Machine Learning 1 exercises 3

Jiayun's version

February 25, 2021

1 Lagrange Multipliers

1.a Find θ with constraint: $\theta^T b = 0$, and give a geometrical interpretation.

First we write down the Lagrange function,

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

$$\mathcal{L}(\theta, \lambda) = n\theta^T\theta - \sum_{k=1}^{N} \theta^T x_k - \sum_{k=1}^{N} x_k^T\theta + \sum_{k=1}^{N} x_k^T x_k + \lambda(\theta^T b)$$

Apply KKT condition,

$$\nabla_\theta \mathcal{L} = 2n\theta - \sum_{k=1}^{N} x_k - \sum_{k=1}^{N} x_k + \lambda b \doteq 0$$

$$\nabla_\lambda \mathcal{L} = \theta^T b \doteq 0$$

$$\theta^T b = 0$$

From first equation we could get $\theta = \bar{x} - \frac{\lambda}{2n} b$, we denoted m by $\bar{x}$, because it means clearer than m. combined with the second and third equation we could get $\theta^T b = (\bar{x} - \frac{\lambda}{2n} b)^T b = 0$, in which $\lambda = 2n\frac{\bar{x}b}{b^T b}$. With b is unit vector, then we know $\hat{\theta} = \bar{x} - \frac{\bar{x}^T b}{b^T b} b = \bar{x} - \bar{x}^T b b$.
From geometrical aspect, the second term of the result is actually the projection vector of $\bar{x}$ to b, so the $\hat{\theta}$ is comprised from data mean minus projection of data mean to b, therefore it's orthogonal to b as is given by the constraint.

1.b minimizes $J(\theta)$ subject to $\|\theta - c\|^2 = 1$, where c is a vector in R^d different from m ($\bar{x}$).

we write out the Lagrange function,

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

get the KKT conditions as below:

$$\nabla_\theta \mathcal{L} = \sum_{k=1}^{N} (2\theta - 2x_k) + \lambda(2\theta - 2c) \doteq 0$$

$$\nabla_\lambda \mathcal{L} = \lambda(\|\theta - c\|^2 - 1) \doteq 0$$

from the first equation we could calculate out: $\theta = \frac{2\sum_{k=1}^{N} x_k + 2\lambda c}{2(n+\lambda)} = \frac{n\bar{x} + \lambda c}{n+\lambda}$, if we use the notation of the mean. Then it follows:

$$\theta = \frac{n}{n+\lambda}\bar{x} + \frac{\lambda}{n+\lambda} c$$

Substitute the θ into the second KKT equation we got,

$$\|\theta - c\|^2 = \left\| \frac{n}{n+\lambda}\bar{x} - \frac{n}{n+\lambda} c \right\|^2 = 1$$

By using the operation laws of norm, we could factor λ out of the norm sign,

$$\left\| \frac{n\bar{x} - nc}{n+\lambda} \right\|^2 = (|\frac{n}{n+\lambda}|)^2 \left\|\bar{x} - c\right\|^2 = 1$$

Then we could get λ, and because of n and value of the norm are positive,

$$\lambda = \sqrt{n^2 \left\|\bar{x} - c\right\|^2} - n = n\left\|\bar{x} - c\right\| - n$$

Return to our representation of θ:

$$\theta = \frac{1}{\|\bar{x} - c\|}\bar{x} + (1 - \frac{1}{\|\bar{x} - c\|})c$$

Geometrical interpretation: because of value of norm is positive, and the factors of two vectors summing to $\frac{1}{\|\bar{x}-c\|} + (1 - \frac{1}{\|\bar{x}-c\|}) = 1$, so the result θ laid on the line (affine space), which built by the two vectors $\bar{x}$ and c

2 Principal Component Analysis

3 Bounds on Eigenvalues

3.a

Because of the property of trace:

$$tr(S) = \sum_{i=1}^{d} S_{ii} = \sum_{i=1}^{d} \lambda_i$$

The matrix S is defined as $S = \sum_{k=1}^{N}(x_k - m)(x_k - m)^T$. For any $\omega \in R^d$ we have:

$$\omega^T S\omega = \sum_{k=1}^{N} \omega^T(x_k - m)(x_k - m)^T\omega = \sum_{k=1}^{N}((x_k - m)^T\omega)^T(x_k - m)^T\omega = \sum_{k=1}^{N} \left\|(x_k - m)^T\omega\right\|^2 \geq 0$$

Thus S is positive semi-definite, so $\lambda_i \geq 0$ for i from 1 to d. Then $\lambda_1 \leq \sum_{i=1}^{d} \lambda_i = \sum_{i=1}^{d} S_{ii}$.

3.b

This holds only when $\lambda_1 = \sum_{k=1}^{d} \lambda_i$, and that indicates $\lambda_i = 0$ for i from 2 to d. And thus the rank of the matrix S is 1, due to the dimension of the kernel is $d - 1$.

3.c

We use spectral decomposition of the matrix S,

$$S = W\Lambda W^T = \begin{bmatrix}\omega_1 & \dots & \omega_d\end{bmatrix} \begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_d\end{bmatrix} \begin{bmatrix}\omega_1 \\ \vdots \\ \omega_d\end{bmatrix} = \sum_{i=1}^{d} \lambda_i\omega_i\omega_i^T$$

Because λ_1 is the maximum eigenvalue with $W^T = W^{-1}$, it's obvious , $S = \sum_{i=1}^{d} \lambda_i\omega_i\omega_i^T$ we can see that in this matrix every entrance is less than or equal to the corresponding entrance in the matrix $\sum_{i=1}^{d} \lambda_1\omega_i\omega_i^T = \lambda_1 WW^T = \lambda_1 I$. This proved $\lambda_1 \geq \max_{i=1}^{d} S_{ii}$, thus λ_1 is lower bounded by $\max_{i=1}^{d} S_{ii}$.

3.d

Following the previous low bound we could state, when all the eigenvalues are the same, then in the equation relaxation step $\sum_{i=1}^{d} \lambda_i\omega_i\omega_i^T \leq \sum_{i=1}^{d} \lambda_1\omega_i\omega_i^T$, the equality can be determined. Say the S has only one eigenvalue with d multiplicity.

4 Iterative PCA

To show the error will decay with exponential rate, firstly we need to derive a relationship between errors of two iteration steps.

$$\varepsilon_k(\omega_T) = |\frac{[S\omega_{T-1}]^T u_k}{[S\omega_{T-1}]^T u_1}|$$

$$\varepsilon_k(\omega_T) = |\frac{\omega_{T-1}^T \sum_{i=1}^d u_i u_i^T \lambda_i u_k}{\omega_{T-1}^T \sum_{i=1}^d u_i u_i^T \lambda_i u_1}| = |\frac{\omega_{T-1}^T S u_k}{\omega_{T-1}^T S u_1}| = |\frac{\omega_{T-1}^T \lambda_k u_k}{\omega_{T-1}^T \lambda_1 u_1}| = |\frac{\lambda_k}{\lambda_1}||\frac{\omega_{T-1}^T u_k}{\omega_{T-1}^T u_1}| = |\frac{\lambda_k}{\lambda_1}|\varepsilon_k(\omega_{T-1})$$

We use time notation as subscript to avoid any vague meaning with transpose. Then from the derivation above, we could recursively build the equation:

$$\varepsilon_k(\omega_T) = |\frac{\lambda_k}{\lambda_1}|^T \varepsilon_k(\omega_0)$$

T is here power of fraction.