Machine Learning 1
Exercise 4

Group: BSSBCH

November 13, 2017

Matthias Bigalke, 339547, maku@win.tu-berlin.de
Tolga Buz, 346836, buz_tolga@yahoo.de
Alejandro Hernandez, 395678, alejandrohernandezmunuera@gmail.com
Aitor Palacios Cuesta, 396276, aitor.palacioscuesta@campus.tu-berlin.de
Christof Schubert, 344450, christof.schubert@campus.tu-berlin.de
Daniel Steinhaus, 342563, dany.steinhaus@googlemail.com

1 Lagrange Multipliers

(a)

Find the parameter θ that minimizes $J(\theta)$ subject to the constraint $\theta^T b = 0$.

We have

$$\mathcal{L} = \sum_{k=1}^{n} ||\boldsymbol{\theta} - \boldsymbol{x}_k||^2 + \lambda \boldsymbol{\theta}^T \boldsymbol{b}$$

From this we get

$$\begin{aligned} & 0 = \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L} && (1) \\ \Leftrightarrow \quad & 0 = \sum_{k=1}^{n} 2(\boldsymbol{\theta} - \boldsymbol{x}_k) + \lambda \boldsymbol{b} && (2) \\ \Leftrightarrow \quad & 0 = \boldsymbol{\theta} - \overline{\boldsymbol{x}} + \frac{\lambda}{2n} \boldsymbol{b} && (3) \\ \Leftrightarrow \quad & \boldsymbol{\theta} = \overline{\boldsymbol{x}} - \frac{\lambda}{2n} \boldsymbol{b} && (4) \end{aligned}$$

and

$$
\begin{aligned}
& 0 = \frac{\partial}{\partial \boldsymbol{\lambda}} \mathcal{L} \tag{5}\\
\Leftrightarrow \quad & 0 = \boldsymbol{\theta}^T \boldsymbol{b} \tag{6}\\
\overset{(4)}{\Leftrightarrow} \quad & 0 = (\overline{\boldsymbol{x}} - \frac{\lambda}{2n}\boldsymbol{b})^T \boldsymbol{b} \tag{7}\\
\Leftrightarrow \quad & 0 = \overline{\boldsymbol{x}}^T \boldsymbol{b} - \frac{\lambda}{2n}\boldsymbol{b}^T \boldsymbol{b} \tag{8}\\
\overset{\boldsymbol{b} \neq \boldsymbol{0}}{\Leftrightarrow} \quad & \lambda = 2n \frac{\overline{\boldsymbol{x}}^T \boldsymbol{b}}{\boldsymbol{b}^T \boldsymbol{b}} \tag{9}\\
& \tag{10}
\end{aligned}
$$

With (4) and (10) we finally get

$$\boldsymbol{\theta} = \overline{\boldsymbol{x}} - \frac{\overline{\boldsymbol{x}}^T \boldsymbol{b}}{\boldsymbol{b}^T \boldsymbol{b}} \boldsymbol{b} = \overline{\boldsymbol{x}} - \frac{\overline{\boldsymbol{x}}^T \boldsymbol{b}}{||\boldsymbol{b}||} \cdot \frac{\boldsymbol{b}}{||\boldsymbol{b}||}$$

Geometrical interpretation

$\frac{\overline{\boldsymbol{x}}^T \boldsymbol{b}}{||\boldsymbol{b}||}$ discribes a projection from $\overline{\boldsymbol{x}}$ on $\boldsymbol{b}$. $\frac{\overline{\boldsymbol{x}}^T \boldsymbol{b}}{||\boldsymbol{b}||} \cdot \frac{\boldsymbol{b}}{||\boldsymbol{b}||}$ gives us the projection point. So this constraint gives us a minimum at the shifted emperical mean in opposite direction of $\boldsymbol{b}$ by the value of the projection.

(b)

Find the parameter θ that minimizes $J(\theta)$ subject to the constraint $||\boldsymbol{\theta} - \boldsymbol{c}||^2 = 1$.

