IB120/201 - Lab 13

Clustering Pt. 2

Due Date: May 1, 2020

University of California, Berkeley *GSI: Naveed Ziari*

We will further our understanding of clustering method in this lab by implementing principal component analysis and the expectation-maximization algorithm

Background

Principal Component Analysis

Principal component analysis (PCA) is linear transformation of data that produces orthogonal bases that contain the highest variance. Each basis, or axis, is referred to as a *principal component*. The main advantage of this method is that it accounts for correlated features in the dataset. If, for instance, the first principal component explains 80% of the variance in a high-throughput dataset, then it provides insight into the quality and information contained within the dataset. In regards to machine learning, PCA reduces computational complexity and mitigates the problem overfitting by eliminating redundancies.

Expectation-Maximization

The expectation-maximization (EM) algorithm is an iterative method of maximum likelihood estimation of the parameters in a statistical model. In many cases, models are contingent upon latent (hidden) variables. In each iteration, the algorithm performs an *expectation* step (E), whereby computes the expectation of the likelihood for the current parameter values, and a *maximization* step (M), which updates the parameter values that maximizes the liklihood function in the E step. Expectation-maximization differs from other traditional ML algorithms because it maximizes a likelihood function rather than minimizing a cost function.

If we have a model with observable data $\mathbf{X}$ and an observed (latent) label $\mathbf{Z}$ parameterized by $\boldsymbol{\theta}$, then the likelihood is equal to the following expression:

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{X}) = \mathbb{P}(\mathbf{X}|\boldsymbol{\theta}) = \int \mathbb{P}(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) d\mathbf{Z}$$

Obtaining a value for $\mathbb{P}(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})d\mathbf{Z}$ is oftentimes not possible, so it necessitates and the expectation-maximization iterative procedure. The expectation step defines $\mathcal{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right)$ as the expected value of the log-likelihood function for $\boldsymbol{\theta}$.

$$\mathcal{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right) = \mathbb{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(t)}}[\log \mathcal{L}(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})]$$

In the maximization step, select the parameters that maximize the quantity in the expectation step. Repeat this procedure until convergence.

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\theta} \mathcal{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right)$$

Questions

Please submit your assignment in the form of a iPython notebook similar to lecture. Each question has its own block.

1. Perform PCA on the `iris` dataset found in `lab_13_data_01.xlsx`, with all the features excluding `Species`.

2. Visualize the data on a scatter plot with dimensions as the principal components and color each data point according to its specie. What can you conclude about the efficacy of PCA on this dataset? *hint: think of explained variance of each principal component.*

3. Load the dataset `lab_13_data_02.xlsx` and use a Gaussian mixture model from the `sklearn` package to estimate the means of the dataset. You may want to visualize the dataset first to see how many components it has.

4. What does the `.weights_` attribute of a `GaussianMixture` object mean?

5. Rotate the PCA matrix by 30 degrees counterclockwise. You need to specify a rotation matrix (provided below) and perform matrix multiplication. Explain quantitatively if your transformed PCA matrix still has principal components.

$$\mathbf{R} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

6. How do eigenvectors come into play when doing rotations and projections? And how about when performing PCA?

7. Name an algorithm frequently used in computational biology that is a special case of the EM algorithm or that uses the EM technique.

8. How is EM different than K-means clustering? *hint: think about probabilities*