CS 6780: Advanced Machine Learning
Homework 3

Due date: 12pm. **Nov 11, 2010** (Printed submission.)

Note:
1. Write your name and net-id in BIG letters on the first page. Do not write anything else on the first-page of the homework.
2. For derivation questions, you need to show all the necessary steps.
3. For data analysis / programming parts, you need to submit: (a) the specific values and plots requested, (b) explanation in the text if needed (don't give us the code printouts only!), and (c) also provide the code in the appendix.

1 GDA

(a) [10 points] For this part of the problem only, you may assume n (the dimension of x) is 1, so that $\Sigma = [\sigma^2]$ is just a real number, and likewise the determinant of Σ is given by $|\Sigma| = \sigma^2$. Given the dataset, we claim that the maximum likelihood estimates of the parameters are given by

$$\phi = \frac{1}{m}\sum_{i=1}^{m} 1\{y^{(i)} = 1\} \tag{1}$$

$$\mu_0 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}} \tag{2}$$

$$\mu_1 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}} \tag{3}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \tag{4}$$

The log-likelihood of the data is

$$l(\phi, \mu_0, \mu_1, \Sigma) = \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma) \tag{5}$$

$$= \log \prod_{i=1}^{m} p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi) \tag{6}$$

By maximizing $l(.)$ with respect to the four parameters, prove that the maximum likelihood estimates of ϕ, μ_0, μ_1, and Σ are indeed as given in the formulas above. (You may assume that there is at least one positive and one negative example, so that the denominators in the definitions of μ_0 and μ_1 above are non-zero.)

(b) [Bonus extra credit points] Without assuming that n = 1, show that the maximum likelihood estimates of ϕ, μ_0, μ_1, and Σ are as given in the formulas in part (a). [Note: If youre fairly sure that you have the answer to this part right, you dont have to do part (a), since thats just a special case.]

2 Regression

We have two different datasets: (X_1, y_1) and (X_2, y_2). One way to learn the parameters is to train two independent linear regressions:
$\theta_{1,ind} = \arg\min(X_1\theta_1 - y_1)^T(X_1\theta_1 - y_1)$
$\theta_{2,ind} = \arg\min(X_2\theta_2 - y_2)^T(X_1\theta_2 - y_2)$.

However, the two datasets are somewhat related, therefore we would like to train two different parameters θ_1 and θ_2, but would want θ_1 and θ_2 to be close. Derive the closed form expression for θ_1 that minimizes $J(\theta_1, \theta_2)$ as given below.

$J(\theta_1, \theta_2) = (X_1\theta_1 - y_1)^T(X_1\theta_1 - y_1) + (X_2\theta_2 - y_2)^T(X_2\theta_2 - y_2) + \lambda(\theta_1 - \theta_2)^T(\theta_1 - \theta_2)$

Note that closed form expression for θ_1 should only depend on X_1, X_2, y_1, y_2. (10 points)

3 GLM

Consider using GLM with a response variable from any member of the exponential family in which $T(y) = y$, and the canonical response function for the family. I.e., $\eta = \theta^T x$ and $h(x) = E[y|x] = \int_y p(y|x; \theta) y$.

Show that stochastic gradient ascent on the log-likelihood $\log p(y|X;\theta)$ results in the update rule:
$\theta_{j+1} = \theta_j - \alpha(h(x^{(i)}) - y^{(i)})x^{(i)}$

Hint: $\exp(a(\eta)) = \int_y b(y)\exp(\eta y)$. ($a(\eta)$ plays the role of normalizing the exponential family distribution.) Compute $a'(\eta)$. (10 points)

4 Isomap

Isomaps could be sometimes useful for visualizing data (and could also be used as a dimensionality reduction technique to run further learning algorithms).

(a) Go to `http://waldron.stanford.edu/` $\sim$ *isomap/*, and download the code. Run it on the Swiss-roll data set (`http://waldron.stanford.edu/` $\sim$ *isomap/datasets.html*). Add IID Gaussian noise $N(0, \sigma^2)$ to the data, and run isomap algorithm on it. Increase $\sigma^2 \geq 0$, and report the value of σ when the isomap algorithm fails to find a good embedding.[1] (7 points)

(b) Take the dataset from your course project (or your dataset related to your research) and run isomap on that dataset.[2] Report what you observe. (8 points)

5 ICA

In this question, we will prove that the sources are not separable if they are Gaussian. (This includes sources that are Gaussian and with arbitrary covariance matrices.)

