PCA, K-means, and pLSA (Exercises in MIE 2A)

Ryota Tomioka
tomioka@mist.i.u-tokyo.ac.jp

July 10, 2013

1 Principal Component Analysis

Let $\boldsymbol{x}_i \in \mathbb{R}^D$ $(i = 1, \ldots, N)$ be a collection of data points (e.g., images). The sample covariance matrix $\boldsymbol{\Sigma}$ is defined as

$$\boldsymbol{\Sigma} := \frac{1}{n} \sum_{i=1}^{N} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top,$$

where $\bar{\boldsymbol{x}}$ is the sample mean $\bar{\boldsymbol{x}} := \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i$.

Principal components are defined as the leading eigenvectors of the sample covariance matrix [1, 5], namely

$$\boldsymbol{\Sigma}\boldsymbol{u}_k = \lambda_k \boldsymbol{u}_k \quad (k = 1, \ldots, D),$$

where $\boldsymbol{u}_k$ is called the kth principal component and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$.

Exercise 1. *For any $K = 1, \ldots, D$, prove that $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K] \in \mathbb{R}^{D \times K}$ ($\boldsymbol{u}_k$ are the eigenvectors of $\boldsymbol{\Sigma}$) is the solution of the maximization problem*

$$\begin{aligned} \underset{\boldsymbol{U} \in \mathbb{R}^{D \times K}}{\text{maximize}} \quad & \mathrm{Tr}\left(\boldsymbol{U}^\top \boldsymbol{\Sigma} \boldsymbol{U}\right), \\ \text{subject to} \quad & \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_K. \end{aligned}$$

Exercise 2. *Let $\bar{\boldsymbol{X}}$ be the centered data matrix, namely*

$$\bar{\boldsymbol{X}} := [\boldsymbol{x}_1 - \bar{\boldsymbol{x}}, \boldsymbol{x}_2 - \bar{\boldsymbol{x}}, \ldots, \boldsymbol{x}_N - \bar{\boldsymbol{x}}] \in \mathbb{R}^{D \times N}.$$

Prove that the left singular vectors $\tilde{\boldsymbol{u}}_1, \ldots, \tilde{\boldsymbol{u}}_D$ obtained by the singular value decomposition of $\bar{\boldsymbol{X}}$ coincides with the eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_D$. The singular value decomposition (SVD) of $\bar{\boldsymbol{X}}$ is defined as

$$\bar{\boldsymbol{X}} = \tilde{\boldsymbol{U}} \boldsymbol{S} \boldsymbol{V}^\top,$$

where $\tilde{\boldsymbol{U}}^\top \tilde{\boldsymbol{U}} = \boldsymbol{I}_D$, $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_N$, and $\boldsymbol{S}$ is diagonal.

Exercise 3. *Let $\boldsymbol{V}^\top = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N]$ be the column slices of the matrix $\boldsymbol{V}^\top$ in the above SVD. Prove that the K-dimensional vector obtained by taking the first K coordinates of $\boldsymbol{v}_i$ is* exactly *the projection of the ith data point $\boldsymbol{x}_i$ on the subspace spanned by the first K principal components $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K$ up to scaling.*

2 K-means Clustering

Let $\boldsymbol{x}_i \in \mathbb{R}^D$ $(i = 1, \ldots, N)$ be a collection of data points. K-means clustering [4] is a clustering algorithm that iteratively minimize the K-means objective

$$f((\boldsymbol{\mu}_k)_{k=1}^K) = \sum_{i=1}^N \min_{k=1,\ldots,K} d(\boldsymbol{x}_i, \boldsymbol{\mu}_k),$$

where $\boldsymbol{\mu}_k$ $(k = 1, \ldots, K)$ are the *cluster centers* and $d(\cdot, \cdot)$ is a distance function. Here we take the squared Euclidean distance

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2.$$

The algorithm can be described as follows:

1. Randomly initialize the cluster assignments $Z = (z_{k,i})$ ($z_{k,i} = 1$ if the ith data point belongs to the kth cluster).
2. Iterate until convergence:
   (a) For a fixed cluster assignment, update the cluster centers by minimizing the sum of distances, namely
   $$\boldsymbol{\mu}_k \leftarrow \operatorname*{argmin}_{\boldsymbol{\mu} \in \mathbb{R}^D} \sum_{i: z_{k,i}=1} d(\boldsymbol{x}_i, \boldsymbol{\mu}) \quad (k = 1, \ldots, K). \tag{1}$$
   (b) For fixed cluster centers, assign each data point to the nearest cluster, namely
   $$z_{k,i} \leftarrow \operatorname*{argmin}_k \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2.$$

