Exercise 11 of Machine Learning [IN 2064]

Name: Yiman Li
Matr-Nr: 03724352
cooperate with Kejia Chen(03729686)

Problem 1

After adding the data of new user Leslie, and using the SVD decomposition function in MATLAB, we get that:

$$\boldsymbol{M}' = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{1}$$

namely:

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 3 & 0 & 0 & 4 \end{bmatrix} = \begin{bmatrix} 0.1375 & -0.0170 \\ 0.4126 & -0.0511 \\ 0.5502 & -0.0681 \\ 0.6877 & -0.0852 \\ 0.0417 & 0.5627 \\ 0.0521 & 0.7033 \\ 0.0208 & 0.2813 \\ 0.1739 & 0.3079 \end{bmatrix} \begin{bmatrix} 12.5217 & 0 \\ 0 & 9.9377 \end{bmatrix} \begin{bmatrix} 0.5602 & -0.0874 & -0.2886 & -0.3028 & 0.7096 \\ 0.6019 & 0.0056 & 0.5067 & 0.6172 & -0.0050 \end{bmatrix} \tag{2}$$

Here, $\boldsymbol{U}$ connects people to concepts, and $U(8,1) = 0.1739$ is smaller than some of the other entries in the first column, because Leslie doesn't rate science fiction very high compared to John and Jack. In contrast, Leslie may more likely to see romance movie since $U(8,2) = 0.3079$ is larger than $U(8,1)$.

Problem 2

We already know that

$$p(\boldsymbol{z}) = \quad \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I}) \tag{3}$$

$$p(\boldsymbol{x}|\boldsymbol{z}) = \quad \mathcal{N}(\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \Phi) \tag{4}$$

For the original data, we have the corresponding log likelihood as below:

$$\begin{aligned} \ln p(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{\Phi}) &= \sum_{n=1}^{N} \ln p(\boldsymbol{x}_n|\boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) \\ &= -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{C}| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \mu)^T\boldsymbol{C}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}) \end{aligned} \tag{5}$$

where $\boldsymbol{C}$ is defined by

$$\boldsymbol{C} = \boldsymbol{W}\boldsymbol{W}^T + \boldsymbol{\Phi} \tag{6}$$

- Setting the derivative of the log likelihood with respect to $\boldsymbol{\mu}$ equal to zero gives the expected result $\boldsymbol{\mu} = \bar{\boldsymbol{x}}$ where $\bar{\boldsymbol{x}}$ is the mean value of the data.

- Maximization with respect to $\boldsymbol{W}$ according to [1], we have

$$\boldsymbol{W}_{ML} = \boldsymbol{U}_M(\boldsymbol{L}_M - \boldsymbol{\Phi})^{\frac{1}{2}}\boldsymbol{R} \tag{7}$$

  where $\boldsymbol{U}_M$ is a $D \times M$ matrix whose columns are given by any subset of the eigenvectors, $\boldsymbol{L}_M$ has elements given by the corresponding eigenvalues λ_i, and $\boldsymbol{R}$ is an arbitary $M \times M$ orthogonal matrix.

- Maximization with respect to $\boldsymbol{\Phi}$ according to [1], we have

$$\phi_i = \frac{1}{D-M}\sum_{i=M+1}^{D}\lambda_i \tag{8}$$

  where D is the original data dimension and M is the number of principle eigenvectors.

After doing a linear transformation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, the mean value and the covariance is now $\boldsymbol{A}\boldsymbol{\mu}_{ML}$ and $\boldsymbol{A}\boldsymbol{\Phi}_{ML}\boldsymbol{A}^T$, which is exactly the corresponding maximum likelihood solution for the transformed data set. And when set the derivative with respect to $\boldsymbol{W}$ in the transformed space, since the elements in matrix $\boldsymbol{L}_M$ is scaled by $\boldsymbol{A}\boldsymbol{A}^T$ and then take the power of $\frac{1}{2}$, so the new solution is now $\boldsymbol{A}\boldsymbol{W}_{ML}$.

And since the matrix $\boldsymbol{A}$ is orthogonal, so the quantity $\boldsymbol{W}\boldsymbol{W}^T$ appears in the covariance matrix $\boldsymbol{C}$ takes the form

$$\tilde{\boldsymbol{W}}\tilde{\boldsymbol{W}}^T = \boldsymbol{W}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{W}^T = \boldsymbol{W}\boldsymbol{W}^T \tag{9}$$

and hence is independent of $\boldsymbol{A}$. We can also get this solution by simple matching patterns the MLE solutions.