Specifically, assume two zero mean sources s_1 and s_2. I.e., $s = \{s_1, s_2\} \in \Re^2$ and $s \sim N(0, \Sigma)$. After mixing, we record $x = As$ for $A \in \Re^{2\times 2}$. Our goal is to find A and s given x for m recordings.

(a) [3 points] Given s is a zero mean Gaussian, prove that x is a zero-mean Gaussian too.

[1] Hint: (a) Use atleast 2000 points; trying to use all of them might take quite long, (b) Use `Isomap(.)`; if you use `IsomapII`, then make sure you are using the compiled mex version of dijsktra file.

[2] Only for theoretical projects where datasets do not apply at all, you can do the following instead: Download the faces data `http://waldron.stanford.edu/` $\sim$ *isomap/face_data.mat.Z*, and the isomap code on this data.

(b) [7 points] Since x is a Gaussian, $E(xx^T)$ completely defines the distribution. Prove that given $E(xx^T)$, it is impossible to figure out if the source was $N(0, \Sigma)$ with mixing matrix A or if the source was $N(0, R^T \Sigma R)$ with mixing matrix AR. (R is a rotation matrix and it follows $RR^T = R^T R = I$.)

6 Clustering

(a) Spectral clustering. For the version of the algorithm described in the class (similar to Ng, Jordan and Weiss, 2000), the eigenvalues of the normalized affinity matrix $L = D^{-.5} W D^{-.5}$ for 20 data-points are provided. Estimate the number of clusters in the data. (Explain in a line.)

Eigenvalues = `0, .1, .9, .95, 0, 1, 0, .05, 0, .1, .9, .1, .2, .1, 1, 0, 0, .1, .2, .85`

(5 points)

(b) Consider `A. Mixture of Gaussians model with EM` and `B. k-means clustering`. You have to prove that 'A' becomes same as 'B' with the following changes in 'A':

- Fix $\Sigma = \sigma^2 I$ for all clusters.
- Replace $w_j^{(i))} = p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$ in the E-step with:

$$j^* = \arg\max_j \quad p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$

$$w_j^{(i)} = 1\{j = j^*\}$$

This is called "Hard-EM".

(5 points)

7 Non-Negative Matrix Factorization

Many of the variables that we deal with are constrained to be positive. PCA does not put any such constraint on the variables. Here we have a data matrix $X \in \Re^{m \times n}$ with m data-points in n dimensions. We want to factor X as

$$X \approx WH \tag{7}$$

where W is $m \times k$, H is $k \times n$, $k \leq \max(m, n)$, and $x_{ij}, w_{ik}, h_{kj} \geq 0$. Assume that x_{ij} has a Poisson distribution with mean $(WH)_{ij}$ (i.e., ij^{th} entry of matrix WH), and each x_{ij} is IID.

(a) Derive the log-likelihood $L(W, H)$ of the model above. Is $L(W, H)$ convex in (W, H)? (5 points)

(b) Kullback-Leibler divergence between two distributions $P(i, j)$ and $Q(i, j)$ is defined as

$$D_{KL}(P||Q) = \sum_{ij} P(i, j) \log \frac{P(i, j)}{Q(i, j)} \tag{8}$$

Prove that maximizing $L(W, H)$ is equivalent to minimizing $D_{KL}(X||WH)$, where we consider $X(i, j)$ and $WH(i, j)$ as normalized distributions. (I.e., if $\sum_{ij} X(i, j) = 1$ and $\sum_{ij} WH(i, j) = 1$. (7 points)

(c) We will use the following alternating algorithm (Lee and Seung, 2001), in which we iterate between optimizing for W (assuming H constant) and optimizing for H (assuming W constant). Prove that for $\delta g_H(W)/\delta w_{ik} = 0$ and for $\delta h_W(H)/\delta h_{kj} = 0$ respectively, the following update rule will converge:

$$w_{ik} \leftarrow w_{ik} \frac{\sum_{j=1}^{n} h_{kj} x_{ij}/(WH)_{ij}}{\sum_{j=1}^{n} h_{kj}} \tag{9}$$

$$h_{kj} \leftarrow h_{kj} \frac{\sum_{i=1}^{m} w_{ik} x_{ij}/(WH)_{ij}}{\sum_{i=1}^{m} w_{ik}} \tag{10}$$

Here $g_H(W) = L(W, H)$ with H constant, and $h_W(H) = L(W, H)$ with W constant. (8 points)

(d) Write code for finding W and H given X. (Provide the code in the appendix.) Find the Non-negative Matrix Factorization of the face-data in previous homework.[3] I.e., X is $m \times n$ matrix of the face-data (where each row is a face). For $k = 10$, display the 10 basis images H. (15 points)

[3]You may have to choose smart methods to initialize.