Technische Universität Berlin
Institut für Softwaretechnik
und theoretische Informatik

Wintersemester 2019/2020
Prof. Dr. Klaus-Robert Müller

Machine Learning I - Homework III

Jacky 391049, Viktor 392636, Duc 395220, Laura 391342, Laura 392032

---

1. Let $(x_k)_{k=1}^n \subset \mathbb{R}^d$ be a data set of n samples. We consider the objective (??) function

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

to be minimized with respect to the parameter $\theta \in \mathbb{R}^d$. It can shown that in absence of constraints for θ, the θ^* that minimizes this objective is given by the empirical mean $\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$.

(a) Using the method of LAGRANGE multipliers, find the parameter θ that minimizes $J(\theta)$ subject to the constraint $\theta^T b = 0$, where $b \in \mathbb{R}^d$. Give a geometrical interpretation to your solution.

We define the LAGRANGIAN (use $y \coloneqq \theta$ for convenience)

$$\mathcal{L}(y, \lambda) = \sum_{k=1}^{n} \|y - x_k\|^2 - \lambda y^T b$$

and compute its gradient:

$$\nabla\mathcal{L}(y, \lambda) = \begin{pmatrix} \sum_{k=1}^{n} 2(y - x_k) - \lambda b \\ y^T b \end{pmatrix}.$$

Setting this equal to zero yields

$$\sum_{k=1}^{n} 2(y - x_k) = \lambda b \iff 2ny - 2\sum_{k=1}^{n} x_k = \lambda b \implies y = \frac{1}{n}\sum_{k=1}^{n} x_k + \frac{\lambda}{2n} \cdot b$$

Calculating $y^T b$ yields $\frac{1}{n}\sum_{k=1}^{n} x_k^T b + \frac{\lambda}{2n} b^T b$, setting zero yields $\lambda = -\frac{2}{\|b\|^2}\sum_{k=1}^{n} x_k^T b$. Plugging this in yields

$$y = \frac{1}{n}\sum_{k=1}^{n} x_k - \frac{1}{\|b\|^2 n}\sum_{k=1}^{n} x_k^T b \cdot b$$

This solution to the minimization problem is the projection of the minimum of J on the hyperplane corresponding to b

(b) Using the same method, find the parameter θ that minimizes $J(\theta)$ subject to $\|\theta - c\|^2 = 1$, where $c \in \mathbb{R}^d$. Give a geometrical interpretation to your solution.

Again we will use LAGRANGE multipliers to minimize $J(\theta)$ subject to the constrain $||\theta - c||^2 = 1$. The LAGRANGIAN in this case is (setting $\theta = y$)

$$\mathcal{L}(y, \lambda) = \sum_{k=1}^{n} ||y - x_k||^2 - \lambda(||y - c||^2 - 1)$$

Then we obtain

$$\nabla\mathcal{L}(y, \lambda) = \begin{pmatrix} 2(ny - \sum_{k=1}^{n} x_k - \lambda(y - c)) \\ ||y - c||^2 - 1 \end{pmatrix}$$

Setting this zero yields:

$$\begin{aligned} 0 &= 2(ny - \sum_{k=1}^{n} x_k - \lambda(y - c)) \\ \Leftrightarrow 0 &= ny - \sum_{k=1}^{n} x_k - \lambda(y - c) \\ \Leftrightarrow y &= \frac{\sum_{k=1}^{n} x_k - \lambda c}{(n - \lambda)} \end{aligned}$$

We further observe

$$1 = ||y - c|| = ||\frac{\sum_{k=1}^{n} x_k - \lambda c}{(n - \lambda)} - c|| = ||\frac{\sum_{k=1}^{n} x_k - \lambda c - (n - \lambda)c}{(n - \lambda)}|| = \frac{1}{n - \lambda}||\sum_{k=1}^{n} x_k - cn||$$

And therefore $\lambda = -||\sum_{k=1}^{n} x_k - cn|| + n$ and $y = \frac{\sum_{k=1}^{n} x_k - \lambda c}{(||\sum_{k=1}^{n} x_k - cn||)}$. This solution is the projection of the minimizer of J on the closed unit ball around c.

2. We consider a data set $(x_k)_{k=1}^{n} \subset \mathbb{R}^d$. The empirical mean m and the scatter matrix S are given by

$$m = \frac{1}{n}\sum_{k=1}^{n} x_k \quad \text{and} \quad S = \sum_{k=1}^{n}(x_k - m)(x_k - m)^T.$$

Let λ_1 be the largest eigenvalue of the matrix S. It quantifies the amount of variation in the data on the first principal component. Because computation of the full scatter matrix and respective eigenvalues can be slow, it can be useful to relate them to the diagonal elements of the scatter matrix $\{S_{ii}\}$ than can be computed in linear time.

(a) Show that $\sum_{k=1}^{d} S_{ii}$ is an upper bound to the eigenvalue λ_1.

