CS 7641 CSE/ISYE 6740 Homework 1

Rakesh Surapaneni

September 23, 2015

1 Probability [15 pts]

(a) Stores A, B, and C have 50, 75, and 100 employees and, respectively, 50, 60, and 70 percent of these are women. Resignations are equally likely among all employees, regardless of stores and sex. Suppose an employee resigned, and this was a woman. What is the probability that she has worked in store C? [5 pts]

Solution: Assumptions and definitions:

- Let W denote the event that the person resigning is a Woman. P(W) is probability that a woman resigns.
- Let 'a' denotes an event that Store A employee resigns. Similarly events 'b' and 'c' are defined.
- P(W/a) = 0.5 (since there are 50% of employees in store A is Woman and resignations are equally likely irrespective of stores and sex).
- Similarly P(W/b) = 0.6 and P(W/c) = 0.7
- $$P(a) = \frac{50}{50+75+100} = \frac{50}{225}$$
- Similarly P(b) = 75/225, P(c) = 100/225

We have to calculate P(c/W) (Probability that store C employee resigns given a Woman resigns)

$$P(W) = \sum_{x=a,b,c} P(W/x)P(x) = P(W/a)*P(a)+P(W/b)*P(b)+P(W/c)*P(c) = \frac{0.5*50+0.6*75+0.7*100}{225} = \frac{140}{225}$$

$$P(c/W) = \frac{P(W/c)P(c)}{P(W)} = \frac{0.7*\frac{100}{225}}{\frac{140}{225}} = \frac{70}{140} = 0.5$$

Hence if an employee resigned, and this was a woman, the probability that she has worked in store C is **0.5**.

(b) A laboratory blood test is 95 percent effective in detecting a certain disease when it is, in fact, present. The test also yields a false positive result for 1 percent of the healthy persons tested. That is, if a healthy person is tested then with probability 0.01 the test result will imply he has the disease. If 0.5 percent of the population actually has the disease, what is the probability a person has the disease given that his test result is positive? [5 pts]

Solution: Assumptions and definitions:

- Let x denote the event that the result is positive.
- y denote the event that the result is negative.
- D denote the event that there is disease and N denotes that there is no disease.
- We are given that P(x/D) = 0.95 (accuracy in case of disease).
- P(x/N) = 0.01 (false positives).
- P(D) = 0.005 P(N) = 0.995 (0.5% have the disease).
- We have to find **P(D/x)**.

$$\begin{aligned} P(x) &= P(xandD) + P(xandN) \\ &= P(D) * P(x/D) + P(N) * P(x/N) \\ &= 0.005 * 0.95 + 0.995 * 0.01 \\ P(x) &= 0.0147 \\ p(D/x) &= \frac{P(x/D)P(D)}{P(x)} - -ByBayesTheorem \\ &= \frac{0.95 * 0.005}{0.0147} - -Byusing1 \\ &= \frac{0.475}{1.47} = 0.3231292517 \end{aligned}$$

Hence probability a person has the disease given that his test result is positive is **0.3231292517**.

[c-d] On the morning of September 31, 1982, the won-lost records of the three leading baseball teams in the western division of the National League of the United States were as follows:

Team	Won	Lost
Atlanta Braves	87	72
San Francisco Giants	86	73
Los Angeles Dodgers	86	73

Each team had 3 games remaining to be played. All 3 of the Giants games were with the Dodgers, and the 3 remaining games of the Braves were against the San Diego Padres. Suppose that the outcomes of all remaining games are independent and each game is equally likely to be won by either participant. If two teams tie for first place, they have a playoff game, which each team has an equal chance of winning.

(c) What is the probability that Atlanta Braves wins the division? [2 pts]

Solution: There are four possible cases when Atlanta Braves can win the division.

