CS 7641 CSE/ISYE 6740 Homework 2

Yao Xie

Deadline: Feb. 13, Sat., 11:55pm

- Submit your answers as an electronic copy on T-square.
- No unapproved extension of deadline is allowed. Zero credit will be assigned for late submissions. Email request for late submission may not be replied.
- For typed answers with LaTeX (recommended) or word processors, extra credits will be given. If you handwrite, try to be clear as much as possible. No credit may be given to unreadable handwriting.
- Explicitly mention your collaborators if any.

1 PCA for face recognition [20 points]

1. Perform data analysis on the Yale face dataset (on Tsquare) for subject 14. Plot the mean face and the first 6 eigenfaces for subject 14.
2. Now use subject14.test.gif to perform face recognition using the following procedure.

For doing face recognition through PCA we proceed as follows. Given the test image, we project it using the first component to obtain the coefficient vector, and compare $|z^\intercal u_{1,i}|$, $i = 1, 2$, where $u_{1,i}$ is the first dominant component for ith person. Are we able to recognize the person correctly using the first principle component?

Remark: you have to perform downsampling of the image by a factor of 4 to turn them into a lower resolution image before we do anything. See the example MATLAB code.

2 PCA: yet another interpretation [20 points]

In class, we showed that PCA finds the "variance maximizing" directions onto which to project the data. In this problem, we find another interpretation of PCA.

Suppose we are given a set of points $x_1, \cdots, x_n$. Let us assume that we have as usual preprocessed the data to have zero-mean and unit variance in each coordinate. For a given unit-length vector v, let $f_v(x)$ be the projection of point x onto the direction given by v. I.e., if $\mathcal{V} = \{\alpha v : \alpha \in \mathbb{R}\}$, then

$$f_v(x) = \arg\min_{u \in \mathcal{V}} \|x - u\|^2.$$

Show that the unit-length vector v that minimizes the mean squared error between projected points and original points corresponds to the first principal component for the data. I.e., show that

$$\arg\min_{\|v\|} \sum_{i=1}^{n} \|x_i - f_v(x_i)\|^2$$

gives the principle component.

Remark. If we are asked to find a k-dimensional subspace onto which to project the data so as to minimize the sum of squares distance between the original data and their projections, then we should choose the k-dimensional subspace spanned by the first k principal components of the data. This problem shows that this result holds for the case of $k = 1$.

3 Order of faces using ISOMAP [20 points]

The objective of this homework is to develop familiarity with the ISOMAP algorithms discussed in class. The file images.tar.gz contains 698 images, corresponding to different poses of the same face. Each image is given as a 64×64 luminosity map, hence represented as a vector in R4096. This vector is stored as a row in the file. [This is one of the datasets used in J.B. Tenenbaum, V. de Silva, and J.C. Langford, Science 290 (2000) 2319-2323]

(a) Choose a distance between images (i.e. in this case a distance in $\mathbb{R}^{4096}$. Construct a proximity graph with vertices corresponding to the images, and connecting each image to the k nearest neighbors in the dataset, for a suitable k. (Notice that as a result, each vertex is in general connected to more than k neighbors.)

(b) Implement the ISOMAP algorithm and apply it to this graph to obtain a $d = 2$-dimensional embedding. You may refer to the code isomapII.m Present a plot of this embedding. Find three points that are close to each other and show what they look like. Do you see any similarity among them?

4 Density Estimation [15 points]

Consider a histogram-like density model in which the space x is divided into fixed regions for which density $p(x)$ takes constant value h_i over ith region, and that the volume of region i in denoted as Δ_i. Suppose we have a set of N observations of x such that n_i of these observations fall in regions i.

(a) What is the log-likelihood function?

(b) Derive an expression for the maximum likelihood estimator for h_i.

Hint: This is a constrained optimization problem. Remember that $p(x)$ must integrate to unity. Since $p(x)$ has constant value h_i over region i, which has volume Δ_i. The normalization constraint is $\sum_i h_i \Delta_i = 1$. Use Lagrange multiplier by adding $\lambda \left(\sum_i h_i \Delta_i - 1\right)$ to your objective function.

(c) Mark T if it is always true, and F otherwise. Briefly explain why.

- Non-parametric density estimation usually does not have parameters.
- Histogram is an efficient way to estimate density for high-dimensional data.
- Parametric density estimation assumes the shape of probability density.

5 EM for Mixture of Gaussians [25 points]

Mixture of K Gaussians is represented as

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k), \tag{1}$$

where π_k represents the probability that a data point belongs to the kth component. As it is probability, it satisfies $0 \leq \pi_k \leq 1$ and $\sum_k \pi_k = 1$. In this problem, we are going to represent this in a slightly different

manner with explicit latent variables. Specifically, we introduce 1-of-K coding representation for latent variables $z^{(k)} \in \mathbb{R}^K$ for $k = 1, ..., K$. Each $z^{(k)}$ is a binary vector of size K, with 1 only in kth element and 0 in all others. That is,

$$\begin{aligned} z^{(1)} &= [1; 0; ...; 0] \\ z^{(2)} &= [0; 1; ...; 0] \\ &\vdots \\ z^{(K)} &= [0; 0; ...; 1]. \end{aligned}$$

For example, if the second component generated data point x^n, its latent variable z^n is given by $[0; 1; ...; 0] = z^{(2)}$. With this representation, we can express $p(z)$ as

$$p(z) = \prod_{k=1}^{K} {\pi_k}^{z_k},$$

where z_k indicates kth element of vector z. Also, $p(x|z)$ can be represented similarly as

$$p(x|z) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k}.$$

By the sum rule of probability, (1) can be represented by

$$p(x) = \sum_{z \in Z} p(z)p(x|z). \tag{2}$$

where $Z = \{z^{(1)}, z^{(2)}, ..., z^{(K)}\}$.

(a) Show that (2) is equivalent to (1).

(b) In reality, we do not know which component each data point is from. Thus, we estimate the responsibility (expectation of z_k^n) in the E-step of EM. Since z_k^n is either 1 or 0, its expectation is the probability for the point x_n to belong to the component z_k. In other words, we estimate $p(z_k^n|x_n)$. Derive the formula for this estimation by using Bayes rule. Note that, in the E-step, we assume all other parameters, i.e. π_k, μ_k, and Σ_k, are fixed, and we want to express $p(z_k^n|x_n)$ as a function of these fixed parameters.

(c) In the M-Step, we re-estimate parameters π_k, μ_k, and Σ_k by maximizing the log-likelihood. Given N i.i.d (Independent Identically Distributed) data samples, derive the update formula for each parameter. Note that in order to obtain an update rule for the M-step, we fix the responsibilities, i.e. $p(z_k^n|x_n)$, which we have already calculated in the E-step.

Hint: Use Lagrange multiplier for π_k to apply constraints on it.

(d) EM and K-Means

K-means can be viewed as a particular limit of EM for Gaussian mixture. Considering a mixture model in which all components have covariance ϵI, show that in the limit $\epsilon \to 0$, maximizing the expected complete data log-likelihood for this model is equivalent to minimizing objective function in K-means:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \|x_n - \mu_k\|^2,$$

where $\gamma_{nk} = 1$ if x_n belongs to the k-th cluster and $\gamma_{nk} = 0$ otherwise.

(e) General setting

Consider a mixture of distribution of the form

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

where the elements of x could be discrete or continuous or a combination of these. Express the mean and covariance of the mixture distribution using the mean μ_k and covariance Σ_k of each component distribution $p(x|k)$.