PRML chapter12

1. Principal Component Analysis
   (a) Maximum variance formulation
   - Our goal is to project data D onto a space having a dimentionality M
   - each data point $\mathbf{x}_n$ is projected onto $\mathbf{u}_1^{\mathrm{T}}\mathbf{x}_n$
   - mean of the projected data is $\mathbf{u}_1^{\mathrm{T}}\overline{\mathbf{x}}$
   - the variance of the projected data is $\frac{1}{N}\sum_{n=1}^{N}\left\{\mathbf{u}_1^{\mathrm{T}}\mathbf{x}_n - \mathbf{u}_1^{\mathrm{T}}\overline{\mathbf{x}}\right\}^2 = \mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1$
   - maximize $\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1$with respect to u

$$\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 + \lambda_1\left(1 - \mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1\right)$$

   - the variance will be a maximum when we set u equal to the maximum eigen vector
   - principal component analysis requires the mean and variance, also needs largest M eigenvectors

   (b) Minimum-error formulation
   - Introduce a complete orthonormal set of D-dimensional vectors u

$$\mathbf{x}_n = \sum_{i=1}^{D}\left(\mathbf{x}_n^{\mathrm{T}}\mathbf{u}_i\right)\mathbf{u}_i$$

   - we approximate the data

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

   - our goal is to minimize

$$J = \frac{1}{N}\sum_{n=1}^{N}\left\|\mathbf{x}_n - \widetilde{\mathbf{x}}_n\right\|^2$$

   - first, minimize it with respect to z, then minimize it with respect to u

   (c) Applications of PCA
   - comprehension of the data

$$\widetilde{\mathbf{x}}_n = \sum_{i=1}^{M}\left(\mathbf{x}_n^{\mathrm{T}}\mathbf{u}_i\right)\mathbf{u}_i + \sum_{i=M+1}^{D}\left(\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{u}_i\right)\mathbf{u}_i = \overline{\mathbf{x}} + \sum_{i=1}^{M}\left(\mathbf{x}_n^{\mathrm{T}}\mathbf{u}_i - \overline{\mathbf{x}}^{\mathrm{T}}\mathbf{u}_i\right)\mathbf{u}_i$$

   - data pre-processing

   (d) PCA for high-dimensional data
   - if the number of data points is smaller than the dimentionality of the data space, we have to take different approach because the computational cost $O(D^3)$

- take the following algorithms and solve it with $O(N^3)$
- let X be a matrix whose nth row is $(\mathbf{x}_n - \overline{\mathbf{x}})^{\mathrm{T}}$
- under $\mathbf{v}_i = \mathbf{X}\mathbf{u}_i$,

$$\frac{1}{N}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{v}_i = \lambda_i \mathbf{v}_i$$

2. Probabilistic PCA
   - PCA can also be expressed as the maximum likelihood solution of a probabilistic latent variable model
   - First we gave the prior distribution over z as $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$
   - the conditional distribution of observed value x, $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$
   - $P(x) = \int P(x|z)p(z)dz$ provide us the parameters

   (a) Maximum likelihood PCA
   - we want to maximize log likelihood function

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2) = \sum_{n=1}^{N} \ln p\left(\mathbf{x}_n|\mathbf{W}, \boldsymbol{\mu}, \sigma^2\right)$$

$$= -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{C}| - \frac{1}{2}\sum_{n-1}^{N}\left(\mathbf{x}_n - \boldsymbol{\mu}\right)^{\mathrm{T}}\mathbf{C}^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}\right)$$

   - the calculation is very complex, Tipping and Bishop(1999)

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

   - the number of independant parameters are controlled automatically

   (b) EM algorithm for PCA
   - it has an advantage when we treat high-dimentinoal data
   - complete log-likelihood function takes the form

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2) = \sum_{n=1}^{N}\left\{\ln p\left(\mathbf{x}_n|\mathbf{z}_n\right) + \ln p\left(\mathbf{z}_n\right)\right\}$$

   - E Step

$$E\left[\mathbf{z}_n\right] = \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}\left(\mathbf{x}_n - \overline{\mathbf{x}}\right)$$

$$E\left[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right] = \sigma^2\mathbf{M}^{-1} + E\left[\mathbf{z}_n\right]E\left[\mathbf{z}_n\right]^{\mathrm{T}}$$

   - M Step

$$\mathbf{W}_{\mathrm{new}} = \left[\sum_{n=1}^{N}\left(\mathbf{x}_n - \overline{\mathbf{x}}\right)E\left[\mathbf{z}_n\right]^{\mathrm{T}}\right]\left[\sum_{n=1}^{N}E\left[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right]\right]^{-1}$$

$$\sigma^2_{new} = \frac{1}{ND}\sum_{n=1}^{N}\left\{\left\|\mathbf{x}_n - \overline{\mathbf{x}}\right\|^2 - 2E\left[\mathbf{z}_n\right]^{\mathrm{T}}\mathbf{W}_{\mathrm{new}}^{\mathrm{T}}\left(\mathbf{x}_n - \overline{\mathbf{x}}\right) + Tr\left(E\left[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right]\mathbf{W}_{\mathrm{new}}^{\mathrm{T}}\mathbf{W}_{\mathrm{new}}\right)\right\}$$

   (c) Bayesian PCA
   - we want to decide M with Bayesian approach

- choose model with Bayesian approach

(d) Factor analysis

- factor analysis has

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\Psi})$$

- We can determine $\mu, \mathbf{W}, \boldsymbol{\Psi}$ in the factor analysis model by maximum likelihood
- use EM algorithm

3. Kernel PCA

- we want to obtain non-linear gerenalization
- the principal vector is defined as

$$\mathbf{S}\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

where

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}}$$

- sample covariance matrix is

$$\mathbf{C} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\phi}\left(\mathbf{x}_n\right) \boldsymbol{\phi}\left(\mathbf{x}_n\right)^{\mathrm{T}}$$

- eigen vector of matrix C is

$$\mathbf{v}_i = \sum_{n=1}^{N} a_{in} \boldsymbol{\phi}\left(\mathbf{x}_n\right)$$

Kernel function gives us a solution for a by solving the following eigenvalue problem

$$\mathbf{K}\mathbf{a}_i = \lambda_i N \mathbf{a}_i$$

- the projection of x onto eigenvector i is written as

$$y_i(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{v}_i = \sum_{n=1}^{N} a_{in} \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \boldsymbol{\phi}\left(\mathbf{x}_n\right) = \sum_{n=1}^{N} a_{in} k\left(\mathbf{x}, \mathbf{x}_n\right)$$

4. Nonlinear latent value model

- consider the models based on non-linear and non-Gaussian distributions

(a) Independent component analysis

- the example of non-linear latent variables model
- In this models, observed variables are related linearly to the latent variables but the latent distribution is non-Gaussian.
- latent variables are independent so

$$p(\mathbf{z}) = \prod_{j=1}^{M} p\left(z_j\right)$$

- no need to consider the noise, because the number of observed variables and latent variables are same
- The success of this approach requires that the latent variables have non-Gaussian distributions.

(b) Auto associative neural networks

- Create the newral network model whose input and output are D dimention
- The model tells us the features of the dataset.
- tells us more information than PCA but calculation amount is large