

# MACHINE LEARNING 1: ASSIGNMENT 3

Tom Nick 340528  
Niklas Gebauer 340942

## Exercise 1

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

$$J(\theta) = \sum_{k=1}^n \|\theta - x_k\|^2$$

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

Therefore we will define the Lagrangian function and set its gradient to zero in order to solve for  $\lambda$  and  $\theta$ :

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

We will at first solve the first entry of our gradient for  $\theta$ :

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

Let's plug this into the second entry to solve for  $\lambda$ :

$$\begin{aligned}\theta^T b &= 0 \\ \Leftrightarrow \left(\frac{1}{n}[\sum_{k=1}^n x_k] - \frac{\lambda}{2n}b\right)^T b &= 0 \\ \Leftrightarrow \left(\frac{1}{n}[\sum_{k=1}^n x_k^T] - \frac{\lambda}{2n}b^T\right)b &= 0 \\ \Leftrightarrow \frac{1}{n}[\sum_{k=1}^n x_k^T]b - \frac{\lambda}{2n}b^T b &= 0 \\ \Leftrightarrow \frac{1}{n}[\sum_{k=1}^n x_k^T]b &= \frac{\lambda}{2n}b^T b \\ \Leftrightarrow \frac{2}{b^T b}[\sum_{k=1}^n x_k^T]b &= \lambda\end{aligned}$$

Finally, we can plug this  $\lambda$  into our formula describing  $\theta$ , yielding the result:

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

Similar to the minimal solution for  $\theta$  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  $\theta$  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 ( $\|\theta - c\|^2 = 1, c \in \mathbb{R}^d$ ):

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

Solve the first entry for  $\theta$ :

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

Let's plug this into the second entry to solve for  $\lambda$ :

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

We can now compute two solutions for  $\theta$  using  $\lambda_1$  and  $\lambda_2$ :

$\lambda_1$  :

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

$\lambda_2 :$

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

This time we again see something similar to the mean, but  $\theta$  is dependent 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 that still has length one). Since we know from the objective without constraint that the empirical mean is the minimum, it's clear that  $\theta_1$  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.

$\theta_2$  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  $\theta_1$ , 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  $\text{tr}$  of  $S$  is an upper bound for  $\lambda_1$  which denotes the highest eigenvalue of  $S$ , the trace of  $S$  is defined as  $\sum_{i=1}^d S_{ii}$ .

We use four 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 ( $AA^T = (AA^T)^T \Leftrightarrow AA^T = (A^T)^T A^T \Leftrightarrow AA^T = AA^T$ ),  $(x_k - m)(x_k - m)^T$  results in a symmetric matrix therefore as well. The sum of two symmetric matrices results in another symmetric matrix,  $S$  is therefore symmetric.

### 2. $S$ is positive semidefinite

For each vector  $a \in \mathbb{R}^d$ ,  $a^T S a \geq 0$ :

$$\begin{aligned}
a^T S a &= a^T \left( \sum_{k=1}^n (x_k - m)(x_k - m)^T \right) a \\
&= \sum_{k=1}^n a^T (x_k - m)(x_k - m)^T a \\
&= \sum_{k=1}^n ((x_k - m)^T a)^T ((x_k - m)^T a) \\
&= \sum_{k=1}^n ((x_k - m)^T a)^2
\end{aligned}$$

Since  $((x_k - m)^T a)^2 \geq 0$  it holds that  $a^T S a \geq 0$ .

3. Thus  $S$  can be decomposed into  $Q\Lambda Q^T$  where  $\Lambda$  is a diagonal with the eigenvalues of  $S$  as values.

4.  $\text{tr}(AQA^T) = \text{tr}(Q)$  when  $A$  is orthogonal  
 $\text{tr}(AQA^T) = \text{tr}((AQ)A^T) = \text{tr}(A^T(AQ)) = \text{tr}(Q)$

With 3. and 4. we get:

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

The trace of  $S$  is therefore the sum of all its eigenvalues, which is trivially an upper-bound for the largest eigenvalue  $\lambda_1$ , if we consider fact 2: Since  $S$  is positive semidefinite, all of its eigenvalues are positive. So the largest eigenvalue  $\lambda_1$  can't be larger than  $\text{tr}(S) = \sum_{i=1}^d S_{ii}$ .