- **case 1:** (Event C1) Braves wins all three games with Padres.

$$P(C1) = 1/2 * 1/2 * 1/2 = 1/8$$

- **case 2:** (Event C2) Braves wins two out of three games . If either Giants or dodgers win three games(1/8 probability each), it leads to additional playoff. Then Braves win with probability 1/2 in playoff. P(C2) = Prob of braves winning playoff(1/2) * prob of giant or dodgers winning 3 games(1/8+ 1/8)* prob of braves winning 2 games in initial set(3/8 (3 possibilities 1/8 each))

$$P(C2) = 1/2 * (2/8) * 3/8 = 3/64$$

- **case 3:** (Event C3) Braves wins two out of three games(3/8) and giant or dodgers wins only 2 games(3/8+3/8). There is no need for playoff since braves are already leading.

$$P(C3) = 3/8 * 3/4 = 9/32$$

- **case 4:** (Event C4) Braves wins one out of three games(3/8) and giant or dodgers wins only 2 games(3/8+3/8), it leads to additional playoff. Then Braves win with probability 1/2 in playoff.

$$P(C4) = 1/2 * (6/8) * 3/8 = 9/64$$

Probability of Braves winning the tournament = (1/8+3/64 + 9/32 + 9/64)
= **0.59375**

(d) What is the probability to have an additional playoff game? [3 pts]

Solution: Assuming that Padres are not good enough to enter playoffs, the only teams that can compete in playoff are Atlanta Braves vs either giants or dodgers.
(There is no possibility of payoffs b/n giants and Dogers since they are at equal points and they have odd number of games b/n them.)
Following are the cases when Braves and giants/dodgers can enter playoffs

- **case 1:** (Event C1) ATL Braves wins 2 games(3/8) and either Dodgers or Giants wins 3 games(1/8+1/8) (P(C1) = 3/32)

- **case 2:** (Event C2) ATL Braves wins 1 games(3/8) and neigher dodgers or giants win three games (3/4) P(C2 = 9/32)

Overall probability of playoff = 3/32+9/32 = 12/32 = 0.375
Hence probability to have an additional playoff game is **0.375**

2 Maximum Likelihood [15 pts]

Suppose we have n i.i.d (independent and identically distributed) data samples from the following probability distribution. This problem asks you to build a log-likelihood function, and find the maximum likelihood estimator of the parameter(s).

(a) Poisson distribution [5 pts]

The Poisson distribution is defined as

$$P(x_i = k) = \frac{\lambda^k e^{-\lambda}}{k!}(k = 0, 1, 2, ...).$$

What is the maximum likelihood estimator of λ? **Solution:**

$$P(x_i = k) = \frac{\lambda^k e^{-\lambda}}{k!}(k = 0, 1, 2, ...).$$

We have n total number of values for k.
Log likelihood estimator function

$$L(sample/\lambda) = Log(\prod_{k=0}^{n-1} P(x_i = k))(k = 0, 1, 2, ...).$$

$$= Log(\prod_{i=0}^{n-1} \frac{\lambda_i^x e^{-\lambda}}{x_i!})$$

$$= \sum_{i=0}^{n-1} Log(\frac{\lambda_i^x e^{-\lambda}}{x_i!})$$

$$= \sum_{k=0}^{n-1} (x_i(Log(\lambda)) - \lambda + log(x_i!))$$

The arg max at which the log-likelihood is maximized, derivative is 0.

$$\frac{\partial L}{\partial \lambda} = \sum_{k=0}^{n-1} \frac{x_i}{\lambda} - 1 = 0$$

$$=> \sum_{i=0}^{n-1} \frac{x_i}{\lambda} - 1 = 0$$

$$=> \lambda = \sum_{i=0}^{n-1} x_i/n$$

Hence maximum likelihood estimation of λ is

$$\boxed{\lambda = \sum_{i=0}^{n-1} x_i/n}$$

(b) Multinomial distribution [5 pts]

The probability density function of Multinomial distribution is given by

$$f(x_1, x_2, \dots, x_k; n, \theta_1, \theta_2, \dots, \theta_k) = \frac{n!}{x_1!x_2!\cdots x_k!}\prod_{j=1}^{k}\theta_j^{x_j},$$

where $\sum_{j=1}^{k}\theta_j = 1, \sum_{j=1}^{k} x_j = n$. What is the maximum likelihood estimator of $\theta_j, j = 1, \dots k$?

