SOUTH UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA

Intro to Big Data Science: Assignment 6

Due Date: May 19, 2022

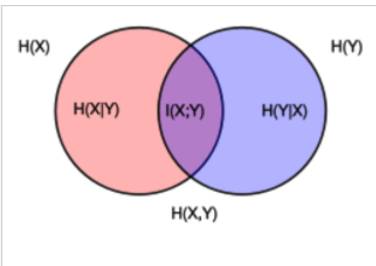
✏ **Exercise 1**

Log into "cookdata.cn", and enroll the course "数据科学导引". Finish the online exercise there.

✏ **Exercise 2** Recall the definition of information entropy, $H(P) = -\sum_{i=1}^{n} p_i \log p_i$, which means the maximal information contained in probability distribution P. Let X and Y be two random variables. The entropy $H(X,Y)$ for the joint distribution of (X,Y) is defined similarly. The conditional entropy is defined as:

$$\begin{aligned} H(X|Y) &= -\sum_j P(Y = y_j) H(X|Y = y_j) \\ &= -\sum_j P(Y = y_j) (\sum_i P(X = x_i|Y = y_j) \log P(X = x_i|Y = y_j)) \end{aligned}$$

1. Show that $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$.
2. The mutual information (information gain) is defined as $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$. Show that if X and Y are independent, then $I(X;Y) = 0$



3. Define the Kullback-Leibler divergence as $D_{KL}(P\|Q) = -\sum_{i=1}^{n} p_i \log \frac{q_i}{p_i}$. Show that $I(X;Y) = D_{KL}(p(X,Y)\|p(X)p(Y))$.
4. (Optional) Furthermore, show that $D_{KL}(P\|Q) \geqslant 0$ for any P and Q by using Jensen's inequality. As a result, $I(X;Y) \geqslant 0$.

✏ **Exercise 3** (EM Algorithm, you may need to carefully read Section 8.5.2 in the book "Elements of Statistical Learning" before solving this problem)

Imagine a class where the probability that a student gets an "A" grade is $\mathbb{P}(A) = \frac{1}{2}$, a "B" grade is $\mathbb{P}(B) = \mu$, a "C" grade is $\mathbb{P}(C) = 2\mu$, and a "D" grade is $\mathbb{P}(D) = \frac{1}{2} - 3\mu$. We are told that c students get a "C" and d students get a "D". We don't know how many students got exactly an "A" or exactly a "B". But we do know that h students got either an "A" or "B". Let a be the number of students getting "A" and b be the number of students getting "B". Therefore, a and b are unknown parameters with $a + b = h$. Our goal is to use expectation maximization to obtain a maximum likelihood estimate of μ.

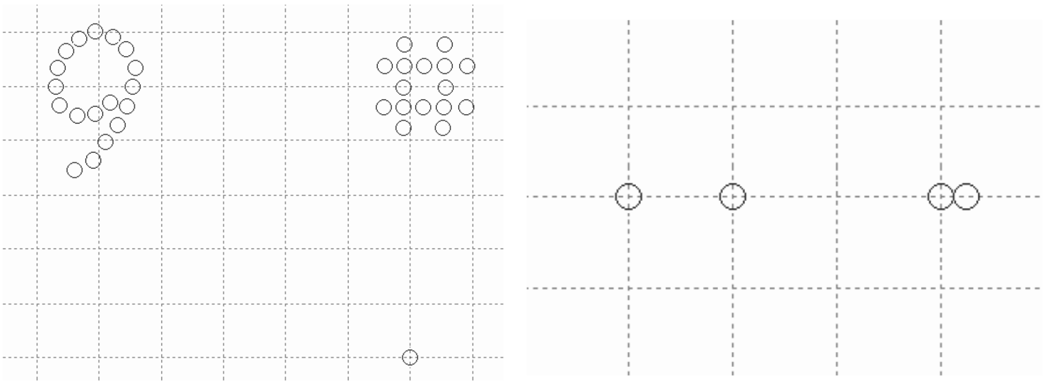
1. Use Multinoulli distribution to compute the log-likelihood function $l(\mu, a, b)$.
2. Expectation step: Given $\hat{\mu}^{(m)}$, compute the expected values $\hat{a}^{(m)}$ and $\hat{b}^{(m)}$ of a and b respectively.
3. Maximization step: Plug $\hat{a}^{(m)}$ and $\hat{b}^{(m)}$ into the log-likelihood function $l(\mu, a, b)$ and calculate for the maximum likelihood estimate $\hat{\mu}^{(m+1)}$ of μ, as a function of $\hat{\mu}^{(m)}$.
4. Iterating between the E-step and M-step will always converge to a local optimum of μ (which may or may not also be a global optimum)? Explain why in short.