Exercise 4. *Prove that update* (1) *with the squared Euclidean distance* $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$ *can be written as follows:*

$$\boldsymbol{\mu}_k \leftarrow \frac{\sum_{i=1}^N z_{k,i} \boldsymbol{x}_i}{\sum_{i=1}^N z_{k,i}} \quad (k = 1, \ldots, K).$$

3 Probabilistic Latent Semantic Analysis (pLSA)

Let $\boldsymbol{x}_i \in \mathbb{Z}_+^W$ $(i = 1, \ldots, D)$ be a collection of documents. Each document is represented as *bag-of-words*; that is, $\boldsymbol{x}_i = (x_{ji})$ and x_{ji} is the number of occurrences of the jth word $(j = 1, \ldots, W)$ in the ith document, where W is the number of words (or size of the vocabulary).

In this bag-of-words representation, each document can be considered as a realization from some multinomial distribution. Probabilistic Latent Semantic Analysis (pLSA) [3] assumes that the underlying multinomial distributions are mixtures of topics, which can be considered as stereotypical distributions corresponding to e.g., sports, economics, etc. More precisely, we model the probability of observing the jth word in the ith document as follows:

$$P(\text{word} = j | \text{document} = i) = \sum_{k=1}^K \phi_{jk} \pi_{ki},$$

where ϕ_{jk} and π_{ki} are parameters defined as follows:

$$\phi_{jk} := P(\text{word} = j | \text{topic} = k),$$
$$\pi_{ki} := P(\text{topic} = k | \text{document} = i).$$

Consequently, the log likelihood of the data $\boldsymbol{x}_i$ $(i = 1, \ldots, D)$ can be written as follows:

$$\log P = \sum_{i=1}^{N} \sum_{j=1}^{W} x_{ji} \log \left(\sum\nolimits_{k=1}^{K} \phi_{jk} \pi_{ki} \right) \tag{2}$$

Since it is not easy to directly maximize the log likelihood (2) with respect to ϕ_{jk} and π_{ki}, we employ the expectation-maximization (EM) algorithm [2]. More precisely, we construct a lower bound of the log likelihood as follows (*exercise!*):

$$\log P \geq \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{K} x_{ji} q_{kji} \log \left(\frac{\phi_{jk} \pi_{ki}}{q_{kji}} \right). \tag{3}$$

Since the above inequality is true for any q_{kji} that satisfies $q_{kji} \geq 0$ and $\sum_{k=1}^{K} q_{kji} = 1$, we take q_{kji} that maximizes the right-hand side of the lower bound (3) as follows:

$$q_{kji} \leftarrow \frac{\phi_{jk} \pi_{ki}}{\sum_{k=1}^{K} \phi_{jk} \pi_{ki}}. \tag{4}$$

This is called the E-step. The obtained q_{kji} can be considered as the posterior mean probability of the jth word in the ith document coming from the kth topic.

Next, we maximize the right-hand side of the lower bound (3) for the above q_{kji} as follows:

$$\phi_{jk} \leftarrow \frac{x_{ji} q_{kji}}{\sum_{j=1}^{W} x_{ji} q_{kji}} \tag{5}$$

$$\pi_{ki} \leftarrow \frac{x_{ji} q_{kji}}{\sum_{k=1}^{K} x_{ji} q_{kji}} \tag{6}$$

This is called the M-step. The overall algorithm is to iterate the E- and M-steps until convergence.

Exercise 5. *Prove inequality* (3)*. Hint: Jensen's inequality.*

Exercise 6. *Derive the update equations* (4)*,* (5)*, and* (6)*. Note that the constraints* $\sum_{j=1}^{W} \phi_{jk} = 1$ *and* $\sum_{k=1}^{K} \pi_{ki}$ *must be satisfied.*

4 Assignments (due *July 17th*)

Write up on either one of the following two topics using A4 papers and submit it to the post box of 数理第六研究室. The deadline is Wednesday, July 17th.

1. Solve exercises 1–6.
2. Implement K-means or pLSA. Apply the algorithm to some data set (provided ones or other data set). Empirically analyze and discuss the results.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Stat. Soc. B*, pages 1–38, 1977.

[3] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296, 1999.

[4] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297, 1967.

[5] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.