CS420 Machine Learning – Assignment 1

Xiang Gu
5130309729

1 PCA Algorithms

- Eigendecompose the sample covariance matrix: One straightforward solution is to calculate all eigenvectors and eigenvalues of the sample covariance matrix $\bar{\mathbf{\Sigma}} = \frac{1}{N}\mathbf{X}\mathbf{X}^T = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T$ (assuming all data points $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ are de-centralized) through eigendecomposition. Formally, $\bar{\mathbf{\Sigma}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ and the eigenvalues in $\mathbf{\Lambda}$ are in descending order. Then the first column of $\mathbf{V}$ is the first principal component.
- Or, Apply Singular Value Decomposition (SVD) to the data matrix $\mathbf{X}$: We can directly apply SVD on the data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_1, \ldots, \mathbf{x}_N] = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Then the first column of $\mathbf{U}$ (left singular vector matrix) is the first principal component.

2 Factor Analysis (FA)

According to the Bayes rule, the conditional probability density function of random vector $\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_p]^T$ given random vector $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_q]^T$ is

$$\begin{aligned} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) &= \frac{f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})f_{\mathbf{Y}}(\mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})} \\ &= \frac{N_q(\mathbf{x}|\mathbf{A}\mathbf{y} + \boldsymbol{\mu}, \mathbf{\Sigma}_e)N_p(\mathbf{y}|\mathbf{0}, \mathbf{\Sigma}_y)}{f_{\mathbf{AY}+\boldsymbol{\mu}+\mathbf{E}}(\mathbf{A}\mathbf{y} + \boldsymbol{\mu} + \mathbf{e})} \end{aligned} \tag{1}$$

Now we know what is hard is the denominator – determining the distribution of a linear combination of two multivariate normal distribution $\mathbf{Y}$ and $\mathbf{E}$ and a constant vector $\boldsymbol{\mu}$. I looked up online at *this lecture notes* and figured out the answer to be:

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \frac{N_q(\mathbf{x}|\mathbf{A}\mathbf{y} + \boldsymbol{\mu}, \mathbf{\Sigma}_e)N_p(\mathbf{y}|\mathbf{0}, \mathbf{\Sigma}_y)}{N_q(\boldsymbol{\mu} + \boldsymbol{\mu}_e, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T + \mathbf{\Sigma}_e)} \tag{2}$$

Usually speaking, the mean of the noise vector is zero $\boldsymbol{\mu}_e = \mathbf{0}$.

3 Independent Component Analysis (ICA)

First we look at the central limit theorem (CLT). CLT is basically concerned with the tendency of estimations of the mean of independently drawn variables of any arbitrary distribution to follow a Gaussian distribution. This matters because in real world samples we often observe data that is in fact a composite of many underlying factors, and based on CLT we understand that linear combination of independent variables create an aggregate variable that tends to be Gaussian in nature.

In independent component analysis (ICA), we want to separate independent factors that underlie the data (which is Gaussian-like according to CLT) – reversing the CLT. Since the linear combination of independent variables is more Gaussian than the original variables, it follows that non-Gaussianity is required to identify the underlying independent variables.

4 Feature Reduction with FA

In this section, we followed the assignment and experimented feature reduction with FA model.

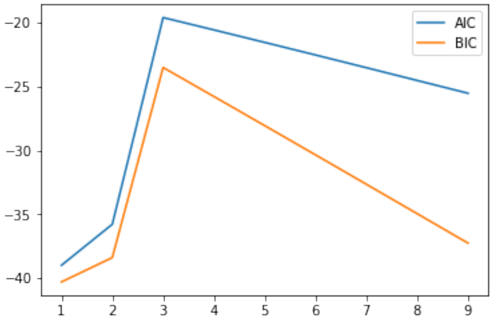

Figure 1: Results of AIC and BIC criteria on the feature reduction process. Both criteria reach their highest value at the dimensionality same as that of the true underlying latent variable (3 in this case).

