MATH 368
Dr. Andreas Alpers

UNIVERSITY OF LIVERPOOL

Project 2
Submit by: 14.05.2021
(deadline might slightly change, watch out for announcements)

Face Recognition Using Support Vector Machines

1. Part A

In this first part of the project consider the $m = 11$ two-dimensional data points

$$x_1 = (-1, 2), \quad x_2 = (-0.8, 0.7), \quad x_3 = (0.5, -1), \quad x_4 = (-2, 0), \quad x_5 = (2, 0),$$
$$x_6 = (0, 3), \quad x_7 = (1, 4), \quad x_8 = (4, 6), \quad x_9 = (5, 3), \quad x_{10} = (4, 4.5), \quad x_{11} = (5.5, 4.0)$$

with corresponding labels $y_1 = y_2 = \cdots = y_6 = -1$ and $y_7 = y_8 = \cdots = y_{11} = 1$.

T1: Write a matlab function

```
function [w,beta] = svm(training_points, training_labels)
```

that takes as input a $d \times m$ matrix `training_points` (containing the d-dimensional data points in the columns) and an m-dimensional row-vector `training_labels` (its ith component containing the label $y_i \in \{\pm 1\}$ of the data point x_i), and which —by using matlab's `quadprog` command[1]— computes the maximum margin separating hyperplane $H_{w,\beta}$ for the data points, if it exists (in this case w and β should be returned, otherwise the message '`Data is not linearly separable.`' should be printed).

T2: Apply your function `svm` to the data and generate a plot that shows the hyperplanes $H_{w,\beta-1}$, $H_{w,\beta}$, $H_{w,\beta+1}$, and the 11 data points (points with label -1 shown in red, points with label $+1$ shown in blue).

T3: Given an answer to the following question (either by giving a proof or counterexample): Is it TRUE or FALSE that the maximum margin separating hyperplane $H_{w,\beta}$ is always (i.e., for any dataset) perpendicular to the line connecting the two closest points from both label classes?

T4: We consider now (but only in this task) an extension of the support vector machine approach. For this, add to each data point $x_i = (\xi_{i1}, \xi_{i2})$, $i = 1, \ldots, 11$, a third component that equals the sum of the squared first two components. So x_i becomes $x_i' = (\xi_{i1}, \xi_{i2}, \xi_{i3})$ with $\xi_{i3} = \xi_{i1}^2 + \xi_{i2}^2$. Apply your function `svm` to this three-dimensional data set to obtain $H_{w,\beta}$. With this $H_{w,\beta}$ classify at least 250 points $t_i' = (\tau_{i1}, \tau_{i2}, \tau_{i1}^2 + \tau_{i2}^2)$ in the range $-2 \leq \tau_{i1}, \tau_{i2} \leq 6$. Generate a two-dimensional plot of the (τ_{i1}, τ_{i2}), $i = 1, \ldots, 11$, colored according to this classification in three dimensions. Is the decision boundary between the two label classes linear?

[1] To use this command, make sure you have matlab's optimization toolbox installed.

2. Part B

Here we consider face recognition. As dataset we will use the images from *LFWcrop*, which is a cropped version of the popular *Labeled Faces in the Wild (LFW)* dataset. You can find the zipped images in the file `lfwcrop_grey.zip` downloadable from the website https://conradsanderson.id.au/lfwcrop/.

T5: Each of the images in the dataset is a 64×64 image. Using matlab's `im2double` command, each of these images can be converted into a $\mathbb{R}^{64\times 64}$ matrix. If we assemble them into a $\mathbb{R}^{64\cdot 64}$ vector we can use the support vector machine approach classifying points in 4096-dimensional space. Taking this approach, use your function `svm` from T1 and train it on the data set consisting of the 60 images

`George_W_Bush_0001.pgm`, ..., `George_W_Bush_0030.pgm` (label them with -1),
`Tony_Blair_0001.pgm`, ..., `Tony_Blair_0030.pgm` (label them with $+1$).

Now classify the 10 images `Tony_Blair_0081.pgm`, ..., `Tony_Blair_0090.pgm`. Report which of these images are correctly classified as showing Tony Blair, and which are misclassified.

3. What and how to submit

Each group is expected to submit a **single PDF** containing answers/code for (a)-(e) stated below (submit via Canvas as group submission). **Each member** of the group is expected to take part in the peer moderation on Canvas using **Buddycheck** upon submission (failure to do so, will result in a subtraction of 5 points from your individual grade after moderation).

(a) [**20 points**] The documented matlab code for your function `svm` in T1.

(b) [**20 points**] The plot requested in T2.

(c) [**20 points**] Answer (proof or counterexample) to the question in T3.

(d) [**20 points**] (i) The plot and (ii) the answer to the question 'is the decision boundary linear' requested in T4.

(e) [**20 points**] (i) The matlab code for training the svm in T5 and (ii) report which of the images are correctly classified as showing Tony Blair, and which are misclassified.

Academic Integrity

The work submitted must be academically sound, without elements of plagiarism or other features of poor practice. More resources on Academic Integrity, and the consequences of Academic Misconduct, can be found here, and the examiner expects that all students complete the KnowHow Academic Integrity module (to be found on Canvas, under KnowHow: Study for Success) before submitting any assignment.
Advice and responsibilities for students working in groups are given in Paragraph 5 of the Code of Practice on Assessments Appendix G. The examiner expects that all students familiarise themselves with this section of the document before commencing group work.