Exercise Sheet 3. Machine learning

1. (a) find θ that minimizes $J(\theta)=\sum_{k=1}^{n}\|\theta-x_k\|^2$, subject to $\theta^Tb=0$

by Lagrange multiplier.

let $L := J(\theta)+\lambda\theta^Tb$

$$=\sum_{k=1}^{n}\|\theta-x_k\|^2+\lambda\theta^Tb$$

$$=\sum_{k=1}^{n}(\theta\theta^T-2\theta^Tx_k+x_k^Tx_k)+\lambda\theta^Tb$$

$$\frac{\partial L}{\partial\lambda}=\theta^Tb=0 \quad \text{— ①}$$

$$\frac{\partial L}{\partial\theta}=\sum_{k=1}^{n}(2\theta-2x_k)+\lambda b=0 \quad \text{— ②}$$

multiply ② by θ^T: $\sum_{k=1}^{n}(2\theta^T\theta-2\theta^Tx_k)+\lambda\theta^Tb=0$

by ①: $\theta^Tb=0$.

so: $\sum_{k=1}^{n}(2\theta^T\theta-2\theta^Tx_k)=0$, $\Rightarrow$ $\theta^T\theta=\theta^T\cdot\frac{1}{n}\sum_{k=1}^{n}x_k$

In order to make sure $J(\theta)$ gets its minimization at θ, we need to make sure the second derivation $\frac{\partial^2 L}{\partial\theta^2}>0$:

$$\frac{\partial^2L}{\partial\theta^2}=\frac{\partial}{\partial\theta}\sum_{k=1}^{n}(2\theta-2x_k)+\lambda b$$

$$=\sum_{k=1}^{n}2 = 2n>0.$$

so: the parameter θ that minimizes $J(\theta)$ should be any parameter $\in R^d$ that satisfies: $\begin{cases}\theta^Tb=0\\ \theta^T\theta=\theta^T\frac{1}{n}\sum_{k=1}^{n}x_k.\end{cases}$ //.

(b) find θ that minimizes $J(\theta)=\sum_{k=1}^{n}\|\theta-x_k\|^2$ subject to $\|\theta-c\|^2=1$.

let $L := J(\theta)+\lambda(\|\theta-c\|^2-1)$

$$=\sum_{k=1}^{n}(\theta\theta^T-2\theta^Tx_k+x_k^Tx_k)+\lambda(\theta^T\theta-2\theta^Tc+c^Tc)-\lambda$$

$$\frac{\partial L}{\partial\lambda}=\|\theta-c\|^2-1=(\theta-c)(\theta-c)^T-1=0 \quad \text{— ①}$$

$$\frac{\partial L}{\partial\theta}=\sum_{k=1}^{n}(2\theta-2x_k)+\lambda(2\theta-2c)=0 \quad \text{— ②}$$

multiply ② by $(\theta-c)^T$: $\sum_{k=1}^{n}(2\theta-2x_k)(\theta-c)^T+\lambda(2\theta-2c)(\theta-c)^T=0$

by ①: $(\theta-c)(\theta-c)^T=1$.

so. $\sum_{k=1}^{n}(2\theta-2x_k)(\theta-c)^T+2\lambda=0$ $\Rightarrow$ $(n\theta-\sum_{k=1}^{n}x_k)(\theta-c)^T+\lambda=0$

Similiar, we need to make sure $\frac{\partial^2L}{\partial\theta^2}>0$:

$$\frac{\partial^2L}{\partial\theta^2}=\frac{\partial}{\partial\theta}\sum_{k=1}^{n}(2\theta-2x_k)+\lambda(2\theta-2c)=2n+2\lambda>0. \quad n>\lambda.$$

so: the parameter θ that minimizes $J(\theta)$ subject to $||\theta - c||^2 = 1$ should be any parameter that satisfies: $\begin{cases} ||\theta - c||^2 = 1 \\ (n\theta - \sum_{k=1}^{n} x_k)(\theta - c)^T + \lambda = 0 \\ \lambda < n \end{cases}$ //.

2. (a) by the definition of Trace: $Tr(S) = \sum_{i=1}^{d} S_{ii}$

and: $Tr(S) = \sum_{i=1}^{m} \lambda_i$, where $\lambda_i = eig(S)$

as for all i $(i<m)$: $\lambda_i \geq 0$

so: $\sum_{i=1}^{d} S_{ii} = Tr(S) = \sum_{i=1}^{m} \lambda_i = \lambda_1 + \lambda_2 + \cdots + \lambda_m$.

so: $\lambda_1 \leq \sum_{i=1}^{d} S_{ii}$. //.

(b) when "=" holds, ie: $\lambda_1 = \sum_{i=1}^{d} S_{ii}$.

then: $m=1$. which means S has only one eigenvalue.

The whole database $x_i \in R^d$, x_i can be considered as lying in one-dimension. ie: even though $x_i \in R^d$, they can be transformed into another space which has dimension 1. $\{x_i \in R^1\}$. //.

(c). The reason that $\max_{i=1}^{d} S_{ii}$ is a lower bound of λ_1 is: $S_{11}, S_{22}, \cdots S_{dd}$ can be viewed as the variance of original data x, which measure the component of each data x_i in X.

$\lambda_1, \lambda_2, \cdots \lambda_m$ is the eigenvalue of S. by the idea of PCA, λ is also the component of the reshaped data in X.

PCA extracts the most principle component of a database. so λ_1, which is the largest eigenvalue, should be larger than any S_{ii}. $(i \leq d)$

ie: $\lambda_1 \geq \max_{i=1}^{d} S_{ii}$. //.

(d) when "=" holds. ie: $\lambda_1 = \max_{i=1}^{d} S_{ii}$.

The reshaped data by PCA is of no difference to the original data. Or the data can not be transformed to anothe space R^t, where $t<d$. //.

3. (a). by $v = S^{0.5} w$: $v^T v = (S^{0.5} w)^T (S^{0.5} w) = w^T S w$.

since S is invertible: $S^{-0.5} v = S^{-0.5} S^{0.5} w$. $\Rightarrow$ $w = S^{-0.5} v$.

so: $J(w) = ||Sw|| - \frac{1}{2} w^T S w$

$= ||S \cdot S^{-0.5} v|| - \frac{1}{2} w^T S w$

$= ||S^{0.5} v|| - \frac{1}{2} v^T v$

$\frac{\partial J}{\partial v} = ||S^{0.5}|| - \frac{1}{2} v$.

$\cancel{J(v) = x\theta S}$.

so: $V \leftarrow V + y\frac{\partial J}{\partial V}$.

$V \leftarrow V + y\left(\|S^{0.5}\| - \frac{1}{2}V\right)$. $\quad V \leftarrow \left(1 + y\|S^{0.5}\|V^{-1} - \frac{1}{2}y\right)V$.

by: $W \leftarrow \frac{SW}{\|SW\|}$ and $W = S^{-0.5}V$:

$$S^{-0.5}V \leftarrow \frac{S \cdot S^{-0.5}V}{\|S \cdot S^{-0.5}V\|}$$

$$S^{0.5}S^{-0.5}V \leftarrow \frac{S^{0.5}S \cdot S^{-0.5}V}{\|S^{0.5}V\|}$$

so: $V \leftarrow \frac{SV}{\|S^{0.5}V\|}$

i.e.: $1 + y\|S^{0.5}\|V^{-1} - \frac{1}{2}y \longrightarrow \frac{S}{\|S^{0.5}V\|}$ ——— ③

But here we cannot find a proper way to prove ③.

(b)