MACHINE LEARNING 1: ASSIGNMENT 3

Tom Nick 340528
Niklas Gebauer 340942

Exercise 1

(a) In the following we want to minimize the objective function

$$J(\theta) = \sum_{k=i}^{n} \|\theta - x_k\|^2$$

subject to the constraint $\theta^T b = 0$, where $x_1, ..., x_k, \theta, b \in \mathbb{R}^d$.

Therefore we will define the Lagrangian function and set its gradient to zero in order to solve for λ and θ:

$$\mathcal{L}(\theta, \lambda) = J(\theta) + \lambda\theta^T b = \sum_{k=i}^{n} \|\theta - x_k\|^2 + \lambda\theta^T b$$
$$\nabla\mathcal{L} = \begin{pmatrix} [\sum_{k=i}^{n} 2(\theta - x_k)] + \lambda b \\ \theta^T b \end{pmatrix} \stackrel{!}{=} \vec{0}$$

We will at first solve the first entry of our gradient for θ:

$$[\sum_{k=i}^{n} 2(\theta - x_k)] + \lambda b = 0$$
$$\Leftrightarrow 2n\theta - 2[\sum_{k=i}^{n} x_k] + \lambda b = 0$$
$$\Leftrightarrow 2[\sum_{k=i}^{n} x_k] - \lambda b = 2n\theta$$
$$\Leftrightarrow \frac{1}{n}[\sum_{k=i}^{n} x_k] - \frac{1}{2n}\lambda b = \theta$$

Let's plug this into the second entry to solve for λ:

$$\theta^T b = 0$$
$$\Leftrightarrow (\frac{1}{n}[\sum_{k=i}^{n} x_k] - \frac{\lambda}{2n} b)^T b = 0$$
$$\Leftrightarrow (\frac{1}{n}[\sum_{k=i}^{n} x_k^T] - \frac{\lambda}{2n} b^T) b = 0$$
$$\Leftrightarrow \frac{1}{n}[\sum_{k=i}^{n} x_k^T] b - \frac{\lambda}{2n} b^T b = 0$$
$$\Leftrightarrow \frac{1}{n}[\sum_{k=i}^{n} x_k^T] b = \frac{\lambda}{2n} b^T b$$
$$\Leftrightarrow \frac{2}{b^T b}[\sum_{k=i}^{n} x_k^T] b = \lambda$$

Finally, we can plug this λ into our formula describing θ, yielding the result:

$$\begin{aligned}
\theta &= \frac{1}{n}[\sum_{k=i}^{n} x_k] - \frac{1}{2n}\lambda b \\
\Leftrightarrow \theta &= \frac{1}{n}[\sum_{k=i}^{n} x_k] - \frac{1}{2n}(\frac{2}{b^T b}[\sum_{k=i}^{n} x_k^T]b)b \\
\Leftrightarrow \theta &= \frac{1}{n}([\sum_{k=i}^{n} x_k] - \frac{[\sum_{k=i}^{n} x_k^T b]}{b^T b}b)
\end{aligned}$$

Similar to the minimal solution for θ of the unconstrained objective, we again find the the empirical mean in our solution. But this time we subtract a scaled version of the vector b. This means, that the solution still has to lie close to the empirical mean. If we imagine a plot with contour lines of $J(\theta)$ there has to be a 'valley' which is the empirical mean and a line cutting the surface which specifies the points where our constraint holds. The second term of our solution will move θ from total minimum (the empirical mean) to the point where the line lies deepest in the plane and where the lines and J's gradient both are perpendicular to the line. So we have the minimal solution that still fulfills the constraint.