✏ **Problem 4** (Spectral Clustering)

1. We consider the 2-clustering problem, in which we have N data points $x_{1:N}$ to be grouped in two clusters, denoted by A and B. Given the N by N affinity matrix W (**Remember that in class we define the affinity matrix in the way that the diagonal entries are zero for undirected graphs**), consider the following two problems:
   - Min-cut: minimize $\sum\limits_{i\in A}\sum\limits_{j\in B} W_{ij}$;
   - Normalized cut: minimize $\frac{\sum_{i\in A}\sum_{j\in B} W_{ij}}{\sum_{i\in A}\sum_{j=1}^{N} W_{ij}} + \frac{\sum_{i\in A}\sum j\in B W_{ij}}{\sum_{i=1}^{N}\sum_{j\in B} W_{ij}}$.

   a) The data points are shown in Figure (a) above. The grid unit is 1. Let $W_{ij} = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}$, give the clustering results of min-cut and normalized cut respectively (Please draw a rough sketch and give the separation boundary in the answer book).

   b) The data points are shown in Figure (b) above. The grid unit is 1. Let $W_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}}$, describe the clustering results of min-cut algorithm for $\sigma^2 = 50$ and $\sigma^2 = 0.5$ respectively (Please draw a rough sketch and give the separation boundaries for each case of σ^2 in the answer book).



(a) (b)

2. Now back to the setting of the 2-clustering problem shown in Figure (a). The grid unit is 1.

   a) If we use Euclidean distance to construct the affinity matrix W as follows:

   $$W_{ij} = \begin{cases} 1, & \text{if } \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \leqslant \sigma^2; \\ 0, & \text{otherwise.} \end{cases}$$

   What σ^2 value would you choose? Briefly explain.

   b) The next step is to compute the first $k = 2$ dominant eigenvectors of the affinity matrix W. For the value of σ^2 you chose in the previous question, can you compute analytically the eigenvalues corresponding to the first two eigenvectors? If yes, compute and report the eigenvalues. If not, briefly explain.

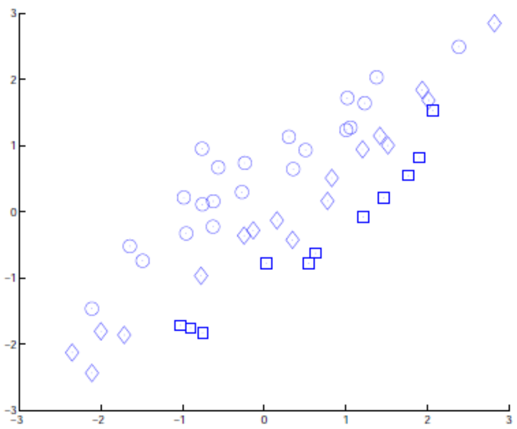
✏ **Exercise 5** (Dimensionality Reduction)

1. (PCA vs. LDA) Plot the directions of the first PCA (plot (a)) and LDA (plot (b)) components in the following figures respectively.

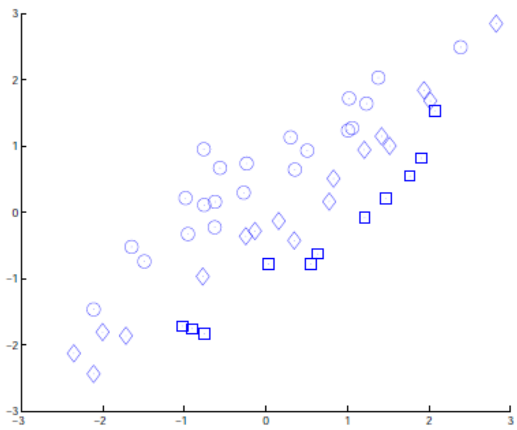
2. (PCA and SVD) Given 6 data points in 5D space, $(1,1,1,0,0)$, $(-3,-3,-3,0,0)$, $(2,2,2,0,0)$, $(0,0,0,-1,-1)$, $(0,0,0,2,2)$, $(0,0,0,-1,-1)$. We can represent these data points by a 6×5 matrix $\mathbf{X}$, where each row corresponds to a data point:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ -3 & -3 & -3 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & -1 & -1 \end{pmatrix}$$

   a) What is the sample mean of the data set?



(c) First PCA component



(d) First LDA component

b) What is the SVD of the data matrix $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ satisfy $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_2$? Note that the SVD for this matrix must take the following form, where a, b, c, d, σ_1, σ_2 are the parameters you need to decide.

$$\mathbf{X} = \begin{pmatrix} a & 0 \\ -3a & 0 \\ 2a & 0 \\ 0 & b \\ 0 & -2b \\ 0 & b \end{pmatrix} \times \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \times \begin{pmatrix} c & c & c & 0 & 0 \\ 0 & 0 & 0 & d & d \end{pmatrix}$$

c) What is first principle component for the original data points?

d) If we want to project the original data points $\{\mathbf{x}_i\}_{i=1}^6$ into 1D space by principle component you choose, what is the sample variance of the projected data $\{\hat{\mathbf{x}}_i\}_{i=1}^6$?

e) For the projected data in d), now if we represent them in the original 5-d space, what is the reconstruction error $\frac{1}{6}\sum_{i=1}^6 \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2$?