Solution:

$$f(x_1, x_2, \dots, x_k; n, \theta_1, \theta_2, \dots, \theta_k) = \frac{n!}{x_1!x_2!\cdots x_k!}\prod_{j=1}^{k}\theta_j^{x_j},$$

$$=> L = log(f) = log(\frac{n!}{x_1!x_2!\cdots x_k!}\prod_{j=1}^{k}\theta_j^{x_j})$$

$$= log(\frac{n!}{x_1!x_2!\cdots x_k!}) + \sum_{j=1}^{k} x_j log\theta_j \tag{1}$$

We know that

$$\sum_{j=1}^{k} x_j = n$$

$$=> x_k = n - \sum_{j=1}^{k-1} x_j \&\& \tag{2}$$

$$\sum_{j=1}^{k}\theta_j = 1$$

$$=> theta_k = 1 - \sum_{j=1}^{k-1}\theta_j \tag{3}$$

from , and , we can say that

$$L = log(\frac{n!}{x_1!x_2!\cdots x_k!}) + \sum_{j=1}^{k-1} x_j log\theta_j + x_k log(1 - \sum_{j=1}^{k-1}\theta_j)$$

At arg max L

$$\frac{\partial L}{\partial \theta_j} = 0 for all j from 1 to k - 1$$

(Note that last theta depends on other values and hence not an independent variable)

$$=> \frac{x_j}{\theta_j} - \frac{x_k}{1 - \sum_{j=1}^{k-1}\theta_j} = 0$$

$$=> \frac{x_j}{\theta_j} - \frac{x_k}{\theta_k} = 0$$

for all j = 1,2,3,4...k-1
Let

$$=> \frac{x_j}{\theta_j} = c$$

$$=> x_j = c * \theta_j$$

summing and applying and

$$=> \sum x_j = c * sum \theta_j$$

$$=> c = n$$

Therefore

$$\theta_j = \frac{x_j}{n}$$

(c) Gaussian normal distribution [5 pts]

Suppose we have n i.i.d (Independent and Identically Distributed) data samples from a univariate Gaussian normal distribution $\mathcal{N}(\mu, \sigma^2)$, which is given by

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

What is the maximum likelihood estimator of μ and σ^2?

Solution: Let

$$x_1, x_2, ...x_n$$

be the n samples which are observed. log Likelihood function:

$$L = log \prod_{j=1}^{k} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_j-\mu)^2}{2\sigma^2}\right).$$

$$= n * log \frac{1}{\sigma\sqrt{2\pi}} - \sum_{j=1}^{n} \frac{(x_j-\mu)^2}{2\sigma^2}.$$

$$= n/2 * log \frac{1}{\sigma^2 2\pi} - \sum_{j=1}^{n} \frac{(x_j-\mu)^2}{2\sigma^2}.$$

To maximising L using

$$\mu$$

we get

$$\frac{\partial L}{\partial \mu} = 0$$

$$=> \sum_{j=1}^{n} \frac{-2(x_j-\mu)}{2\sigma^2} = 0$$

$$=> \sum_{j=1}^{n} \frac{-2(x_j-\mu)}{2\sigma^2} = 0$$

$$=> \sum_{j=1}^{n} \frac{-2(x_j-\mu)}{=} 0$$

$$=> \mu = \frac{\sum_{j=1}^{n} x_j}{n} - - > \textbf{Answer}$$

To maximise L using

$$\sigma^2$$

we get

$$\frac{\partial L}{\partial \sigma^2} = 0$$

$$=> \sum_{j=1}^{k} - \frac{-(x_j - \mu)^2}{2\sigma^4} - n/2 * \frac{\sigma^2 2\pi}{\sigma^{2^2} 2\pi} = 0$$

