Optimization for Machine Learning
01SQKMV
Practice Sheet 1

Visualization of data

Exercise 1 (Bar Graph)**.** A simple, jet effective, way of visualizing data that vary over time is to draw a bar graph that represents them. A bar graph is a graph that presents data with rectangular bars with heights or lengths proportional to the values that they represent.

Download the file `norway_new_car_sales_by_make.csv`, which contains the monthly car sales in Norway for the period 2007-2017 by make.

Draw a bar graph representing the number of cars sold by Volkswagen, Toyota, Peugeot, Tesla, and Fiat in Feb. 2014, Feb. 2015 and Feb 2016.

Can you extract any information about car sales in Norway for the years 2017-2018 from these data?

Exercise 2 (Scatter Plot)**.** A scatter plot is a type of diagram using Cartesian coordinates to display values for two variables for a set of data. This type of plots is very useful when the task is to describe how data are organized (or clustered) without having any measure of the output (*unsupervised learning*).

Download the file `weight-height.csv`, which contains the sex, height and weight of 10000 people.

Draw a scatter plot representing such data, by using different colors form males and females. Can this plot be used to identify whether a person is a male or female just on the basis of his/her height and weight?

Scatter plots can be used to represent n-dimensional data as well by using the so called *scatter matrix*, which represents the pairwise dependencies among variables in an array. In particular, each entry of such an array is the scatter plot of one of the variables vs another variable.

Download the file `Iris.csv`, which contains data about the sepal length, sepal width, petal length, and petal width of 150 irises of 3 different species and draw the scatter matrix of such data.

Exercise 3 (Visualizing 3D Data)**.** A contour line of a function of two variables is a curve along which the function has a constant value. A contour plot is a bi-dimensional plot in which multiple contour lines are depicted at the same time.

Consider the polynomial

$$z = x^4y^2 + x^2y^4 - 3x^2y^2 + 1.$$

Draw the surface and the contour plots of such a function in the range $[-1.3, 1.3] \times [-1.3, 1.3]$. Can you identify the minima of such a polynomial from such a plot?

Exercise 4 (Parallel Coordinates Plot)**.** Visualize data in an n-dimensional space can be challenging if $n > 3$. A simple way to represent n-dimensional data is to produce a parallel coordinates plot, which consists of n parallel lines, typically vertical and equally spaced. A point in the n-dimensional space is represented as a polyline with vertices on the parallel axes; the position of the vertex on the i-th axis corresponds to the i-th coordinate of the point. This type of plot is useful when attempting at classifying an object on the basis of its features (*supervised learning*).

Download the file `Iris.csv`, which contains data about the sepal length, sepal width, petal length, and petal width of 150 irises of 3 different species.

Draw a parallel coordinates plot representing the sepal length, sepal width, petal length, and petal width of the flowers, by using different colors for each specie. Then, draw a modified parallel coordinates plot in order to show only the median, 25%, and 75% quartile values for each flower. Can you use this plot to distinguish an iris on the basis of its sepal length, sepal width, petal length, and petal width?

Hint: use the command `parallelcoords`.

Exercise 5 (Projection onto lower dimensional subspaces)**.** Given n-dimensional data $x^{(1)}, \dots, x^{(N)}$, a simple way to store and represent them is the so called *data matrix*, that is the $n \times N$ dimensional matrix

$$x = [\ x^{(1)} \quad \cdots \quad x^{(N)}\ ].$$

In order to better analyze the differences among the collected data, define the centered datum $x_c^{(i)} = x^{(i)} - \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$ and the *centered data matrix*

$$x_c = [\ x_c^{(1)} \quad \cdots \quad x_c^{(N)}\ ].$$

Since visualizing the centered data matrix can be challenging if $n > 3$, it may be convenient to represent the projection of the data either along a selected direction or onto a two-dimensional subspace.

Download the file `Iris.csv`, which contains data about the sepal length, sepal width, petal length, and petal width of 150 irises of 3 different species.

1. Represent the projection of the centered data matrix corresponding to such a dataset along the axis $W_1 = \mathrm{Span}(v)$, where

$$v = [\ 1 \quad -1 \quad \tfrac{1}{2} \quad \tfrac{3}{4}\ ]^\top.$$

2. Represent the projection of the data onto the subspace $W_2 = \mathrm{Span}(v_1, v_2)$, where

$$v_1 = [\ \tfrac{3}{5} \quad 0 \quad -\tfrac{4}{5} \quad 0\ ]^\top, \quad v_2 = [\ 0 \quad \tfrac{3}{5} \quad 0 \quad -\tfrac{4}{5}\ ]^\top$$

and onto the subspace

$$W_3 = \left\{ x \in \mathbb{R}^4 : \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} x = 0 \right\}.$$

Hint: Given a vector v such that $\|v\|_2 = 1$, the projection of the datum $x^{(i)}$ along $\mathrm{Span}(v)$ is given by $s^{(i)} = v^\top x^{(i)}$.

On the other hand, given orthonormal vectors v_1, v_2, the projection of the datum $x^{(i)}$ along $\mathrm{Span}(v_1, v_2)$ is given by $s^{(i)} = A^\top x^{(i)}$, where $A = [\ v_1 \quad v_2\ ]$.