension, and compare the results with those using the original feature vectors. Repeat using the optimal two-dimensional subspace. In this space, the PDFs and the effect of the rejection level etc. can be visualised. Make creative use of plots to illustrate your experiments.

---

[1] http://www.dsp.sun.ac.za/pr813/lectures/3_bayes/3_bayes.pdf

[2] http://www.dsp.sun.ac.za/pr813/lectures/lecture03/pr813_lecture03.pdf

[3] http://www.dsp.sun.ac.za/pr813/datasets/pr813_datasets.html