(b) Now we will repeat the same procedure for the objective above with a different constraint ($\|\theta - c\|^2 = 1, c \in \mathbb{R}^d$):

$$\begin{aligned}
\mathcal{L}(\theta, \lambda) &= J(\theta) + \lambda\|\theta - c\|^2 - \lambda \\
\nabla\mathcal{L} &= \begin{pmatrix} [\sum_{k=i}^{n} 2(\theta - x_k)] + \lambda 2(\theta - c) \\ \|\theta - c\|^2 - 1 \end{pmatrix} \stackrel{!}{=} \vec{0}
\end{aligned}$$

Solve the first entry for θ:

$$\begin{aligned}
&[\sum_{k=i}^{n} 2(\theta - x_k)] + \lambda 2(\theta - c) = 0 \\
\Leftrightarrow\ & 2n\theta - 2[\sum_{k=i}^{n} x_k] + 2\lambda\theta - 2\lambda c = 0 \\
&\Leftrightarrow 2n\theta + 2\lambda\theta = 2[\sum_{k=i}^{n} x_k] + 2\lambda c \\
\Leftrightarrow\ & \theta = \frac{1}{n+\lambda}([\sum_{k=i}^{n} x_k] + \lambda c)
\end{aligned}$$

Let's plug this into the second entry to solve for λ:

$$\begin{aligned}
\|\theta - c\|^2 - 1 &= 0 \\
\Leftrightarrow \|\frac{1}{n+\lambda}([\sum_{k=i}^{n} x_k] + \lambda c) - c\|^2 - 1 &= 0 \\
\Leftrightarrow \|\frac{1}{n+\lambda}[\sum_{k=i}^{n} x_k] + \frac{\lambda}{n+\lambda}c - \frac{n+\lambda}{n+\lambda}c\|^2 - 1 &= 0 \\
\Leftrightarrow \|\frac{1}{n+\lambda}[\sum_{k=i}^{n} x_k] - \frac{n}{n+\lambda}c\|^2 &= 1 \\
\Leftrightarrow \|\frac{1}{n+\lambda}([\sum_{k=i}^{n} x_k] - nc)\|^2 &= 1 \\
\Leftrightarrow (\frac{1}{n+\lambda}([\sum_{k=i}^{n} x_k] - nc))^T(\frac{1}{n+\lambda}([\sum_{k=i}^{n} x_k] - nc)) &= 1 \\
\Leftrightarrow \frac{1}{(n+\lambda)^2}([\sum_{k=i}^{n} x_k] - nc)^T([\sum_{k=i}^{n} x_k] - nc) &= 1 \\
\Leftrightarrow \|[\sum_{k=i}^{n} x_k - c]\|^2 &= (\lambda + n)^2 \\
\Leftrightarrow \|[\sum_{k=i}^{n} x_k - c]\|^2 &= \lambda^2 + 2n\lambda + n^2 \\
\Leftrightarrow 0 &= \lambda^2 + 2n\lambda + n^2 - \|[\sum_{k=i}^{n} x_k - c]\|^2 \\
\overset{pq-formula}{\Rightarrow} \frac{-2n}{2} \pm \sqrt{\frac{2n}{2}^2 - (n^2 - \|[\sum_{k=i}^{n} x_k - c]\|^2)} &= \lambda_{1,2} \\
\Leftrightarrow -n \pm \sqrt{\|[\sum_{k=i}^{n} x_k - c]\|^2} &= \lambda_{1,2} \\
\Rightarrow -n + \|[\sum_{k=i}^{n} x_k - c]\| &= \lambda_1 \\
\Rightarrow -n - \|[\sum_{k=i}^{n} x_k - c]\| &= \lambda_2
\end{aligned}$$

We can now compute two solutions for θ using λ_1 and λ_2:

λ_1 :

$$\begin{aligned}
\theta_1 &= \frac{1}{n+\lambda_1}([\sum_{k=i}^{n} x_k] + \lambda_1 c) \\
&= \frac{1}{n + (-n + \|[\sum_{k=i}^{n} x_k - c]\|)}([\sum_{k=i}^{n} x_k] + (-n + \|[\sum_{k=i}^{n} x_k - c]\|)c) \\
&= \frac{1}{\|[\sum_{k=i}^{n} x_k - c]\|}([\sum_{k=i}^{n} x_k] - nc + \|[\sum_{k=i}^{n} x_k - c]\|c) \\
&= \frac{1}{\|[\sum_{k=i}^{n} x_k - c]\|}[\sum_{k=i}^{n} x_k - c] + c
\end{aligned}$$

