CAP 6610 Machine Learning, Spring 2022

Homework 4

Due 4/20/2022 11:59PM

1. (15%) *Alternating optimization for PCA.* Recall that the principal component analysis (PCA) tries to find the projection matrix $\boldsymbol{\Theta}$ and embeddings $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ as the solution of

$$\underset{\boldsymbol{\Theta}, \boldsymbol{y}_i}{\text{minimize}} \quad \sum_{i=1}^{n} \|\boldsymbol{\phi}_i - \boldsymbol{\Theta}\boldsymbol{y}_i\|^2.$$

   We can define matrices $\boldsymbol{\Phi} = [\ \boldsymbol{\phi}_1\ \boldsymbol{\phi}_2\ \cdots\ \boldsymbol{\phi}_n\ ]$ and $\boldsymbol{Y} = [\ \boldsymbol{y}_1\ \boldsymbol{y}_2\ \cdots\ \boldsymbol{y}_n\ ]$ and rewrite the problem as

$$\underset{\boldsymbol{\Theta}, \boldsymbol{Y}}{\text{minimize}} \quad \|\boldsymbol{\Phi} - \boldsymbol{\Theta}\boldsymbol{Y}\|^2. \tag{1}$$

   Without imposing the constraint $\boldsymbol{\Theta}^\top\boldsymbol{\Theta} = \boldsymbol{I}$, derive an alternating optimization algorithm for (1).

2. (15%) *Naive Bayes GMM.* 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 $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ where each of the covariance matrices $\boldsymbol{\Sigma}_c$ is diagonal. This means given the latent variable $\boldsymbol{y}$, elements of the ambient variable $\boldsymbol{x}$ are conditionally independent, resembling the naive Bayes model. Given a set of observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, derive the expectation-maximization algorithm for estimating the model parameters $\pi_c, \boldsymbol{\mu}_c$, and $\boldsymbol{\Sigma}_c$ for $c = 1, \ldots, k$.

3. (40%) *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?

4. (30%) *Mixture of multinomials.* Consider the latent variable model with the following probability distribution:

$$\Pr(y_i = c) = \frac{1}{k}, \quad \Pr(\boldsymbol{x}_i | y_i = c) \sim \text{Multi}(\boldsymbol{p}_c, L_i),$$

meaning that y_i is categorical, with k possible outcomes and equal probability; $(\boldsymbol{x}_i | y_i = c)$ follows a multinomial distribution by drawing from $\boldsymbol{p}_c$ L_i times, i.e.,

$$p(\boldsymbol{x}_i | y_i = c) = \frac{L_i!}{\prod_{j=1}^{d} x_{ij}!} \prod_{j=1}^{d} {p_{cj}}^{x_{ij}}.$$

Given data samples $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$:

(a) Write out the maximum likelihood formulation for estimating $\boldsymbol{p}_1, ..., \boldsymbol{p}_k$. Simplify the objective function as much as possible.

(b) Derive an expectation-maximization algorithm for approximately solving the aforementioned problem.

(c) Implement this algorithm and try it on the 20 News Group data set with $k = 20$ (on the raw word-count data, without tf-idf preprocessing). Show the top 10 words in each cluster.