Lehrstuhl für Informatik VI
Rheinisch–Westfälische Technische Hochschule Aachen
Prof. Dr.–Ing. H. Ney

Friday, November 3rd, 2017

kim@cs.rwth-aachen.de
kitza@cs.rwth-aachen.de

4. Exercise Sheet **Statistical Classification and Machine Learning**

The solutions to the problems indicated by (*...) may be submitted until **Friday, November 10th, 2017**, till 08:00 o'clock by uploading them to L2P. You should work out the solutions in groups of up to **four** students.

Please note: For the submission of programming problems we require that you use Python, C or C++. Your program needs to compile and run platform-independent, we will not debug your code. Submit the output of your program and compiling/running instructions as well.

1. **Training of an Image Recognition System**
   A basic approach to image recognition is to use the intensity matrix of an image directly as observation "vector" for the corresponding image recognition system. Using such an approach, a set of pre-processing steps is needed to normalize the data with respect to, e.g., dynamic range of intensity values, position, size and orientation of the object represented by an image, which is to be recognized. Now assume that these normalization steps are already done, i.e. each image $n = 1, \ldots, N$ of our database is given by its class k_n and its normalized observation vector $x_n = x_{n,1}, \ldots, x_{n,D}$ of fixed dimension D with $x_{n,d} \in \mathbb{R} \quad \forall\ n = 1, \ldots, N$ and $d = 1, \ldots, D$. Now assume a *Gaussian* classification model, i.e. the class conditional probability of a single observation x of an image of class k is given by:

   $$p(x|k) = \mathcal{N}(x|\mu_k, \Sigma)$$

   with (class specific) mean vectors μ_k and a pooled covariance matrix Σ (all classes share the same covariance matrix).

   As attachment to this task you can find data for the so called *US Postal* (USPS) digit recognition task in the L2P room. The corpus consists of handwritten digits extracted from postal codes on letters. The digits are represented by a matrix of 16×16 gray values in the range between 0 and 1000. The data has been transformed into ASCII format and has been formatted such that you will be able to get an idea of the data by displaying the data with a font with fixed width. The original corpus has been pre-processed and divided into a training and a testing part, usps.train and usps.test. In usps.README you find a description of the formats of the training and testing data files as well as the parameter files to be produced. (In this exercise only pooled diagonal variances are used. The parameter file should start with a "d" and all variances are equal.)

   Assume the training data is given by a set of images $n = 1, \ldots, N$ which are represented by their classes k_n and the observation vectors x_n with $x_n = x_{n,1}, \ldots, x_{n,D}$ of fixed dimension D with $x_{n,d} \in \mathbb{R} \quad \forall\ n = 1, \ldots, N$ and $d = 1, \ldots, D$.

See second page on reverse side!

Later in the lecture it is going to be shown that the parameters of the *Gaussian* classification model μ_k, Σ can be calculated by

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N_k} x_n \quad (x_n \text{ out of set where } k_n = k) \tag{1}$$

$$\Sigma_{dd'} = \begin{cases} \sigma_d^2 & \text{for} \quad d = d' \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$\sigma_d^2 = \frac{1}{N} \sum_{n=1}^{N} (x_{nd} - \mu_{k_n,d})^2 \tag{3}$$

with N_k the count of $k_n = k$. In this exercise these equations are assumed as given. The later needed prior $p(k)$ is calculated as

$$p(k) = \frac{N_k}{N} \tag{4}$$

(a) Implement the above estimators for the parameters $\mu_k, \Sigma, p(k)$. The corresponding output parameter file should conform with the format given in usps.README (the estimator will produce diagonal covariance matrices). (* 8P)

(b) Use your program to estimate the parameters using the training data usps.train. (* 2P)

(c) Write an *efficient* implementation of a *Bayes* classifier using multivariate *Gaussian* class-conditional probabilities, which takes the parameter file produced in 1b and the above mentioned data files as input. (* 8P)

$$x \mapsto r(x) = \underset{k}{\operatorname{argmax}} \{p(k)p(x|k)\} \tag{5}$$

Before calculating this equation, try to simplify it, e.g. by removing constant parts. You may use publicly available libraries for common matrix operations.
The output of the classifier should consist of the empirical error rate ("wrong classification"/"all events") and a confusion matrix. An entry $cm_{k,k'}$ in a confusion matrix is the number of times an observation of class k has been classified to be class k', i.e. the sum over all entries of the confusion matrix equals the number of classified observations.

(d) Use your program to calculate the error rates and confusion matrices. (Hint: the resulting error rate is expected to be in the range of 20%.) (* 2P)