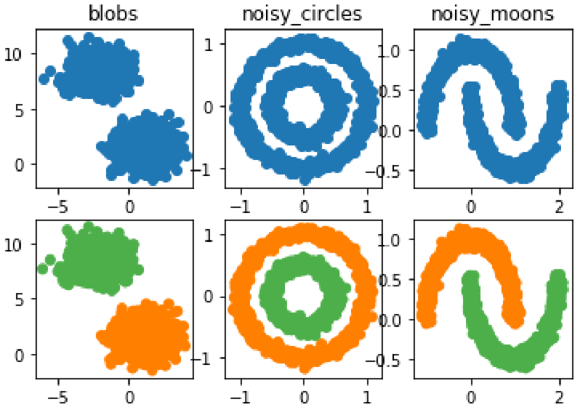

Figure 2: Result of examples when spectral clustering works: a (easy) blob-like clusters, a circle-like clusters, and a moon-like clusters

We first generate a dataset in the exactly same way as described in the assignment:

1. Randomly sample N $\mathbf{y}'s$ from Gaussian density $G(\mathbf{y}|\boldsymbol{\mu}, \mathbf{I})$, with $N = 100$, $\mathbf{dim}(\mathbf{y}) = m = 3, \boldsymbol{\mu} = \mathbf{0}$;
2. Randomly sample N noises vectors $\mathbf{e}'s$ from Gaussian density $G(\mathbf{e}|\mathbf{0}, \sigma^2\mathbf{I})$, with $\sigma^2 = 0.1, \mathbf{e} \in \mathbb{R}^n, n = 10$
3. Get N observable variables $\mathbf{x} = \mathbf{A}\mathbf{y} + \mathbf{e}$

Then we apply FA model with various dimensionalities for the latent variable $1, 2, ..., M$, where $M = 9$.

The result is shown in figure 1. We can clearly see that when $m = 3$, which is also the same as the dimensionality of the true underlying latext variable, both criteria reaches their highest value.

5 Spectral Clustering

- When it works: I experimented three different cluster data – blobs, circle-like clusters and moon-like clusters. I referred to *this scikit-learn page*. See figure 2 for results.
- When it fails: I looked at *this research paper* and picked two examples from it that the author demonstrated scenarios where spectral clustering fails to find the correct cluster. See figure 3 and 4 for the results (I copypasted the results from the original paper).

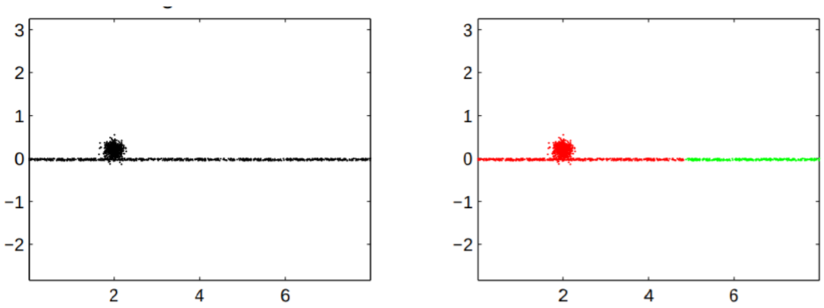

Figure 3: Result of example one that spectral clustering fails. The data is generated from a mixture of a uniform strip distribution and a Guassian distribution. See the code for detailed parameters

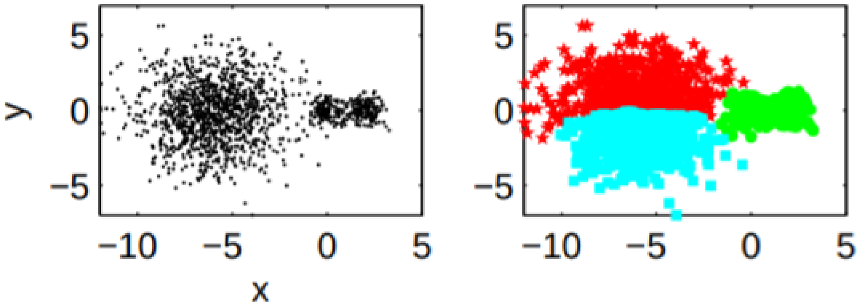

Figure 4: Result of example two that spectral clustering fails. The data is generated from a mixture of three Gaussian distribution. See the code for detailed parameters