Homework 5

Special Topics in Advanced Machine Learning
Spring 2017
Instructor: Anna Choromanska

Homework is due 05/03/2018.

Problem 1 (35 points): Optimization

Implement Kernel Logistic Regression with L_2 regularizer using empirical kernel map, i.e., optimize,

$$J(\omega) = -\sum_{i=1}^{N} \log(\sigma(y_i \omega^\top k_i)) + \lambda \omega^\top \omega,$$

to get ω. Here, k_i is a column vector such that $k_i = [k(x_i, x_1). \ldots, k(x_i, x_j), \ldots, k(x_i, x_N)]^\top$, y_i is a label of data point x_i, and $\sigma(v) = 1/(1 + e^{-v})$. Use RBF (Gaussian) kernel with $\sigma^2 = \frac{1}{N^2} \sum_{i,j=1}^{N} \|x_i - x_j\|^2$.

After ω is obtained, for any test data x, compute $p(y = 1|x) = \sigma(\omega^\top k_x)$, where $k_x = [k(x, x_1), k(x, x_2), \ldots, k(x, x_N)]^\top$. If $p(y = 1|x) > 0.5)$ the predicted label is 1, else it is -1. Report the accuracy.

Use the following methods to optimize $J(\omega)$:

a) [6 points] GD

b) [7 points] SGD (for each iteration use p points to estimate the gradients and explore two settings of p: $p = 1$ and $p = 100$)

c) [10 points] BFGS(randomly sample 4000 training points, i.e. 2000 from each class, and use them to describe the empirical kernel map and construct the approximation of inverse Hessian using BFGS method)

d) [12 points] repeat the same experiment as for BFGS, but instead for LBFGS, where you use a small number of vectors (experiment with a couple of choices) to approximate inverse Hessian

You will use data set "data1.mat". Experiment with various step sizes and pick what works the best for you. Compare how the value of the cost function decreases with time for different methods. Stop the iterations, if the gradient becomes smaller than epsilon (say, $1e-5$). Compare the methods.

Problem 2 (20 points): EM

Consider a random variable x that is categorical with M possible values $1, 2, \ldots, M$. Suppose x is represented as a vector such that $x(i) = 1$ if x takes the ith value, and $\sum_{i=1}^{M} x(i) = 1$. The distribution of x is described by a mixture of K discrete multinomial distributions such that:

$$p(x) = \sum_{k=1}^{K} \pi_k p(x|\mu_k)$$

and

$$p(x|\mu_k) = \prod_{j=1}^{M} \mu_k(j)^{x(j)},$$

where π_k denotes the mixing coefficient for the kth component (aka the prior probability that the hidden variable $z = k$), and μ_k specifies the parameters of the k^{th} component. Specifically, $\mu_k(j)$ represents the probability $p(x(j) = 1|z = k)$, and $\sum_j \mu_k(j) = 1$. Given an observed data set $\{x_i\}, i = 1, 2, \ldots, N$, derive the E and M step of the EM algorithm for optimizing the mixing coefficients and the component parameters $\mu_k(j)$ for this distribution (below we provide the generic formula for the E and M steps, where θ denotes all the parameters of the mixture model).

- E-step (5 points): For each i, calculate $Q_i(z_i) = p(z_i|x_i; \theta)$, i.e. the probability that observation i belongs to each of the K clusters.

- M-step (15 points): Set

$$\theta := \arg\max_{\theta} \sum_{i=1}^{N} \sum_{z_i} Q_i(z_i) \log \frac{p(x_i, z_i; \theta)}{Q_i(z_i)}.$$

Problem 3 (10 points): PCA

Download the "teapots.mat" data set containing 100 images of teapots of size 38×50. To view an image, say the second one in the data set type:
imagesc(reshape(teapotImages(2,:),38,50));
colormap gray;
Compute the data mean and top 3 eigenvectors of the data covariance matrix and show them as images. Reconstruct the data using PCA with least squares error using only the mean and a linear combination of the top 3 eigenvectors. Show 10 different images before and after reconstruction. Discuss results.

Problem 4 (15 points): PCA

Given input vectors $\{x_1, \cdots, x_T\}$ where $x_i \in \mathbb{R}^n$, the goal of principal components analysis (PCA) is to find a low-dimensional approximation of the data minimizing the *quadratic compression loss*. More formally, we want to find an n-dimensional vector m and a rank k projection matrix P, where $k \leq n$, such that the following loss function is minimized:

$$\text{comp}(P, m) = \sum_{t=1}^{T} \|(x_t - m) - P(x_t - m)\|_2^2$$

Differentiating and solving for m gives: $m^* = \bar{x}$, where $\bar{x}$ is the data mean. Show that substituting m^* to the expression for loss function yields:

$$\text{comp}(P) = \text{tr}(C) - \text{tr}(PC)$$

where C is the data covariance matrix and tr is the matrix trace. Furthermore, show that $\text{tr}(PC)$ is maximized if P consists of the k eigenvectors of C with the largest eigenvalues.

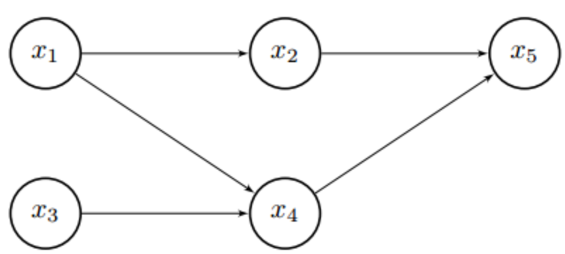
Problem 5 (10 points): Clustering

Lemma 1 *Let $\phi(W_1)$ be the optimum value of the k-means objective for the k-clustering of data set W_1, and let $\phi(W_2)$ be the optimum value of the k-means objective for the k-clustering of data set W_2. Finally, let $\phi(W_1 \cup W_2)$ be the optimum value of the k-means objective for the k-clustering of data set $W_1 \cup W_2$. Prove that*

$$\phi(W_1) + \phi(W_2) \leq \phi(W_1 \cup W_2).$$

Problem 6 (10 points): Bayesian Network Conditional Independence

Consider the Bayesian network below with binary variables representing the following: x_1 student is intelligent, x_2 student is good at taking tests, x_3 student is hard working, x_4 student understands the material, and x_5 student gets good grade.



Write out the factorization of the probability distribution $p(x_1, ..., x_5)$ implied by this directed graph. Then, using the Bayes ball algorithm, indicate for each statement below if it is True or False and justify your answers

- x_2 and x_4 are independent.
- x_2 and x_4 are conditionally independent given x_1, x_3, and x_5.
- x_2 and x_4 are conditionally independent given x_1 and x_3.
- x_5 and x_3 are conditionally independent given x_4.
- x_5 and x_3 are conditionally independent given x_1, x_2, and x_4.
- x_1 and x_3 are conditionally independent given x_5.
- x_1 and x_3 are conditionally independent given x_2.
- x_2 and x_3 are independent.
- x_2 and x_3 are conditionally independent given x_5.
- x_2 and x_3 are conditionally independent given x_5 and x_4.