THU-70250043, Pattern Recognition (Spring 2016) **Homework: 7**

PCA and Non-linear dimensionality reduction

Lecturer: Changshui Zhang zcs@mail.tsinghua.edu.cn

Student: XXX xxx@mails.tsinghua.edu.cn

Problem 1

Maximum-variance and Minimum-error approaches to PCA

You have gone through the K-L and PCA algorithms in the class, but don't be confused by the names of this algorithm: Principal component analysis(PCA) is also known as K-L transform, they are completely the same thing.

PCA can be defined as the orthogonal projection of the data onto a lower dimensional linear space, known as the principal subspace, such that *the variance of the projected data is maximized*; It can also be defined as the linear projection that *minimizes the average projection cost*, defined as the mean squared distance between the data points and their projections. We'll consider both approaches in this problem.

Suppose we have a data set of observations $\{\mathbf{x}_n\}$, where $n = 1, ..., N$, $\mathbf{x}_n \in \mathcal{R}^D$. Our goal is to project the data onto a space having dimensionality $M < D$.

Maximum-variance approach

1.1 To begin with, let's consider the projection onto a one-dimensional space($M = 1$). Define the direction of this space using a D-dimensional vector $\mathbf{u}_1$. Prove that the variance of the projected data is given by the following expression: $\mathbf{u}_1^T S \mathbf{u}_1$, where S is called the data covariance matrix.

1.2 We now maximize the projected variance $\mathbf{u}_1^T S \mathbf{u}_1$ with respect to $\mathbf{u}_1$. Clearly, this has to be a constrained maximization to prevent $||\mathbf{u}_1|| \to \infty$, the appropriate constraint comes from the normalization condition $\mathbf{u}_1^T \mathbf{u}_1 = 1$. We can introduce a Lagrange multiplier denoted as λ_1 and solve this maximization problem. Show the details of solving $\mathbf{u}_1$ and the maximum variance.

1.3 For cases $M \geq 2$, we use proof by induction to show the same principle. Suppose the result above holds for some general value of M, show that it consequently holds for dimensionality $M + 1$.

Hint: To do this, remember that we've already selected the M eigen-vectors corresponding to the M largest eigen-values of the data covariance matrix S. We're now tying to maximize the variance on direction $\mathbf{u}_{M+1}$. The maximization should be done subject to the constraints that $\mathbf{u}_{M+1}$ be orthogonal to the existing vectors $\mathbf{u}_1, ..., \mathbf{u}_M$, and also that it be normalized to unit length.

Minimum-error approach

Suppose we have a *complete orthonormal set* of D-dimensional basis vector $\{\mathbf{u}_i\}$ where $i = 1, ..., D$ that satisfy:

$$\mathbf{u}_i^T \mathbf{u}_j = \delta_{ij} \tag{1}$$

Since this basis is *complete*, each data point can be presented as follows:

$$\mathbf{x}_n = \sum_{i=1}^{D} (\mathbf{x}_n^T \mathbf{u}_i)\mathbf{u}_i \tag{2}$$

Our goal is to approximate this data point using a restricted number $M < D$ of variables corresponding to a projection onto a lower-dimensional subspace. Let's make use of the following expression:

$$\tilde{\mathbf{x}}_n = \sum_{i=1}^{M} z_{ni}\mathbf{u}_i + \sum_{i=M+1}^{D} b_i \mathbf{u}_i \tag{3}$$

Where the $\{z_{ni}\}$ depends on the particular data point, whereas the $\{b_i\}$ are constants same for all data points.

Just as the subtitle, we shall minimize the squared distance between the original data point $\mathbf{x}_n$ and its approximation $\tilde{\mathbf{x}}_n$:

$$J = \frac{1}{N}\sum_{n=1}^{N} ||\mathbf{x}_n - \tilde{\mathbf{x}}_n||_2^2 \tag{4}$$

1.4 Prove that the optimal $\{z_{ni}\}$ complies with:

$$z_{ni} = \mathbf{x}_n^T \mathbf{u}_i \tag{5}$$

And the optimal $\{b_i\}$ complies with:

$$b_i = \bar{\mathbf{x}}^T \mathbf{u}_j \tag{6}$$

Where $\bar{\mathbf{x}} = (1/N)\sum_{n=1}^{N} \mathbf{x}_n$.

