Tutorial 6

Thursday 10th October, 2019

1 Principal Component Analysis

Let $\mathbf{X}$ be a random vector and $\Gamma = \mathbf{E}[(\mathbf{X} - \mathbf{E}[\mathbf{X}])(\mathbf{X} - \mathbf{E}[\mathbf{X}])^T]$ its covariance matrix. Let $\mathbf{e}_1, \ldots, \mathbf{e}_n$ be the n (normalized) eigenvectors of Γ.

- The n principal components of $\mathbf{X}$ are said to be $\mathbf{e}_1^T\mathbf{X}$, $\mathbf{e}_2^T\mathbf{X}$, ..., $\mathbf{e}_n^T\mathbf{X}$. See `https://arxiv.org/abs/1804.10253`.
- Let $p(X_1) = \mathcal{N}(0, 1)$ and $p(X_2) = \mathcal{N}(0, 1)$ and $cov(X_1, X_2) = \theta$. Find all the principal components of the random vector $\mathbf{X} = [X_1, X_2]^T$.

  Solution:

  1. We first note that the 2×2 matrix
  $$\Gamma = \begin{bmatrix} 1 & \theta \\ \theta & 1 \end{bmatrix}$$
  2. To compute the eigenvalues of Γ from the characteristic polynomial:
  $$\left| \begin{bmatrix} 1-\lambda & \theta \\ \theta & 1-\lambda \end{bmatrix} \right| = (1-\lambda)^2 - \theta^2 = 0$$
  $\Rightarrow \lambda_1 = 1 + \theta$ and $\lambda_2 = 1 - \theta$ are the two solutions/eigenvalues.
  3. For eigenvalue λ_1 we find its eigenvector v_1:
  $$\begin{bmatrix} 1 & \theta \\ \theta & 1 \end{bmatrix} v_1 = (1+\theta)v_1$$
  which is easily solvable as
  $$v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$
  or its normalized version
  $$v_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

4. Similarly, for eigenvalue λ_2 we find its eigenvector v_2:

$$\begin{bmatrix} 1 & \theta \\ \theta & 1 \end{bmatrix} v_2 = (1-\theta)v_2$$

which is easily solvable as

$$v_2 = \begin{bmatrix} 1 \\ -1/ \end{bmatrix}$$

or its normalized version

$$v_2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

5. Thus, the principal components of $\mathbf{X}$ are $X_1/\sqrt{2} + X_2/\sqrt{2}$ and $X_1/\sqrt{2} - X_2/\sqrt{2}$.

- Now, let $\mathbf{Y} = \mathcal{N}(\mathbf{0}, \Sigma) \in \Re^p$ where $\Sigma = \lambda^2 I_{p\times p} + \alpha^2 ones(p,p)$ for any $\lambda, \alpha \in \Re$. Here, $I_{p\times p}$ is a $p \times p$ identity matrix while $ones(p,p)$ is a $p \times p$ matrix of 1's. Find atleast one principal component of $\mathbf{Y}$.

2 How would you Kernelize PCA?

How would you Kernelize PCA? See Section 14.5.4 of the Tibshirani book posted on moodle.

Solution: Here is a proof sketch, that we also provided in the slides

1. Consider the Singular Value Decomposition of $X = ADB^T$. Then $nS = XX^T = AD^2A^T$ and $X^TX = BD^2B^T$
2. $U = AD$ is the matrix of principal component variables.
3. Consider the kernel/gram matrix $\mathcal{K}$ of the data, in place of $\mathbf{X}^T\mathbf{X}$

$$\mathcal{K} = \begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & ... & K(\mathbf{x}_1, \mathbf{x}_n) \\ ... & K(\mathbf{x}_i, \mathbf{x}_j) & ... \\ K(\mathbf{x}_m, \mathbf{x}_1) & ... & K(\mathbf{x}_m, \mathbf{x}_m) \end{bmatrix}$$