Answer: As S is symmetric, there exists an eigendecomposition of $S = Q\Lambda Q^T$, where Λ is a diagonal matrix containing the eigenvalues and Q is orthogonal.

The expression $\sum_{k=1}^{d} S_{ii}$ is also called the trace of the matrix, $\mathrm{tr}(S)$. Similar matrices (A and B are similar if there exists a $P \in \mathrm{GL} : A = P^{-1}BP$) have the same trace, i.e. $\mathrm{tr}(\Lambda) = \mathrm{tr}(S)$. It is well know that $\mathrm{tr}(A) = \sum_{k=1}^{n} n_k\lambda_k$ for any matrix A, where λ_k are the eigenvalues of A with algebraic multiplicities $n_k \in \mathbb{N}_{\geq 1}$. Therefore,

$$\lambda_1 \leq n_1\lambda_1 \leq \mathrm{tr}(\Lambda) = \mathrm{tr}(S).$$

(b) State the conditions on the data for which the upper bound is tight.

Answer: If λ_1 is the only non-zero eigenvalue with algebraic multiplicity 1, the answer for (a) immediately yields that the bound is tight. Examples for such matrices $A \in \mathbb{R}^{n\times n}$ must have the characteristic polynomial $p_A(\lambda) = \lambda^{n-1}(\lambda - \lambda_1)$ for some $n \in \mathbb{N}_{\geq 1}$. All such matrices represent projections into a one-dimensional subspace.

(c) Show that $\max_{i=1}^{d} S_{ii}$ is a lower bound to λ_1.

Proof. It is well known that $\lambda_1 = \max_{||x||=1} xS^Tx$. Using this fact one obtains

$$\lambda_1 = \max_{||x||=1} xS^Tx \geq e_iS^Te_i = S_{ii} \quad \forall i \in \{1, ..., d\}$$

where e_i is the i-th unit vector. Thus, it follows immediately that $\max_{i=1}^{d} S_{ii} \leq \lambda_i$. □

(d) State the conditions on the data for which the lower bound is tight.

Answer: Equality holds whenever there exists at least one $i \in \{1, ..., d\}$ such that e_i is an associated eigenvector for the largest eigenwert. To see this, assume e_i is an eigenvector for the largest eigenwert of the scatter matrix S, i.e.

$$Se_i = \lambda_1 e_i = S_{ii}e_i.$$

Therefore, $\lambda_1 = \max_{i=1}^{d} S_{ii}$.

3. When performing principal component analysis, computing the full eigendecomposition of the scatter matrix S is typically slow, and we are often only interested in the few first principal components. An efficient procedure to find the first eigenvector is the power iteration method, which starts with a random vector $w \in \mathbb{R}^d$, and iterative applies the parameter update $w \leftarrow \frac{Sw}{\|Sw\|}$ until some convergence criterion is met.

(a) Show that application of the power iteration method is equivalent to defining the unconstrained objective

$$J(w) = \|Sw\| - \frac{1}{2}w^TSw$$

and performing the gradient ascent $v \leftarrow v + \gamma\frac{\partial I}{\partial v}$, where $v = S^{0.5}w$ is a reparametrization of v, for some learning γ. We assume that the matrix S is invertible.

Answer: First we substitute $v = S^{0.5}w$ in J and use that S is symmetric:

$$J(w) = ||S^{0.5}(S^{0.5}w)|| - w^TS^{0.5}S^{0.5}w \Rightarrow J(v) = ||S^{0.5}v|| - \frac{1}{2}v^Tv$$

Then we obtain

$$\frac{\partial J}{\partial v} = \frac{\partial}{\partial v}(||S^{0.5}v|| - \frac{1}{2}v^Tv) = \frac{Sv}{||S^{0.5}v||} - v$$

Choosing the learning rate $\gamma = 1$ yields

$$v \leftarrow v + \gamma(\frac{Sv}{||S^{0.5}v||} - v)$$
$$S^{0.5}w \leftarrow S^{0.5}w + (\frac{SS^{0.5}w}{||S^{0.5}S^{0.5}w||} - S^{0.5}w)$$

Since S is invertible we obtain

$$w \leftarrow w + (\frac{Sw}{||S^{0.5}S^{0.5}w||} - w) = \frac{Sw}{||Sw||}$$

We observe that this is equivalent to the power iteration

(b) Show that a necessary condition for w to maximize the objective $J(w)$ is to be a unit vector (i.e. $\|w\| = 1$).

Proof. First, the gradient of J is given by

$$\nabla J(w) = \frac{SSw}{||Sw||} - Sw.$$

Setting the gradient of J to zero yields

$$Sw = \frac{1}{||Sw||}SSw \implies w = \frac{1}{||Sw||}Sw,$$

which shows that w must be a unit vector. Note that S^{-1} exists, otherwise the argument does not work. □