Lehrstuhl für Informatik VI **Thursday, November 8th, 2018**
Rheinisch–Westfälische Technische Hochschule Aachen
Prof. Dr.–Ing. H. Ney **{kim,kitza,gao}@cs.rwth-aachen.de**

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

Solutions to the problems indicated by (*...) must be submitted by **08:00 of Thursday, November 15th, 2018** via L2P. Please form a group of up to **three** students! (Group Workspace in L2P)

1. Training of an Image Recognition System

A basic approach to image recognition is to use the intensity matrix of an image as observation "vector" for the corresponding image recognition system. As attachment to this task you can find *US Postal* (USPS) digit recognition data. 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 corpus has been pre-processed and divided into a training and a testing part, `usps.train` and `usps.test`. In `usps.README`, you can find a description of the formats of the training and testing data files as well as the parameter files to be produced.

Each image $n = 1, \ldots, N$ of the training data is given by its class k_n and its observation vector $x_n = x_{n1}, \ldots, x_{nD}$ of fixed dimension D with $x_{nd} \in \mathbb{R}$ for $d = 1, \ldots, D$. Now assume a Gaussian classification model, i.e. the class conditional probability of a single observation $x \in \mathbb{R}^D$ 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).

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 $k_n = k$. The prior $p(k)$ is calculated as

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

Using these equations, do the followings:

(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)

Warning: In this exercise, only pooled diagonal variances are used. The parameter file should start with a "d" and all variances are equal.)

(b) Use your program to estimate the parameters using the training data `usps.train` and store the parameter values in: (* 2P)

- `usps_d.param` (please keep the **correct file format**! see `usps.README`)

Warning: In your parameter file, please do NOT copy and paste the comments for each parameter in `usps.README`, e.g. "`: number of classes (10 digits, 10 itself represents digit zero)`". Write only the parameter values in the file.

(c) Now consider a classifier based on the Bayes decision rule: (* 2P)

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

Expand this equation using the given model assumptions and simplify as much as possible, e.g. by removing constant parts.

(d) Implement the classifier of 1c, which takes the parameter file produced in 1b and the data files `usps.train` or `usps.test` as input. You may use publicly available libraries for common matrix operations. Please make it **efficient**. (* 6P)

The output of the classifier should consist of the **empirical error rate**:

$$\text{empirical error rate} = \frac{\text{wrong classifications}}{\text{all events}}$$

and the **confusion matrix** M. An entry $M_{kk'}$ in a confusion matrix M is the number of times when an observation of class k is classified to be class k' by your classifier. Accordingly, the sum over all entries of the confusion matrix equals the number of classified observations.

(e) Use your program to calculate the error rates and confusion matrices. (Hint: the resulting error rate is expected to be around 20%.) Store the results in the following file names: (* 2P)

- Error rate: `usps_d.error` (just a single line with the error rate, nothing else)
- Confusion matrix: `usps_d.cm`
  (each line with a row of M, with columns tab-separated; do NOT print Python list directly)

Warning:

- Only **Python** is allowed for your implementations.
- Please follow instructions above for each output file format. Otherwise, you may get penalized even if your result is correct.
- We will also check the readability of your code. Please make the code structure clear and understandable with descriptive variable/function names. Put sufficient comments in your code.