4. Then, $nS = XX^T = AD^2A^T$ and $\mathcal{K} = X^TX = BD^2B^T$ and the projections XU of our data X onto those components $U = AD$ can be computed from the eigendecomposition of $\mathcal{K}$
5. Do eigenvalue decomposition of $\mathcal{K}$. Take the eigenvectors, $\mathbf{v}_1, \ldots, \mathbf{v}_k$.
6. Kernel-PCA computes not the principal components themselves, but the projections of our data onto those components:

- See Extra & Optional slides on Mercer's Theorem and RKHS Kernel for conditions under which $K(\mathbf{x}_i, \mathbf{x}_j)$ will be a **valid Kernel function**
- For more discussion on Kernel PCA, see Section 14.5.4 of the Tibshirani book posted on moodle.

3 EM Algorithm for Mixture of Gaussians (completely optional

Q: Show that the following algorithm for estimating the mean μ_i, the covariance matrix Σ_i and mixture components π_i for a mixture of Gaussians is an instance of the general EM algorithm

ANSWER:

Initialize $\mu_i^{(0)}$ to different random values and $\Sigma_i^{(0)}$ to I. Now iterate between the following **E Step** and **M Steps**:

E Step:

1. For the posterior $p(z_i \mid \phi(x_j), \mu, \Sigma)$

$$p^{(t+1)}(z_i \mid \phi(x_j), \theta) = \frac{\pi_i \mathcal{N}\left(\phi(x); \mu_i^{(t)}, \Sigma_i^{(t)}\right)}{\sum_{l=1}^{K} \pi_l \mathcal{N}\left(\phi(x); \mu_l^{(t)}, \Sigma_l^{(t)}\right)}$$

M Steps:

1. For the prior π_i

$$\pi_i^{(t+1)} = \frac{1}{n} \sum_{j=1}^{n} p^{(t+1)}(z_i \mid \phi(x_j), \theta)$$

2. For μ_i

$$\mu_i^{(t+1)} = \frac{\sum_{j=1}^{n} p^{(t+1)}(z_i \mid \phi(x_j), \theta) \phi(x_j)}{\sum_{j=1}^{n} p^{(t+1)}(z_i \mid \phi(x_j), \theta)}$$

3. For Σ_i

$$\Sigma_i^{(t+1)} = \frac{\sum_{j=1}^{n} p^{(t+1)}(z_i \mid \phi(x_j), \theta) \left(\phi(x_j) - \mu_i^{(t+1)}\right) \left(\phi(x_j) - \mu_i^{(t+1)}\right)^T}{\sum_{j=1}^{n} p^{(t+1)}(z_i \mid \phi(x_j), \theta)}$$

Q: Note that this algorithm is for the Mixture of Gaussians assuming a difference covariance matrix Σ_i for each class C_i. What will be the algorithm like, if we assume a shared covariance matrix Σ across all classes (that is, the Linear Discriminant Analysis discussed in Section 1.2)?

ANSWER: We will simply build on the solution to the Linear Discriminant case from Section 2.1 and simply replace multiple class-specific estimates Σ_i with a single estimate Σ:

4 Convergence of Hard K-Means Algorithm

Prove the following claim: The K-Means Clustering algorithm will converge in a finite number of iterations.

1. **Proof Sketch:** At each iteration, the K-Means algorithm reduces the objective $\sum_{j=1}^{m}\sum_{l=1}^{K} P_{l,j}\|\phi\left(\mathbf{x}^{(j)}\right) - \mu_l\|^2$ and stops when this objective does not reduce any further.

2. Hint1: $P^{(t+1)} = \underset{P}{\operatorname{argmin}} \sum_{j=1}^{m}\sum_{l=1}^{K} P_{l,j}\|\phi\left(\mathbf{x}^{(j)}\right) - \mu_l^{(t)}\|^2$

3. Hint2: $\mu^{(t+1)} = \underset{\mu}{\operatorname{argmin}} \sum_{j=1}^{m}\sum_{l=1}^{K} P_{l,j}^{(t+1)}\|\phi\left(\mathbf{x}^{(j)}\right) - \mu_l\|^2$

4. Hint3: Only a finite number of combinations of $P_{i,j}$ are possible.