1.5 If we substitute for $\{z_{ni}\}$ and $\{b_i\}$, and make use of the general expansion (2), we obtain:

$$\mathbf{x}_n - \tilde{\mathbf{x}}_n = \sum_{i=M+1}^{D} \{(\mathbf{x}_n - \bar{\mathbf{x}})^T \mathbf{u}_i\}\mathbf{u}_i \tag{7}$$

Take equation (7) into equation (4) to get the distortion measure J purely related to $\{\mathbf{u}_i\}$ and the data covariance matrix S. Then, show that the distortion measure J with constraints (1) can be further written as:

$$L = Tr\{USU\} + Tr\{H(I - U^T U)\} \tag{8}$$

Where, U is a matrix of dimension $D \times (D-M)$ whose columns are given by $\mathbf{u}_i$. Matrix H is the Lagrange multipliers, one for each constraint.

1.6 Take the derivative of (8) with respect to U to find the optimal $\mathbf{u}_i$.

Hint: The last problem may be a little tricky and involves derivative of trace. You can look up the uploaded book "The Matrix Cookbook".

1.7 Give me your explicit explanations of these two approaches. You can illustrate it with hand-drawn images, either inserted in your report or included in the package is ok.

Problem 2

Probabilistic PCA

In the previous problem, the PCA was based on a linear projection of the data onto a subspace. In this problem, we'll show that PCA can also be expressed as the maximum likelihood solution of a probabilistic latent variable model.

Probabilistic PCA is a simple example of the linear-Gaussian framework, in which all of the marginal and conditional distributions are Gaussian. We can formulate this model by first introducing an explicit latent variable $\mathbf{z}$ corresponding to the principle-component subspace. Next we define a Gaussian prior $p(\mathbf{z})$ over the latent variable with zero-mean and unit-covariance:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|0, I) \tag{9}$$

Similarly, the conditional distribution of the observed variable $\mathbf{x}$ is again Gaussian:

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|W\mathbf{z} + \mu, \sigma^2 I) \tag{10}$$

With $W \in \mathcal{R}^{D\times M}$, $\mu \in \mathcal{R}^D$ governing the mean and σ^2 governing the variance of the conditional distribution.

2.1 Suppose we wish to determine the values of the parameters W, μ and σ^2 using maximum likelihood, so we need an expression for the marginal distribution $p(\mathbf{x})$ of the observed variable. From the sum and product rules of probability, we have:

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} \tag{11}$$

Show that the marginal distribution is given by:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, C) \tag{12}$$

Where $C = WW^T + \sigma^2 I$

2.2 We also require the posterior distribution $p(\mathbf{z}|\mathbf{x})$, show that:

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|M^{-1}W^T(\mathbf{x} - \mu), \sigma^{-2}M) \tag{13}$$

Where $M = W^TW + \sigma^2 I$.

2.3 Write down the log likelihood function of the observed variables:

$$lnp(X|\mu, W, \sigma^2) = \sum_{n=1}^{N} lnp(\mathbf{x}_n|W, \mu, \sigma^2) \tag{14}$$

Then verify that maximizing (14) with respect to μ gives the result $\mu_{ML} = \bar{\mathbf{x}}$ where $\bar{\mathbf{x}}$ is the mean of the data vectors.

2.4 Maximization with respect to W and σ^2 is more complex but has exact closed-form solution:

$$W_{ML} = U_M(L_M - \sigma^2 I)^{\frac{1}{2}}R \tag{15}$$

$$\sigma^2_{ML} = \frac{1}{D-M}\sum_{i=M+1}^{D} \lambda_i \tag{16}$$

Where U_M is a $D \times M$ matrix whose columns are given by the eigenvectors of the data covariance matrix S, and the corresponding eigenvalues are the M largest; L_M is a $M \times M$ diagonal matrix with the largest M eigenvalues of S as its elements; R is an arbitrary $M \times M$ orthogonal matrix; $\lambda_i (i = M+1, ..., D)$ are the smallest eigenvalues.

As what we've derived from 2.2, we have:

$$E[\mathbf{z}|\mathbf{x}] = M^{-1}W_{ML}^{T}(\mathbf{x} - \bar{\mathbf{x}}) \tag{17}$$

M is the same as that of equation (13).

Show that in the limit $\sigma^2 \to 0$, the posterior mean for the probabilistic PCA model becomes an orthogonal projection onto the principal subspace, as in conventional PCA.

Hint: Orthogonal projection in the conventional PCA is $\mathbf{y}_n = L_M^{-\frac{1}{2}} U_M^T(\mathbf{x}_n - \bar{\mathbf{x}})$. *When* $M = D$, *this operation is called whitening or sphereing which makes* $\{y_n\}$ *have zero-mean and unit-variance.*