We have

$$\mathcal{L} = \sum_{k=1}^{n} ||\boldsymbol{\theta} - \boldsymbol{x}_k||^2 + \lambda ||\boldsymbol{\theta} - \boldsymbol{c}||^2 - \lambda$$

From this we get

$$
\begin{aligned}
& 0 = \frac{\partial}{\partial \theta} \mathcal{L} \tag{11}\\
\Leftrightarrow \quad & 0 = \sum_{k=1}^{n} 2\boldsymbol{\theta} - 2\boldsymbol{x}_k + 2\lambda(\boldsymbol{\theta} - \boldsymbol{c}) \tag{12}\\
\Leftrightarrow \quad & 0 = n\boldsymbol{\theta} - \sum_{k=1}^{n} \boldsymbol{x}_k + \lambda\boldsymbol{\theta} - \lambda\boldsymbol{c} \tag{13}\\
\Leftrightarrow \quad & 0 = \boldsymbol{\theta} - \overline{\boldsymbol{x}} + \frac{\lambda}{n}\boldsymbol{\theta} - \frac{\lambda}{n}\boldsymbol{c} \tag{14}\\
\Leftrightarrow \quad & \frac{n+\lambda}{n}\boldsymbol{\theta} = \overline{\boldsymbol{x}} + \frac{\lambda}{n}\boldsymbol{c} \tag{15}\\
\Leftrightarrow \quad & \boldsymbol{\theta} = \frac{n}{n+\lambda}\overline{\boldsymbol{x}} + \frac{\lambda}{n+\lambda}\boldsymbol{c} \tag{16}
\end{aligned}
$$

and

$$
\begin{aligned}
& 0 = \frac{\partial}{\partial \lambda}\mathcal{L} && (17)\\
\Leftrightarrow\quad & 0 = ||\boldsymbol{\theta} - \boldsymbol{c}||^2 - 1 && (18)\\
\Leftrightarrow\quad & 0 = ||\boldsymbol{\theta} - \boldsymbol{c}||^2 - 1 \qquad \text{eq (16)} && (19)\\
\Leftrightarrow\quad & 0 = ||\frac{n}{n+\lambda}\overline{\boldsymbol{x}} + \frac{\lambda}{n+\lambda}\boldsymbol{c} - \boldsymbol{c}||^2 - 1 && (20)\\
\Leftrightarrow\quad & 0 = ||\frac{n}{n+\lambda}\overline{\boldsymbol{x}} - \frac{n}{n+\lambda}\boldsymbol{c}||^2 - 1 && (21)\\
\Leftrightarrow\quad & 1 = \frac{n^2}{(n+\lambda})^2||\overline{\boldsymbol{x}} - \boldsymbol{c}||^2 && (22)\\
\Leftrightarrow\quad & (n+\lambda)^2 = n^2||\overline{\boldsymbol{x}} - \boldsymbol{c}||^2 && (23)\\
\Rightarrow\quad & \lambda = n \cdot (||\overline{\boldsymbol{x}} - \boldsymbol{c}|| - 1) && (24)
\end{aligned}
$$

With Eq (16) and Eq (24) we get

$$
\begin{aligned}
\boldsymbol{\theta} &= \frac{n}{n + n||\overline{\boldsymbol{x}} - \boldsymbol{c}|| - n}\overline{\boldsymbol{x}} + \frac{n||\overline{\boldsymbol{x}} - \boldsymbol{c}|| - n}{n + n||\overline{\boldsymbol{x}} - \boldsymbol{c}|| - n}\boldsymbol{c}\\
&= \frac{1}{||\overline{\boldsymbol{x}} - \boldsymbol{c}||}\overline{\boldsymbol{x}} + \frac{||\overline{\boldsymbol{x}} - \boldsymbol{c}|| - 1}{||\overline{\boldsymbol{x}} - \boldsymbol{c}||}\boldsymbol{c}\\
&= \frac{1}{||\overline{\boldsymbol{x}} - \boldsymbol{c}||}\overline{\boldsymbol{x}} - \frac{1}{||\overline{\boldsymbol{x}} - \boldsymbol{c}||}\boldsymbol{c} + c\\
&= \frac{1}{||\overline{\boldsymbol{x}} - \boldsymbol{c}||}(\overline{\boldsymbol{x}} - \boldsymbol{c}) + c
\end{aligned}
$$

Geometrical interpretation

The derived solution is as expected dependent on the vector c. In addition to the first case, c has now not only additive impact but also anti-proportional, multiplicative impact on $\overline{x}$. The presented solution is the optimum of J considering the constraint g. At this point the contour line of J is tangential to the line describing g.