✏ **Exercise 6**(PCA as factor analysis and SVD, optional)

PCA of a set of data in $\mathbb{R}^p$ provide a sequence of best linear approximations to those data, of all ranks $q \leqslant p$. Denote the observations by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, and consider the rank-$q$ linear model for representing them

$$f(\alpha) = \mu + \mathbf{V}_q \alpha$$

where μ is a location vector in $\mathbb{R}^p$, V_q is a $p \times q$ matrix with q **orthogonal unit vectors** as columns, and α is a q vector of parameters. If we can find such a model, then we can reconstruct each $\mathbf{x}_i$ by a low dimensional coordinate vector α_i through

$$\mathbf{x}_i = f(\alpha_i) + \epsilon_i = \mu + \mathbf{V}_q \alpha_i + \epsilon_i \tag{1}$$

where $\epsilon_i \in \mathbb{R}^p$ are noise terms. Then PCA amounts to minimizing this reconstruction error by least square method

$$\min_{\mu,\{\alpha_i\},\mathbf{V}_q} \sum_{i=1}^{N} \|\mathbf{x}_i - \mu - \mathbf{V}_q \alpha_i\|^2$$

1. Assume $\mathbf{V}_q$ is known and treat μ and α_i as unknowns. Show that the least square problem

$$\min_{\mu,\{\alpha_i\}} \sum_{i=1}^{N} \|\mathbf{x}_i - \mu - \mathbf{V}_q \alpha_i\|^2$$

is minimized by

$$\hat{\mu} = \bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i, \tag{2}$$

$$\hat{\alpha}_i = \mathbf{V}_q^T(\mathbf{x}_i - \bar{\mathbf{x}}). \tag{3}$$

Also show that the solution for $\hat{\mu}$ is not unique. Give a family of solutions for $\hat{\mu}$.

2. For the standard solution (2), we are left with solving

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} \|(\mathbf{x}_i - \bar{\mathbf{x}}) - \mathbf{V}_q\mathbf{V}_q^T(\mathbf{x}_i - \bar{\mathbf{x}})\|^2 = \min_{\mathbf{V}_q} \mathrm{Tr}\Big(\tilde{\mathbf{X}}(\mathbf{I}_p - \mathbf{V}_q\mathbf{V}_q^T)\tilde{\mathbf{X}}^T\Big). \tag{4}$$

Here we introduce the centered sample matrix

$$\tilde{\mathbf{X}} = (\mathbf{I}_N - \frac{1}{N}\mathbf{J}_N)\mathbf{X} = \begin{pmatrix} (\mathbf{x}_1 - \bar{\mathbf{x}})^T \\ \vdots \\ (\mathbf{x}_N - \bar{\mathbf{x}})^T \end{pmatrix} \in \mathbb{R}^{N\times p}$$

where $\mathbf{I}_N$ is $N\times N$ identity matrix and $\mathbf{J}_N$ is a matrix whose entries are all 1's. Recall the singular value decomposition (SVD) in linear algebra: $\tilde{\mathbf{X}} = \mathbf{U}\mathbf{D}\mathbf{V}^T$. Here $\mathbf{U}$ is an $N\times p$ orthogonal matrix ($\mathbf{U}^T\mathbf{U} = \mathbf{I}_p$) whose columns $\mathbf{u}_j$ are called the left singular vectors; $\mathbf{V}$ is a $p\times p$ orthogonal matrix ($\mathbf{V}^T\mathbf{V} = \mathbf{I}_p$) with columns $\mathbf{v}_j$ called the right singular vectors, and $\mathbf{D}$ is a $p\times p$ diagonal matrix, with diagonal elements $d_1 \geqslant d_2 \geqslant \cdots \geqslant d_p \geqslant 0$ known as the singular values.

Show that the solution $\mathbf{V}_q$ to problem (4) consists of the first q columns of $\mathbf{V}$. (Then the optimal $\hat{\alpha}_i$ are given by the i-th row with the first q columns of $\mathbf{UD}$.)

Remark: The model (1), in general, gives the factor analysis in multivariate statistics:

$$\mathbf{x} = \mu + \mathbf{V}_q\alpha + \epsilon$$

In traditional factor analysis, α_j with $j = 1,\dots,q$ is assumed to be Gaussian and uncorrelated as well as ϵ_i with $i = 1,\dots,p$. However, Independent Component Analysis (ICA) instead assumes α_j with $j = 1,\dots,q$ is assumed to be non-Gaussian and independent. Because of the independence, ICA is particularly useful in separating mixed signals.