CAP 6610 Machine Learning, Spring 2020

Homework 4

Due 4/15/2020 11:59PM

1. (30 points) *PCA via successive deflation.* Let $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ be the eigenvectors of the k largest eigenvalues of $\boldsymbol{C} = (1/n)\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$, i.e., the PCA solution. These satisfy

$$\boldsymbol{\theta}_c^\top \boldsymbol{\theta}_d = \begin{cases} 0 & c \neq d \\ 1 & c = d. \end{cases}$$

We will construct a method of finding $\boldsymbol{\theta}_c$ sequentially.

As we showed in class, $\boldsymbol{\theta}_1$ is the first principal eigenvector of $\boldsymbol{C}$, and satisfies $\boldsymbol{C}\boldsymbol{\theta}_1 = \lambda_1 \boldsymbol{\theta}_1$. Now define $\widetilde{\boldsymbol{\phi}}_i$ as the orthogonal projection of $\boldsymbol{\phi}_i$ onto the space orthogonal to $\boldsymbol{\theta}_1$:

$$\widetilde{\boldsymbol{\phi}}_i = (\boldsymbol{I} - \boldsymbol{\theta}_1 \boldsymbol{\theta}_1^\top)\boldsymbol{\phi}_i.$$

Define $\widetilde{\boldsymbol{\Phi}} = [\ \boldsymbol{\phi}_1 \ \cdots \ \boldsymbol{\phi}_n\ ]$ as the *deflated* data matrix, we have

$$\widetilde{\boldsymbol{\Phi}} = (\boldsymbol{I} - \boldsymbol{\theta}_1 \boldsymbol{\theta}_1^\top)\boldsymbol{\Phi}.$$

(a) Using the fact that $\boldsymbol{C}\boldsymbol{\theta}_1 = \lambda_1 \boldsymbol{\theta}_1$ and $\|\boldsymbol{\theta}_1\|^2 = 1$, show that the second moment of the deflated matrix is given by

$$\widetilde{\boldsymbol{C}} = \frac{1}{n}\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}^\top = \boldsymbol{C} - \lambda_1 \boldsymbol{\theta}_1 \boldsymbol{\theta}_1^\top.$$

(b) Show that $\boldsymbol{\theta}_2$ is the eigenvector of the largest eigenvalue of $\widetilde{\boldsymbol{C}}$. Recall that $\boldsymbol{\theta}_2$ is the eigenvector of the second-largest eigenvalue of $\boldsymbol{C}$.

(c) Suppose we have a simple method for finding the leading eigenvector and eigenvalue of a positive semi-definite matrix, denoted by $(\lambda, \boldsymbol{u}) = f(\boldsymbol{C})$. Write some pseudo code for finding the first k principal basis vectors of $\boldsymbol{\Phi}$ that only uses the special f function and simple vector arithmetic, i.e., your code should not use SVD or the eig function. Hint: this should be a simple iterative routine that takes 2–3 lines to write. The input is $\boldsymbol{C}$, k and the function f, the output should be $\boldsymbol{\theta}_c$ and λ_c for $c = 1, \ldots, k$. Do not worry about being syntactically correct.

2. (20 points) Consider a Gaussian mixture model in which the marginal distribution of the latent variable $\boldsymbol{y}$ is $\Pr(\boldsymbol{y} = \boldsymbol{e}_c) = \pi_c, c = 1, \ldots, k$, and the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{y}$ is $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma})$, i.e., each Gaussian component has their own mean $\boldsymbol{\mu}_c$ but they share the same covariance matrix $\boldsymbol{\Sigma}$. Given a set of observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, derive the expectation-maximization algorithm for estimating the model parameters $\pi_c, \boldsymbol{\mu}_c, c = 1, \ldots, k$, and $\boldsymbol{\Sigma}$.

3. (50 points) *20 Newsgroup revisited.* Let us revisit the 20 Newsgroup data set `<http://qwone.com/~jason/20Newsgroups/>`, and apply some of the unsupervised methods by ignoring their labels. We will only consider the training data. You are required to code the algorithms by yourselves in the language of your choice.

   (a) *LSI/PCA via orthogonal iteration.* Implement the orthogonal iteration algorithm that finds the PCA projection matrix $\boldsymbol{\Theta}$ of a data matrix $\boldsymbol{\Phi}$. You are allowed to use a pre-existing function of QR. Apply tf-idf to the term-document matrix to obtain $\boldsymbol{\Phi}$ and feed it into your orthogonal iteration algorithm. Remember to use sparse matrix operations to avoid unneccessary memory/computational complexities. Set $k = 2$ and let the algorithm run until $\boldsymbol{\Theta}$ doesn't change much. Then get $\boldsymbol{Y} = \boldsymbol{\Theta}^\top \boldsymbol{\Phi}$. Each column of $\boldsymbol{Y}$ is a two-dimensional vector that you can plot on a plain. Plot all the documents on a two-dimensional plain, and use a different color for each point that belong to different news groups.

   (b) *GMM via EM.* Implement the EM algorithm for the Gaussian mixture model (with different means and covariances for each Gaussian component). The data matrix $\boldsymbol{\Phi}$ is obtained from LSI with $k_{\text{LSI}} = 100$ using the previous orthogonal iteration algorithm. Run the EM algorithmn for GMM with $k_{\text{GMM}} = 20$ until convergence. For each Gaussian component $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, calculate $\boldsymbol{\Theta}\boldsymbol{\mu}_c$ where $\boldsymbol{\Theta}$ is the PCA projection; the vector $\boldsymbol{\Theta}\boldsymbol{\mu}_c$ should be element-wise nonnegative. For each cluster c, show the 10 terms that have the highest value in $\boldsymbol{\Theta}\boldsymbol{\mu}_c$. The index-term mapping can be found here `<http://qwone.com/~jason/20Newsgroups/vocabulary.txt>`. Does the result make sense?