2 Bounds on Eigenvalues

(a)

Show that $\sum_{i=1}^{d} S_{ii} \geq \lambda_1$.

Assumptions

Since S is a scatter matrix we can infer:

(I) S is symmetrical

(II) S is positive semi definite

(III) The trace elements are positive, because we are using a correlation matrix

Proof

We have

$$\begin{aligned} \lambda_1 \leq \sum_{i=1}^{d} S_{ii} = Tr(S) \overset{(I)}{=} \sum_{i=1}^{d} \lambda_i \\ \Leftrightarrow \quad 0 \leq \sum_{i=2}^{d} \lambda_i \end{aligned}$$

which is fullfilled because with (II) we get $\forall i \in \{1, ..., d\} : \lambda_i \geq 0$.

(b)

The bound becomes tight if the sum $\sum_{i=2}^{d} \lambda_i$ vanishes. This is the case if the data can be explained solely in terms of the first PC.

(c)

In PCA, we maximize $\omega^T S \omega$ subject to the unit vector constraint. If we assume that the eigenvector ω_1 is filled with zeroes at position k (with k corresponding to the elementary feature with largest variance), the first eigenvalue is then given by $\lambda_1 = 1 \cdot var(X_k) = S_{kk} = max_{i=1}^{d} SS_{ii}$ (tight case). By using this method, if we get any other ω_1 that combines multiple dimensions, $\lambda_1 \geq var(X_k) \geq S_{kk} = max_{i=1}^{d} SS_{ii}$ will hold instead.

(d)

If the matrix is diagonal, this means that the dimensions are uncorrelated. Then the eigenvalues have to be the values in the diagonal of the matrix. The maximum of these values is at the same time the lower bound.

3 Iterative PCA

(a)

$$\frac{\partial J(v)}{\partial v} = \frac{\partial J(\omega)}{\partial \omega} \cdot \frac{\partial \omega}{\partial v}$$

$$\frac{\partial J(\omega)}{\partial \omega} = \frac{S^2 \omega}{||S\omega||} - S\omega$$

$$\omega = S^{-\frac{1}{2}} v$$

$$\frac{\partial \omega}{\partial v} = S^{-\frac{1}{2}}$$

$$\frac{\partial J(v)}{\partial v} = S^{-\frac{1}{2}}(\frac{S^2\omega}{||S\omega||} - S\omega)$$

Inserting the partial derivative of J (in respect to v) into the iteration step equation, we can write:

$$S^{\frac{1}{2}}wt + 1 = S^{\frac{1}{2}}w_t + \gamma(S^{-\frac{1}{2}}(\frac{S^2w}{||Sw||} - Sw)$$

$$w_{t+1} = w_t + \gamma(S^{-\frac{1}{2}}(\frac{S^2w}{||Sw||} - Sw)$$

Setting $\gamma = 1$ yields: $w_{t+1} = \frac{Sw_t}{||Sw_t||}$

(b)

We first set the cost function's gradient equal to zero and try to find the $w*$ that satisfies that equation:

$$\frac{\partial J(w)}{\partial w} = \frac{S^2w}{||Sw||} - Sw = 0$$

$$\frac{Sw}{||Sw||} = w$$

$$Sw = ||Sw||w$$

As $||Sw||$ is a scaling factor of the vector w. we can apply the Euclidian norm on both sides and keep $||Sw||$ as a multiplicative factor:
$||Sw|| = ||Sw||\ ||w||$

Obviously, this will only hold if $||w|| = 1$.

4 Programming

See next page.