$$=> \sum_{j=1}^{k} \frac{(x_j - \mu)^2}{2\sigma^2} - n/2 = 0$$

$$=> \sigma^2 = \sum_{j=1}^{k} \frac{(x_j - \mu)^2}{n} - -\textbf{Answer}$$

3 Principal Component Analysis [20 pts]

In class, we learned that Principal Component Analysis (PCA) preserves variance as much as possible. We are going to explore another way of deriving it: minimizing reconstruction error.

Consider data points $\mathrm{x}^n(n = 1, ..., N)$ in D-dimensional space. We are going to represent them in $\{\mathrm{u}_1, ..., \mathrm{u}_D\}$ orthonormal basis. That is,

$$\mathrm{x}^n = \sum_{i=1}^{D} \alpha_i^n \mathrm{u}_i = \sum_{i=1}^{D} (\mathrm{x}^{nT}\mathrm{u}_i)\mathrm{u}_i.$$

Here, α_i^n is the length when x^n is projected onto u_i.

Suppose we want to reduce the dimension from D to $M < D$. Then the data point x^n is approximated by

$$\tilde{\mathrm{x}}^n = \sum_{i=1}^{M} z_i^n \mathrm{u}_i + \sum_{i=M+1}^{D} b_i \mathrm{u}_i.$$

In this representation, the first M directions of u_i are allowed to have different coefficient z_i^n for each data point, while the rest has a constant coefficient b_i. As long as it is the same value for all data points, it does not need to be 0.

Our goal is setting u_i, z_i^n, and b_i for $n = 1, ..., N$ and $i = 1, ..., D$ so as to minimize reconstruction error. That is, we want to minimize the difference between x^n and $\tilde{\mathrm{x}}^n$ over $\{\mathrm{u}_i, z_i^n, b_i\}$:

$$J = \frac{1}{N}\sum_{n=1}^{N} \|\mathrm{x}^n - \tilde{\mathrm{x}}^n\|^2.$$

(a) What is the assignment of z_j^n for $j = 1, ..., M$ minimizing J? [5 pts]

Solution:

$$J = \frac{1}{N}\sum_{n=1}^{N} \|\mathrm{x}^n - \tilde{\mathrm{x}}^n\|^2.$$

Since all u_i's are Ortho-normal to each other,$u_i u_j = 0$ for all i!= j Hence by expanding J we get

$$J = \frac{1}{N}\sum_{n=1}^{N} (\mathrm{x}^n - \tilde{\mathrm{x}}^n)^T(\mathrm{x}^n - \tilde{\mathrm{x}}^n).$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{M}(\sum (\alpha_i^n - z_i^n)u_i + \sum_{i=M+1}^{D}(\alpha_i^n - b_i)u_i)^T(\sum_{i=1}^{M}(\alpha_i^n - z_i^n)u_i + \sum_{i=M+1}^{D}(\alpha_i^n - b_i)u_i)$$

$$J = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{M}(\alpha_i^n - z_i^n)^2 u_i^T u_i + \sum_{i=M+1}^{D}(\alpha_i^n - b_i)^2 u_i^T u_i \tag{4}$$

(since cross multiplication terms where u_i and u_j are multiplied are 0 due to ortho-normal property)
To minimize J w.r.t z_j^n , we differential and make it 0 for all the n and j values.

$$\frac{\partial J}{\partial z_j^n} = \frac{1}{N}2(\alpha_i^n - z_i^n)(-1) = 0$$

$$=> \boxed{z_i^n = \alpha_i^n} -- > \textbf{Answer}$$

$$=> \boxed{z_i^n = x^{n^T} u_i} -- > \textbf{Answer}$$

(b) What is the assignment of b_j for $j = M+1, ..., D$ minimizing J? [5 pts]