Now, if $\boldsymbol{A}$ is orthogonal and $\boldsymbol{\Phi}$ a scaled indentity matrix, the model characteristics are also preserved since

$$\boldsymbol{A}\boldsymbol{\Phi}_x\boldsymbol{A}^T = \sigma^2\boldsymbol{I}\boldsymbol{A}\boldsymbol{A}^T = \sigma^2\boldsymbol{I}^2 = \sigma^2\boldsymbol{I} \tag{10}$$

Problem 3

Suppose we make a matrix transformation formulated by $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W} + \boldsymbol{b}$, where $\boldsymbol{W} \in \mathbb{R}^{D\times D}$, and $\boldsymbol{b} \in \mathbb{R}^{N\times D}$, then we have

$$\begin{aligned} Var(\boldsymbol{Y}) &= \boldsymbol{W}^T(\frac{1}{N}\boldsymbol{X}^T\boldsymbol{X} - \bar{\boldsymbol{x}}\boldsymbol{x}^T)\boldsymbol{W} \\ &= \boldsymbol{W}^T(\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T)\boldsymbol{W} \end{aligned} \tag{11}$$

- a) With $\boldsymbol{S}$ being the identity matrix $\boldsymbol{I}$, which means that $\boldsymbol{Y}$ remains the same as $\boldsymbol{X}$ after the transfomation, so there are still 70% of the variance preserved.
- b) With $\boldsymbol{R}$ being the row orthogonal matrix, we have $Var(\boldsymbol{Y}) = \boldsymbol{R}^T(\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T)\boldsymbol{R}$, with $\boldsymbol{\Gamma}' = \boldsymbol{R}^T\boldsymbol{\Gamma}$. Here $\boldsymbol{R}^T\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\boldsymbol{R} = \boldsymbol{I}$, and $\boldsymbol{\Gamma}^T\boldsymbol{R}\boldsymbol{R}^T\boldsymbol{\Gamma} = \boldsymbol{I}$, which means that $\boldsymbol{\Gamma}'$ is still a orthonormal matrix, so there are 70% of the variance being preserved.
- c) With $\boldsymbol{P} = diag(5, -5, \cdots, 5, -5)$, then the new covariance matrix is now scaled by a factor 25, which will not change what the matrix looks like, so it still preserve 70% of the variance.
- d) With $\boldsymbol{Q} = diag(1, 2, 3, \cdots, D-1, D)$, which means the new covariance matrix is now $diag(\lambda_1, 4\lambda_2, 9\lambda_3, \cdots)$, so we have no idea what the order of principle components in new matrix will be, which means the preserved variance can not be told without additional information.
- e) Since adding the same value to the original matrix will not change the covatriance matrix, so we still preserve 70% of original variance.
- f) Since $rank(\boldsymbol{A}) = 5$, which means that the data lies in a 5-dimensional space or subspace, so the top 5 principle components captures all the variance, that is to say 100%.

Problem 4

a) The mean of the N points is shown as below:

$$\bar{\boldsymbol{x}} = \frac{1}{N} \cdot \boldsymbol{X}^T \cdot \mathbf{1}_N = [\ 2 \quad 1 \quad 1\ ] \tag{12}$$

Shift the points by their mean $\bar{\boldsymbol{x}}$, we get

$$\tilde{\boldsymbol{X}} = \boldsymbol{X} - \bar{\boldsymbol{x}} = \begin{bmatrix} 2 & 2 & 1 \\ 0 & 0 & -3 \\ 2 & -2 & 1 \\ -4 & 0 & 1 \end{bmatrix} \tag{13}$$

So the variance is

$$\boldsymbol{\Sigma}_{\boldsymbol{X}} = Var(\tilde{\boldsymbol{X}}) = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 6 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{14}$$

- The first principle component is $[1\ 0\ 0]^T$, with variance value 6;
- The second principle component is $[0\ 0\ 1]^T$, with variance value 3;
- The third principle component is $[0\ 1\ 0]^T$, with variance value 2.

b) According to the function 14, we can see that the top-2 principle conponents are 6 and 3, so we obtain the projected data $\boldsymbol{Y}$ as below:

$$\boldsymbol{Y} = \begin{bmatrix} 2 & 2 & 1 \\ 0 & 0 & -3 \\ 2 & -2 & 1 \\ -4 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 0 & -3 \\ 2 & 1 \\ -4 & 1 \end{bmatrix} \tag{15}$$

c)When new data point have exactly the same PCA perfomance, which means that the covariance remains the same after adding this new data point(scaling factor will not change what the covariance matrix looks like), so we can say that $\boldsymbol{x}_5 = [2\ 1\ 1]$

Problem 5

To be seen in the end.

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.