TU München
Fakultät für Informatik
PD Dr. Rudolph Triebel
John Chiotellis

Machine Learning for Robotics and Computer Vision
Summer term 2016

Homework Assignment 4

Topic 1: Boosting
May 30, 2016

Exercise 1: Adaboost (Programming)

Download the file 'banknote_auth.zip' available at the course's website. The data are features of banknotes and the labels indicate whether a banknote is forged or not. The dataset is taken from `https://archive.ics.uci.edu/ml/datasets/banknote+authentication` with some duplicate entries removed. Implement the AdaBoost algorithm with decision stumps as weak classifiers.

a) To begin train on 50% of the data with 20 weak classifiers.

b) Generate a plot of the classification error with respect to the number of weak classifiers. What do you observe?

c) Add more weak classifiers. Does the error decrease? What's the optimal number of weak classifiers to use?

d) Now keep the number of weak classifiers fixed and try different training/testing set sizes. How does it affect the classification accuracy?

Topic 2: Mixture Models and EM

Exercise 2: Expectation-Maximization for GMM

In the standard EM algorithm, we first define the responsibilities γ as

$$\gamma_{nk} = p(z_{nk} = 1|x_n) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} \quad , z_{nk} \in \{0, 1\}, \sum_{k=1}^{K} z_{nk} = 1$$

a) Find the optimal means, covariances and mixing coefficients that maximize the data likelihood. How can we interpret the results?

b) Define the complete-data-log-likelihood. What is the difference to the standard log-likelihood?

Exercise 3: K-Means and Expectation-Maximization (Programming)

Download the fisher-iris.zip file. Inside you will find Fisher's Iris dataset. It consists of 150 flowers that have 4 properties and belong to 3 different species. Implement following exercises in your preferred programming language.

a) Implement the K-Means algorithm to cluster the dataset.

b) Run your algorithm 100 times with randomized initial values for the cluster centers. Each time compare the result with the true labels. How much does the performance change for different initializations?

c) Now implement the Expectation-Maximization algorithm (EM) for Gaussian mixture models (GMM) using a fixed symmetric covariance (supplied as a parameter) and apply it to the same dataset.

d) Again evaluate the algorithm for different initializations of cluster centers and covariances. How do the EM errors compare to K-means on average? Does EM converge faster? Is it more reliable in finding the true means?

The next exercise class will take place on **June 10th, 2016**.

For downloads of slides and of homework assignments and for further information on the course see

`https://vision.in.tum.de/teaching/ss2016/mlcv16`
