CS 475 Machine Learning: Lecture 17
Probabilistic PCA

Prof. Mark Dredze

1 Probabilistic PCA

We are given points ($\mathbf{x}$) in a D dimensional space.
We want to find a M dimensional space, where $M < D$ that corresponds to the high dimensional space.
Point $\mathbf{x}_i$ is represented as $\mathbf{z}_i$ in the low dimensional space.

Assume a Gaussian prior distribution $p(\mathbf{z})$ over the latent variable and a Gaussian conditional distribution $p(\mathbf{x}|\mathbf{z})$ for the observed variable $\mathbf{x}$ conditioned on the latent variable. We'll define $p(\mathbf{z})$ as:

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

The conditional distribution is given by:

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

This distribution is parametrized by $\mathbf{W}, \mu, \sigma$. $\mathbf{W}$ is a $D \times M$ matrix that transforms $\mathbf{z}$ into high dimensional space, μ is a D-dimensional vector that gives the mean of points in the high dimensional space. We assume identity covariance of σ^2. Notice that the elements in $\mathbf{x}$ are based on a naive Bayes model.

This model assumes that high dimensional points are given by:

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \mu + \epsilon$$

where ϵ is a zero-mean Gaussian distributed noise variable with covariance $\sigma^2\mathbf{I}$.

1.1 Likelihood

Let's write the likelihood of this model given a collection of examples. What is $p(\mathbf{x})$? Observe that the expectation is given as:

$$\mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{W}\mathbf{z} + \mu + \epsilon] = \mu$$

since $\mathbf{z}$ and ϵ have expected values of 0. The expected covariance is:

$$cov[\mathbf{x}] = \mathbb{E}[(\mathbf{W}\mathbf{z} + \epsilon)(\mathbf{W}\mathbf{z} + \epsilon)^T] = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

Therefore, the distribution for $p(\mathbf{x})$ can be written as a Gaussian:

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

where:

$$\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

1.2 Maximum Likelihood

We've conveniently used the fact that both $p(\mathbf{z})$ and $p(\mathbf{x}|\mathbf{z})$ are Gaussians to write $p(\mathbf{x})$ directly as a Gaussian. This makes maximum likelihood for our model simpler.

The log likelihood function is written in the familiar form as:

$$\log p(\mathbf{X}|\mu, \mathbf{W}, \sigma^2) = \sum_{i=1}^{N} \log p(\mathbf{x}_i|\mathbf{W}, \mu, \sigma^2)$$

$$= -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\mathbf{C}| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i - \mu)^T\mathbf{C}^{-1}(\mathbf{x}_i - \mu) \quad (1)$$

If we take the derivative with respect to μ, we get the familiar maximum likelihood solution, that:

$$\mu = \bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

Maximization with respect to $\mathbf{W}$ and σ^2 is more complex. For $\mathbf{W}$ we obtain:

$$\mathbf{W} = \mathbf{U}_M(\mathbf{L}_M - \sigma^2\mathbf{I})^{1/2}\mathbf{R}$$

$\mathbf{U}_M$ is a $D \times M$ matrix whose columns are given by any subset of M eigenvectors of the data covariance matrix S:

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

($\mathbf{L}_M$ is an $M \times M$ matrix has values given by the eigenvalues λ_i and $\mathbf{R}$ is an arbitrary $M \times M$ orthogonal matrix. The maximum is given when the M eigenvectors are the largest. We typically assume they are arranged in decreasing order

The columns of $\mathbf{W}$ are said to define the principal subspace.

The corresponding solution for σ^2 is:

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

The matrix $\mathbf{R}$ can be any orthogonal matrix, so we can just assume that $\mathbf{R} = \mathbf{I}$. This means that the columns of $\mathbf{W}$ the principal component eigenvectors scaled by the variance parameters $\lambda_i - \sigma^2$.

Let's study this for a moment. Consider the form of the covariance matrix:

$$\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

Let's consider the variance along some direction specified by unit vector $\mathbf{v}$, which is given by $\mathbf{v}^T\mathbf{C}\mathbf{v}$. Suppose that $\mathbf{v}$ is orthogonal to the principal subspace, meaning it is given by the linear combination of some eigenvectors that were discarded. In this case, we have $\mathbf{v}^T\mathbf{U} = \mathbf{0}$ and hence $\mathbf{v}^T\mathbf{C}\mathbf{v} = \sigma^2$. This means that the model just predicts noise (just the variance) for directions orthogonal to the principal subspace. This variance is exactly the mean of the discarded eigenvalues.

Now instead suppose that $\mathbf{v} = \mathbf{u}_i$, where $\mathbf{u}_i$ is one of the retained eigenvectors. Then $\mathbf{v}^T\mathbf{C}\mathbf{v} = (\lambda_i - \sigma^2) + \sigma^2 = \lambda_i$. This means that we'll use exactly the variance of the

data (eigenvalue) along the principal axes, and approximate the variance in all remaining directions with a single average variance σ^2.

Using the maximum likelihood solution to solve for our parameters, we can write a given point $\mathbf{x}$ in the low dimensional space as $\mathbf{z}$ by taking its expectation according to the model parameters:

$$\mathbb{E}[\mathbf{z}|\mathbf{x}] = \mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x} - \bar{\mathbf{x}})$$

where $\mathbf{M}$ is a $M \times M$ matrix defined as

$$\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}$$

We make this substitution since we need to invert $\mathbf{M}$ and it is easier to invert a smaller matrix than the potentially very large matrix $\mathbf{C}$ (a $D \times D$ matrix).

This algorithm is called Probabilistic PCA.