λ_2 :

$$\begin{aligned}
\theta_2 &= \frac{1}{n+\lambda_2}([\sum_{k=i}^{n} x_k] + \lambda_2 c) \\
&= \frac{1}{n + (-n - \|[\sum_{k=i}^{n} x_k - c]\|)}([\sum_{k=i}^{n} x_k] + (-n - \|[\sum_{k=i}^{n} x_k - c]\|)c) \\
&= -\frac{1}{\|[\sum_{k=i}^{n} x_k - c]\|}([\sum_{k=i}^{n} x_k] - nc - \|[\sum_{k=i}^{n} x_k - c]\|c) \\
&= \frac{1}{\|[\sum_{k=i}^{n} x_k - c]\|}[\sum_{k=i}^{n} c - x_k] + c
\end{aligned}$$

This time we again see something similar to the mean, but θ is dependend on c. In both formulas we have a normalized vector (length = 1) and add up c, so that the constraint holds (if we subtract c again we will have a vector thatstill has length one). Since we know from the objective without constraint that the empirical mean is the minimum, it's clear that θ_1 is the minimizing parameter: it stays close to the (normalized mean) but every data point gets a little offset of -c so that the latter addition of c doesn't take the result too far away from the empirical mean.

θ_2 is rather close to c since it centers all the points around c. Thus it won't yield as small results in our objective function as θ_1, which is closer to the empirical mean.

Exercise 2

The Scatter-matrix S is defined as

$$\sum_{k=1}^{n} (x_k + m)(x_k + m)^T$$

where m is defined as the mean

$$m = \frac{1}{n} \sum_{k=1}^{n} x_k$$

.

(a) It is to show, that the trace tr of S is an upper bound for λ_1 which denotes the highest eigenvalue of S, the trace of S is defined as $\sum_{i=1}^{d} S_{ii}$.

We use three facts to prove this:

1. S **is symmetric**
   It is easy to see, that S is a symmetric matrix - a vector multiplied by its transposed self creates always a symmetric matrix ($AA^T = (AA^T)^T \Leftrightarrow (A^T)^T A^T \Leftrightarrow AA^T$), $(x_k + m)(x_k + m)$ results in a symmetric matrix therefore as well. The sum of two symmetric matrices results in another symmetric matrix, S is therefore symmetric.
2. S **can be decomposed into** $Q\Lambda Q^T$ **where** Λ **is a diagonal with the eigevalues of** S **as values.**
3. $\text{tr}(AQA^T) = \text{tr}(Q)$ when A is orthogonal
   $\text{tr}(AQA^T) = \text{tr}((AQ)A^T) = \text{tr}(A^T(AQ)) = \text{tr}(Q)$

$$\text{tr}(S) = \text{tr}(Q\Lambda Q^T) = \text{tr}(\Lambda) = \sum_{k=1}^{d} \lambda_k$$

The trace of S is therefore the sum of all it's Eigenvalues, which is trivially an upper-bound for the largest Eigenvalue λ_1.

(b) As the upper bound is the sum of all Eigenvalues, it is tighter the smaller the Eigenvalues $\lambda_2 \dots \lambda_d$ relativly to the Eigenvalue λ_1 are. This means for the data, that only one feature is really significant for the variance of the data.

(c) From it is known, that S is decomposable into $Q\Lambda Q^T$ with Λ containing the eigenvalues of S $\lambda_1 \dots \lambda_n$ on its diagnoal.

(d)

Exercise 3

(a)

(b)

(c) (*bonus)