CS28010 Homework 2

Guoxin SUI

November 1, 2017

1 Minimizing error pt2

1.1 Principal component analysis

Suppose we have N data points $x_i, i = 1, ..., N$, where each x_i is a d-dimensional vector $x_i = [x_{i,1}, x_{i,2}, ...x_{i,d}]^T$. You are required find a single line that best represents these N points. We assume $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = 0$. And we use w to denote the direction of the line and $||w|| = 1$. Please find w using the data. (Knowledge about eigen vector, matrix derivatives might be required to finish this problem)

The length of the projection of x_i onto the line is given by $x_i^T w$. The line that best repsesents these N points maximizes the variance of the projections.

So, we look for the w so as to maximize :

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N}({x_i}^T w)^2 &= \frac{1}{N}\sum_{i=1}^{N}(w^T x_i {x_i}^T w) \\ &= w^T \left(\frac{1}{N}\sum_{i=1}^{N}(x_i {x_i}^T) \right) w \end{aligned}$$

The maximization of this subject to $||w|| = 1$ gives the principal eigenvector of $\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i {x_i}^T)$, which is just the empirical covariance matrix of the data.

2 Factor analysis

2.1 Distribution of observed data

Suppose we observe N data points $x_i, i = 1, ..., N$, where each x_i is a d-dimensional vector $x_i = [x_{i,1}, x_{i,2}, ...x_{i,d}]^T$. In order to explain the inner relationship of these data, we specify some factors $y_i, i = 1, ..., N$, where each y_i is a m-dimensional vector $y_i = [y_{i,1}, y_{i,2}, ...y_{i,m}]^T$. We denote the linear relationship between x_i and y_i to be $x_i = Ay_i + \epsilon_i$, where A_{d*m} is a matrix and ϵ_i is the error term. Suppose $y \sim \mathcal{N}(\mu, \Lambda)$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, $E(y\epsilon^T) = 0$. Please compute the marginal distribution of observed data $q(x)$ and the conditional distribution $q(x|y)$.

3 Optional summary work

Please use your words to explain in the settings above, how many matrix A satisfy the condition $x = Ay + \epsilon$.

In probability theory