Universität des Saarlandes
FR 4.7 Sprach- und Signalverarbeitung

Prof. Dietrich Klakow
Pattern and Speech Recognition

Exercise Sheet 4

Machine Learning Basics

Deadline: 6.12.2016, 23:59

k-Means Clustering

Exercise 4.1

In this exercise, you will implement k-means clustering algorithm.

a) Download the data from course website. Load *data_kmeans.txt* and plot the 2 dimensional datapoints. (1 point)

b) Implement k-means algorithm as follows: (6 points)

Let X = $\{x_1, x_2 ... x_n\}$ be the set of datapoints, and C = $\{c_1 ... c_k\}$ be the cluster centers (initialized randomly). Implement the steps described in slides and iterate till cluster centers don't change OR the objective J doesn't change, where

$$J = \sum_{i=1}^{k} \sum_{x \in c_i} ||x - c_i||^2$$

(Refer: https://en.wikipedia.org/wiki/K-means_clustering for more details)

c) Plot the clustering results for k=2. Use different colors to represent the clusters (1 point)

d) What is the smallest k for which J attains the value zero? (Present a theoritical arguement, don't run your code for different values of k)(1 point)

Maximum Likelihood Estimation

Exercise 4.2

A football team scores 2,3,0,2,1 and 5 goals in six matches played. Assuming these samples are drawn from a Poission distribution, find the maximum likelihood estimate for the parameter λ.(Match outcomes are independent of each other) What is the probability that the team will score 2 goals in the next match? (3 points)

Composite functions

Exercise 4.3

Compute the first and second order partial derivatives for the following function $f(x, y) = \log(\sin(xy))$ (3 points)

Classification

Exercise 4.4

In this task, you will use logistic regression to classify Iris plants as into two categories: Setosa and Virginica based on the sepal length and petal width of flowers. (5 points)

Download the Iris dataset from course website. (filename: *iris.data*)

- Delete the last 50 rows, i.e. data corresponding to Versicolor class.
- Use column-1 (sepal length) and column-4 (petal width) from the file as features. The last column contains the ground truth data.
- Modify the code *logistic_regression.py* provided on the course website and use it for the classification task.
- Plot the classification result and decision boundary.

Submission instructions

The following instructions are mandatory. If you are not following them, tutors can decide to not correct your exercise.

Submission architecture

You have to generate a **single ZIP file** respecting the following architecture:

```
tutorial2_<matriculation_nb1>_<matriculation_nb2>_<matriculation_nb3>
|
+--- source
|      |
|      +------- file 1
|      +------- file 2
|      +------- ...
+--- report.pdf
+--- README.txt
```

where

- **source** contains the source code of your project,

- **report.pdf** is the report where you present your solution with **the explanations** and the plots.
- **README** which contains group member informations (name, matriculation numbers and emails) and a **clear** explanation about how to compile and run your source code

The ZIP filename has to be :

`tutorial2_<matriculation_nb1>_<matriculation_nb2>_<matriculation_nb3>.zip`

Some hints

We advice you to follow the following guidelines in order to avoid problems :

- Avoid building complex systems. The exercises are simple enough.
- Do not include any executables in your submission, as this will cause the e-mail server to reject it.

Grading

Send your assignment to the tutor who is responsible of your group:

- Merlin Köhler s9mnkoeh@stud.uni-saarland.de
- Goutam Y G goutamyg@lsv.uni-saarland.de
- Ahmad Taie s8ahtaie@stud.uni-saarland.de

The email subject should start with `[PSR TUTORIAL 4]`