CS28010 Homework 3

Guoxin SUI

November 22, 2017

1 Factor analysis

1.1 Linear factor analysis

We denote the observed data as x, the latent factor as y and the error as ϵ. Suppose $y \sim \mathcal{N}(\mu, \Lambda)$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I), E(y\epsilon^T) = 0$, where A is an $n{*}m$ matrix, n is the dimension of x, m is the dimension of y and $m < n$. Please explain why there is more than one solution that satisfy $E(xx^T) = A\Lambda A^T + \Sigma$. When Σ is not a general positive definite matrix, but a diagonal matrix, how many solution exists? And if $\Sigma = \sigma 2I$, how many solution exists?
There are 4 uncertainties in factor analysis model:

- **rotation uncertainty**: since the covariance matrix Λ of y is diagonal, there is no rotation uncertainty for y.
- **scale uncertainty**: $y \sim N(y|\mu, \Lambda)$, which means $y \in \mathbb{R}^m$. Hence there are always scale uncertainty for y.
- **addition uncertainty**: when Σ is a general positive definite matrix or a diagonal matrix, there exists addition uncertainty. But there is no addition uncertainty when $\Sigma = \sigma^2\mathbf{I}$.
- **dimension uncertainty**: as the dimension of y is m, there is no dimension uncertainty.

Even if assuming $\Sigma = \sigma^2\mathbf{I}$ can cancel the **addition uncertainty**, there are always **rotation uncertainty** and **scale uncertainty** in $A\Lambda A^T$. So there are always multiple solutions for all cases.

1.2 Binary factor analysis

If y is a latent factor where each dimension is an independent variable that subjects to a different Bernoulli distribution, what are the answers to the above three questions?
If y_i subjects to Bernoulli distribution, then for each dimension of y,

$$y_i \in \{0, 1\}$$

Hence there are no longer exists **rotation uncertainty** and **scale uncertainty**, since either rotate matrix or scale matrix will cause $y_i \notin \{0, 1\}$. Therefore,

- if $Sigma$ is not a general positive definite matrix, but a diagonal matrix, there still are **addition uncertainty**. So there are multiple solutions.

- if $\Sigma = \sigma^2\mathbf{I}$, there are no uncertainty. Hence we can get the only one solution.

$$\begin{bmatrix} \alpha_1^T \\ \alpha_2^T \\ \vdots \\ \alpha_k^T \end{bmatrix}$$

2 Projection

2.1 Orthogonal projection

Suppose we have a hyperplane whose orthogonal basis are $\alpha_1, \alpha_2, ..., \alpha_k, k < n$. Now we have a n-dimensional vector x and we want to apply an orthogonal projection on the hyperplane. Please compute the corresponding projection matrix P.
Assuming all vector is cloumn vector($v^T = [v_1, v_2, \cdots, v_n]$)

Denote $\mathbf{A} = [\alpha_1, \alpha_2, \cdots, \alpha_k]$ is a $n \times k$ matrix. Denote the orthogonally projected vector as $\hat{\mathbf{x}}$. Since we have $\alpha_i^T(\mathbf{x} - \hat{\mathbf{x}}) = 0$, then

$$\begin{aligned} \mathbf{A}^T(\mathbf{x} - \hat{\mathbf{x}}) &= \mathbf{0} \\ \mathbf{A}^T\mathbf{x} &= \mathbf{A}^T\hat{\mathbf{x}} \end{aligned}$$

we also have that

$$\hat{\mathbf{x}} = c_1\alpha_1 + c_1\alpha_2 + \cdots + c_k\alpha_k = \mathbf{A}\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix} = \mathbf{Ac}$$

Hence,

$$\begin{aligned} \mathbf{A}^T\mathbf{x} &= \mathbf{A}^T\hat{\mathbf{x}} \\ &= \mathbf{A}^T\mathbf{Ac} \end{aligned}$$

so we get $\mathbf{c} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{x}$ Then

$$\hat{\mathbf{x}} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{x}$$

Therefore, the corresponding projection matrix $\mathbf{P}$ is:

$$\mathbf{P} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$$

3 Clustering

3.1 Comparison between Gaussian mixture model and k-means

Please add constraints to Gaussian mixture model so that it degenerates into k-means algorithm.
Given a training set $x^{(1)}, ..., x^{(m)}$.
K-means:

1. Initialize cluster centroids $\mu_1, \mu_2, ...\mu_k \in R^n$
2. Repeat :
   (a) For every i, set $c^{(i)} := argmin||x^{(i)} - \mu_j||$
   (b) For every j, set $\mu_j := \frac{\sum_{i=1}^m 1\{c^{(i)}=j\}x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)}=j\}}$

EM Algorithm for Gaussian mixture model (GMM):

1. For each i, j, set $w_i^{(i)} := p(z^{(i)=j|x^{(i)}}; \phi, \mu, \Sigma)$
2. M-step : Update the parameters ϕ, μ, Σ

We find that GMM is reminiscent of the K-means clustering algorithm, except that instead of the "hard" cluster assignments $c(i)$(assign a point to a cluster centroid), we instead have the "soft" assignmetns $w_j^{(i)}$(calculate the possibility that a point belongs to each seperated Gaussian model).
To make these two precesses the same :

1. All the single Gaussian models have the same variance σ, such that the maximum possibility that a point belongs to a single Gaussian model depends only on the distance $x^{(i)} - \mu_j$, which is the same as in K-means ;
2. The variance σ tends to be 0, such that $w_i^{(i)}$ tends to have only two values $0, 1$, the "soft" assignment becomes a "hard" asignment. (This condition covers the first condition)

4 Optional summary work

Please compare PCA, FA and ICA.
PCA: Principal Components Analysis project the variables to a lower dimension basis by eigenvector calculation to remove the redundancy.
FA: Factor Analysis is based on a probabilistic model. In a FA model, we imagine that each datapoint is generated by sampling a low dimension multivariate Gaussian and then map it to a high dimension multivaraite Guassian by a linear transform with a noise. The transform of dimension solves the problem that the training set size is significantly smaller than the dimension of the data.
ICA: Independent Components Analysis will also find w new basis in which to represent the data, but the goal is to seperate the independent components by finding the mixing matrix.