CAP 6610 Machine Learning, Spring 2020

Homework 5 (Midterm 2 replacement)

Due 4/30/2020 11:59PM

Each question is worth 5 points, so you only need to answer 3 of the 4 questions to get the full 15 points.

1. *Naive Bayes Gaussian discriminant analysis.* Consider the generative model for (supervised) classification by assuming $\Pr[y_i = c] = \pi_c$ and $\boldsymbol{x}_i | y_i \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$. Without any additional assumptions, this is the Gaussian discriminant analysis (aka quadratic discriminant analysis) that we covered in class. We've discussed one special case when all the $\boldsymbol{\Sigma}_c$ matrices are the same, which leads to the linear discriminant analysis (LDA) model.

   Here we make a different assumption: all the $\boldsymbol{\Sigma}_c$ matrices are diagonal (but not necessarily the same); for multivariate normals it means all variables are independent (conditioned on observing labels).

   (a) Derive the maximum likelihood estimate for the model parameters $\pi_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c, c = 1, \ldots, k$.

   (b) Given a new data point $\boldsymbol{x}_0$, explain how to predict $\widehat{y}_0$. Simplify the expression as much as possible.

2. Consider the lasso regularized k-class logistic regression problem

$$\underset{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^{n} \left(\log \sum_{c=1}^{k} \exp(\boldsymbol{\theta}_c^\top \boldsymbol{\phi}_i) - \boldsymbol{\theta}_{y_i}^\top \boldsymbol{\phi}_i \right) + \frac{\lambda}{2} \sum_{c=1}^{k} \|\boldsymbol{\theta}_c\|_1.$$

   Write the pseudocode of the proximal stochastic gradient descent algorithm for solving it.

   Hint: Denote $f(\boldsymbol{z}) = \log \sum \exp(\boldsymbol{z})$, then $\nabla f(\boldsymbol{z}) = \frac{1}{\sum \exp(\boldsymbol{z})} \exp(\boldsymbol{z})$, where we overload the definition of $\exp(\cdot)$ for vector inputs by taking exponential element-wise. If $g(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x})$, then $\nabla g(\boldsymbol{x}) = \boldsymbol{A}^\top \nabla f(\boldsymbol{A}\boldsymbol{x})$.

3. Consider the latent variable model with the following probability distribution:

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

   meaning that y_i is categorical, with k possible outcomes, and $\Pr(y_i = c) = \pi_c$; $(\boldsymbol{x}_i | y_i = c)$ 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 $\boldsymbol{p}_1, ..., \boldsymbol{p}_k$, and $\boldsymbol{\pi}$. 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.

4. *Multidimensional scaling (MDS).* MDS is another classical approach for unsupervised embedding, and to some extent relates to PCA.

The main idea of MDS is to embed each $\boldsymbol{x}_i$ to $\boldsymbol{y}_i$ so that the pair-wise distances are preserved as much as possible. This can be formulated as the following optimization problem

$$\underset{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n}{\text{minimize}} \quad \sum_{i=1}^{n}\sum_{j=1}^{i-1}(\|\boldsymbol{x}_i - \boldsymbol{x}_j\| - \|\boldsymbol{y}_i - \boldsymbol{y}_j\|)^2.$$

There is no close-form solution for this formulation. A modified formulation called *classical scaling* is proposed:

$$\underset{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n}{\text{minimize}} \quad \sum_{i=1}^{n}\sum_{j=1}^{n}(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{x}}_j - \boldsymbol{y}_i^\top \boldsymbol{y}_j)^2, \tag{1}$$

where $\widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - (1/n)\sum_{j=1}^{n}\boldsymbol{x}_j$ is the centered data.

(a) Use the Eckart-Young theorem to show that an optimal solution of (1) is to take the eigen-decomposition of the matrix $\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, where $\widetilde{\boldsymbol{X}} = [\ \widetilde{\boldsymbol{x}}_1 \ \cdots \ \widetilde{\boldsymbol{x}}_n\ ]$, keep the k largest eigenvalues in $\boldsymbol{\Lambda}$ and the corresponding columns in $\boldsymbol{U}$, and let $\boldsymbol{Y} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^\top$; then each $\boldsymbol{y}_i$ is the ith column of $\boldsymbol{Y}$.

(b) Oftentimes one is directly given the pair-wise distance matrix $\boldsymbol{D}$, where

$$D_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2,$$

without explicitly given the data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Show that we can define the matrix

$$\boldsymbol{B} = \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top\right)\boldsymbol{D}\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top\right),$$

and replace $\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{x}}_j$ with B_{ij} in formulation (1).