Answer:
By using equation 4, we know that

$$J = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{M}(\alpha_i^n - z_i^n)^2 u_i^T u_i + \sum_{i=M+1}^{D}(\alpha_i^n - b_i)^2 u_i^T u_i$$

To find b_j that minimizes J,

$$\frac{\partial J}{\partial b_j} = \frac{1}{N}\sum_{n=1}^{N} 2(\alpha_j^n - b_j) = 0$$

$$=> \frac{1}{N}\sum_{n=1}^{N} 2(\alpha_j^n - b_j) = 0$$

$$=> \sum_{n=1}^{N}(\alpha_j^n) = \sum_{n=1}^{N}(b_j)$$

$$=> \sum_{n=1}^{N}(\alpha_j^n) = N * b_j$$

$$=> \boxed{b_j = \frac{1}{N}\sum_{n=1}^{N}(\alpha_j^n)}$$

$$=> \boxed{b_j = \frac{1}{N}\sum_{n=1}^{N}(x^{n^T} u_j)} - - > \textbf{Answer}$$

(c) Express optimal $\tilde{\mathrm{x}}^n$ and $\mathrm{x}^n - \tilde{\mathrm{x}}^n$ using your answer for (a) and (b). [2 pts]

By using equation 4 and results from above, we know

$$\tilde{\mathrm{x}}^n = \sum_{i=1}^{M} z_i^n u_i + \sum_{i=M+1}^{D} b_i u_i$$

$$\boxed{\tilde{\mathrm{x}}^n = \sum_{i=1}^{M}(x^{n^T} u_i)u_i + \sum_{i=M+1}^{D} \mu_i u_i}$$

where μ_i = mean of all x_i^n's.

$$\mathrm{x}^n - \tilde{\mathrm{x}}^n = \sum_{i=1}^{M}(x^{n^T} u_i)u_i - (x^{n^T} u_i)u_i + \sum_{i=M+1}^{D}(x^{n^T} u_i)u_i - \mu_i u_i$$

$$\boxed{\mathrm{x}^n - \tilde{\mathrm{x}}^n = \sum_{i=M+1}^{D}(x^{n^T} u_i)u_i - \bar{\mathrm{x}}_i u_i}$$

(d) What should be the $\mathbf{u}_i$ for $i = 1, ..., D$ to minimize J? [8 pts]

Hint: Use $S = \frac{1}{N}\sum_{n=1}^{N}(\mathrm{x}^n - \bar{\mathrm{x}})(\mathrm{x}^n - \bar{\mathrm{x}})^T$ for sample co-variance matrix.

Ans:

We know that S is co-variance matrix.

$$u_j^T S u_j = \frac{1}{N}\sum_{n=1}^{N} u_j^T (x^n - \tilde{\mathrm{x}})(x^n - \tilde{\mathrm{x}})^T u_j$$

The above equation is square of ”projection of $x^n - \tilde{\mathrm{x}}$ on top of u_j”.

$$\sum_{j=1}^{D} u_j^T S u_j = \frac{1}{N}\sum_{j=1}^{D}\sum_{n=1}^{N} u_j^T (x^n - \tilde{\mathrm{x}})(x^n - \tilde{\mathrm{x}})^T u_j$$

$$= \frac{1}{N}\sum_{n=1}^{N} \|\mathrm{x}^n - \tilde{\mathrm{x}}^n\|^2.$$

We know that $u_j^T u_j = 1$ for all j's due to the ortho-normal property. Hence we can add Lagrange term just like in PCA derivation as follows

$$J = \frac{1}{N}\sum_{n=1}^{N} \|\mathrm{x}^n - \tilde{\mathrm{x}}^n\|^2 + \lambda(D - \sum_{j=1}^{D} u_i^T u_i)$$

$$J = \sum_{j=1}^{D} u_j^T S u_j + \lambda(D - \sum_{j=1}^{D} u_i^T u_i)$$

$$J = \sum_{j=1}^{D} u_j^T S u_j + \lambda(1 - u_i^T u_i)$$

To maximize J w.r.t u_i, we take derivative.