(b) As the upper bound is the sum of all eigenvalues, it is tighter the smaller the eigenvalues  $\lambda_2 \dots \lambda_d$  relatively to the eigenvalue  $\lambda_1$  are. Especially it is the same if all other eigenvalues are zero. This means for the data, that only one feature is really significant for the variance of the data. If all entries for the other features are identical among the different data points, the bound will be tight.

(c) Let's consider the data's covariance matrix, which is given as:

$$\text{Cov} = \frac{1}{n} \sum_{k=1}^n (x_k - m)(x_k - m)^T$$

It's obvious that the scatter matrix is a scaled (by  $n$ ) version of the covariance matrix:  $n \cdot \text{Cov} = n \cdot \frac{1}{n} \sum_{k=1}^n (x_k - m)(x_k - m)^T = \sum_{k=1}^n (x_k - m)(x_k - m)^T = S$ .

The same facts that hold for  $S$  as in 2 a) also hold for  $\text{Cov}$ : it is symmetric and positive semidefinite and thus decomposable as shown above.

It is known that the diagonal elements of the covariance-matrix are equal to the variance of the respective dimension of our data set. Thus the diagonal elements of  $S$  are the variance of the particular dimension scaled by  $n$ .

The Eigenvector of the largest eigenvalue of  $\text{Cov}$  points in the direction of the biggest variance of our data and the corresponding eigenvalue  $\lambda_{\max}$  is equal to the variance of the data in this direction.

We are now using the eigenvalue-decomposition of both matrices to show that the eigenvalues of  $S$  are the ones of the covariance-matrix scaled by  $n$ :

$$S = n \cdot \text{Cov} = n \cdot Q\Lambda Q^T = Q(n \cdot \Lambda)Q^T = Q\bar{\Lambda}Q^T$$

So now we know that the largest eigenvalue  $\lambda_1$  of  $S$  is equal to the largest eigenvalue  $\lambda_{\max}$  of  $\text{Cov}$  scaled by  $n$ :  $\lambda_1 = n \cdot \lambda_{\max}$ .

Let  $x_k^i, m^i$  be the  $i$ 'th entry of vector  $x_k, m$ . We can show with proof by contradiction, that  $\lambda_1 \geq \max_{i=1}^d S_{ii}$ . To show that, we assume  $\max_{i=1}^d S_{ii} = S_{tt}$  and  $\lambda_1 < S_{tt}$ :

$$\begin{aligned} \lambda_1 &< S_{tt} \\ \Leftrightarrow \lambda_1 &< \sum_{k=1}^n (x_k^t - m^t)(x_k^t - m^t) \\ \Leftrightarrow n \cdot \lambda_{\max} &< \sum_{k=1}^n (x_k^t - m^t)(x_k^t - m^t) \\ \Leftrightarrow \lambda_{\max} &< \frac{1}{n} \sum_{k=1}^n (x_k^t - m^t)^2 \\ \Leftrightarrow \lambda_{\max} &< E\{(x_k^t - m^t)^2\} \\ \Leftrightarrow \lambda_{\max} &< \text{Var}(x_k^t) \end{aligned}$$

This leads to a contradiction, since it means that the largest eigenvalue  $\lambda_{max}$  corresponding to the largest variance in our data is supposed to be smaller than the variance of our data in dimension  $t$ .

So we can conclude  $\lambda_{max} \geq \text{Var}(x_k^t)$  and thus also  $n \cdot \lambda_{max} = \lambda_1 \geq n \cdot \text{Var}(x_k^t) = S_{tt} = \max_{i=1}^d S_{ii}$ .

- (d) The bound will be tight if the maximum variance in the data is aligned along one particular dimension. An example for such a case would be the same data set as in 2 c): if the entries of each dimension except for one are identical for all data points, the bound will be tight.

### Exercise 3

- (a)
- (b)
- (c) (\*bonus)