第九次作业 - Unsupervised Learning 1

ProblemPCA1.docx

Problem I

Problem 1: Given training data $\{x_i\}_{i=1}^N, x_i \in R^p$, $\hat{\mu}$ is the mean of the training data, Σ is its covariance matrix with eigenvectors $\{\mathbf{v}_j\}_{j=1}^p$ and eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_p$. Given an eigenvector $\mathbf{v}_j$, the projection of data to eigenvector $\mathbf{v}_j$ subspace is defined by

$$\{\hat{\beta}_{ji}\}_{i=1}^N = \arg\min_{\beta_{ji}} \sum_{i=1}^N \left\| x_i - \hat{\mu} - \mathbf{v}_j \beta_{ji} \right\|^2 \text{, where } \mathbf{v}_j^T \mathbf{v}_j = 1.$$

1. Derive the solution $\{\hat{\beta}_{ji}\}_{i=1}^N$ to the above optimal problem.
2. Prove that $\sum_{i=1}^N \left\| x_i - \hat{\mu} - \mathbf{v}_k \hat{\beta}_{ki} \right\|^2 < \sum_{i=1}^N \left\| x_i - \hat{\mu} - \mathbf{v}_j \hat{\beta}_{ji} \right\|^2$, if $\lambda_k > \lambda_j$.

1

Assuming that v_j is fixed. We notice that:

$$\{\hat{\beta}_{ji}\}_{i=1}^N = \arg\min_{\beta_{ji}} \sum_{i=1}^N \|x_i - \hat{\mu} - \mathbf{v}_j \beta_{ji}\|^2 = \arg\min_{\beta_{ji}} \sum_{i=1}^N (x_i - \hat{\mu} - v_j \beta_{ji})^\top (x_i - \hat{\mu} - v_j \beta_{ji})$$

We take the derivation of the above equation with respect to β_{ji} , and let it be 0. And we get

$$-(x_i - \hat{\mu})^\top v_j - v_j^\top (x_i - \hat{\mu}) + 2 v_j^\top v_j \beta_{ji} \Rightarrow \beta_{ji} = v_j^\top (x_i - \hat{\mu})$$

2

$$\sum_{i=1}^{N}\|x_i-\hat{\mu}-\mathbf{v}_k\hat{\beta}_{ki}\|^2=\sum_{i=1}^{N}\|x_i-\hat{\mu}-\mathbf{v}_k\mathbf{v}_k^T(x_i-\hat{\mu})\|^2$$

$$\begin{aligned}
&=\sum_{i=1}^{N}\|(\mathbf{I}-\mathbf{v}_k\mathbf{v}_k^T)(x_i-\hat{\mu})\|^2=tr\left((\mathbf{I}-\mathbf{v}_k\mathbf{v}_k^T)\mathbf{X}^T\mathbf{X}(\mathbf{I}-\mathbf{v}_k\mathbf{v}_k^T)^{\mathrm{T}}\right)\\
&=tr\left(\mathbf{X}^T\mathbf{X}(\mathbf{I}-\mathbf{v}_k\mathbf{v}_k^T)\right)=tr(\mathbf{X}^T\mathbf{X})-tr(N\Sigma\mathbf{v}_k\mathbf{v}_k^T)\\
&=tr(\mathbf{X}^T\mathbf{X})-tr(N\lambda_k\mathbf{v}_k\mathbf{v}_k^T)\\
&=tr(\mathbf{X}^T\mathbf{X})-N\lambda_k
\end{aligned}$$

So

$$\sum_{i=1}^{N}\left\|x_i-\hat{\mu}_k-\mathbf{v}_k\hat{\beta}_{ki}\right\|^2=tr(\mathbf{X}^T\mathbf{X})-N\lambda_k<tr(\mathbf{X}^T\mathbf{X})-N\lambda_j=\sum_{i=1}^{N}\left\|x_i-\hat{\mu}_j-\mathbf{v}_j\hat{\beta}_{ji}\right\|^2$$

Problem II

Problem 2: The Non-negative Matrix Factorization $\mathrm{X}=WH$ can be formulated as maximum likelihood of Poisson distribution. Prove that such a formulation is equivalent to minimizing the KL divergence of x_{ij} and $(WH)_{ij}$。

Maximun likelihood of Poisson distribution

Following the poisson distribution, we get:

$$P(x\mid(WH)_{ij})=\frac{(WH)_{ij}^{x}\,e^{-(WH)_{ij}}}{x!}$$

Since we aim to maximize the likelihood if this PDF, we just need to maximize its log probability. that is

$$L(W,H)=\sum_{i=1}^{N}\sum_{j=1}^{p}[x_{ij}\log(WH)_{ij}-(WH)_{ij}]$$

Minimizing the KL divergence

I will use the generalized version of KL divergence in

> Yang, Zhirong, et al. "Kullback-Leibler divergence for nonnegative matrix factorization." *International Conference on Artificial Neural Networks*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

Given a nonnegative input data matrix $\mathbf{X} \in \mathbb{R}_+^{m\times n}$, *Nonnegative Matrix Factorization* (NMF) seeks a decomposition of $\mathbf{X}$ that is of the form $\mathbf{X} \approx \mathbf{WH}$, where $\mathbf{W} \in \mathbb{R}_+^{m\times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r\times n}$ with the rank $r < \min(m,n)$. The matrix $\widehat{\mathbf{X}} = \mathbf{WH}$ is called the unnormalized approximating matrix of $\mathbf{X}$.

In previous work, the approximation has widely been achieved by minimizing one of the two measures: (1) the least square criterion $\varepsilon = \sum_{i,j}(X_{ij} - \widehat{X}_{ij})^2$ and (2) the *generalized Kullback-Leibler divergence* (or *I-divergence*)

$$D_I\left(\mathbf{X}||\widehat{\mathbf{X}}\right) = \sum_{ij}\left(X_{ij}\log\frac{X_{ij}}{\widehat{X}_{ij}} - X_{ij} + \widehat{X}_{ij}\right). \quad (1)$$

So for distribution x and WH

$$\mathbb{D}_{\mathrm{KL}}(x||WH) = \sum_{i=1}^{N}\sum_{j=1}^{p}\left(x_{ij}\log\frac{x_{ij}}{WH_{ij}} - x_{ij} + WH_{ij}\right).$$

Our goal is to find W, H that minimizes the above equation, so we can remove the fixed constant part of it.

Notice that $x_{ij}\log x_{ij}, -x_{ij}$ are both independent of the relevant objective function.

In order to minimize the KL divergence, we only need to maximize the following function

$$L(W,H) = -\left(\sum_{i=1}^{N}\sum_{j=1}^{p}(-x_{ij}\log WH_{ij} + WH_{ij})\right) = \sum_{i=1}^{N}\sum_{j=1}^{p}(x_{ij}\log WH_{ij} - WH_{ij})$$

So we have proven the equivalent of these two methods.