$$\frac{\partial J}{\partial u_i} = u_i^T S u_i \lambda(-2u_i^T)$$

Source: wiki page for matrix calculus

I have checked differentiating $x^T x$ gives $2x^T$. Now this equation is similar to one we have solved in PCA. applying same reduction we get u's being **Eigen-vectors** of S. TO minimize J, we pick D eigen vectors in decreasing order of their Eigen-values. first M values have M highest eigen values.

4 Clustering [20 pts]

[a-b] Given N data points $\mathrm{x}^n (n = 1, \ldots, N)$, K-means clustering algorithm groups them into K clusters by minimizing the distortion function over $\{r^{nk}, \mu^k\}$

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r^{nk} \|\mathrm{x}^n - \mu^k\|^2,$$

where $r^{nk} = 1$ if x^n belongs to the k-th cluster and $r^{nk} = 0$ otherwise.

(a) Prove that using the squared Euclidean distance $\|\mathbf{x}^n - \mu^k\|^2$ as the dissimilarity function and minimizing the distortion function, we will have

$$\mu^k = \frac{\sum_n r^{nk} \mathbf{x}_n}{\sum_n r^{nk}}.$$

That is, μ^k is the center of k-th cluster. [5 pts]

solution: Given that

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r^{nk} \|\mathrm{x}^n - \mu^k\|^2,$$

Clearly J is at an optima when

$$\frac{\partial J}{\partial \mu^i} = 0$$

$$=> \frac{\partial \sum_{n=1}^{N} \sum_{k=1}^{K} r^{nk} \|\mathrm{x}^n - \mu^k\|^2}{\partial \mu^i} = 0$$

$$=> \sum_{n=1}^{N} 2r^{ni} * (x^n - \mu^i) = 0$$

$$=> \mu^i = \frac{\sum_{n=1}^{N} r^{ni} x^n}{\sum_{n=1}^{N} r^{ni}}$$

To prove that this point which is an optima is infact a argmin is by taking very large values of μ. J values for large μ is very high which indicates that this point is a minimum where cost 'J' is minimized. **Hence** minimizing the distortion function 'J', we get

$$\mu^k = \frac{\sum_n r^{nk} \mathrm{x}_n}{\sum_n r^{nk}}.$$

(b) Prove that K-means algorithm converges to a local optimum in finite steps. [5 pts]

- If in any step of iteration, if the J value doesn't decrease, then we have already converged to a local minima.
- There are k^n possible cluster assignments (each point can be assigned to one of k clusters). In each cluster assignment there can be unique minima as shown in part 1 where center is taken as the argmin
- also in each iteration, we change cluster assignment only if the cost decreases. Recaculating the cluster center will further reduce the 'J' value.
- Hence using above argument since there are finite possible cost values and in each iterations cost decreases, it will eventually reach Minima where we define algorithm to converge.

[c-d] In class, we discussed bottom-up hierarchical clustering. For each iteration, we need to find two clusters $\{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_m\}$ and $\{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_p\}$ with the minimum distance to merge. Some of the most commonly used distance metrics between two clusters are:

- Single linkage: the minimum distance between any pairs of points from the two clusters, i.e.

$$\min_{\substack{i=1,\ldots,m \\ j=1,\ldots,p}} \|\mathrm{x}_i - \mathrm{y}_j\|$$

- Complete linkage: the maximum distance between any parts of points from the two clusters, i.e.

$$\max_{\substack{i=1,\ldots,m \\ j=1,\ldots,p}} \|\mathrm{x}_i - \mathrm{y}_j\|$$

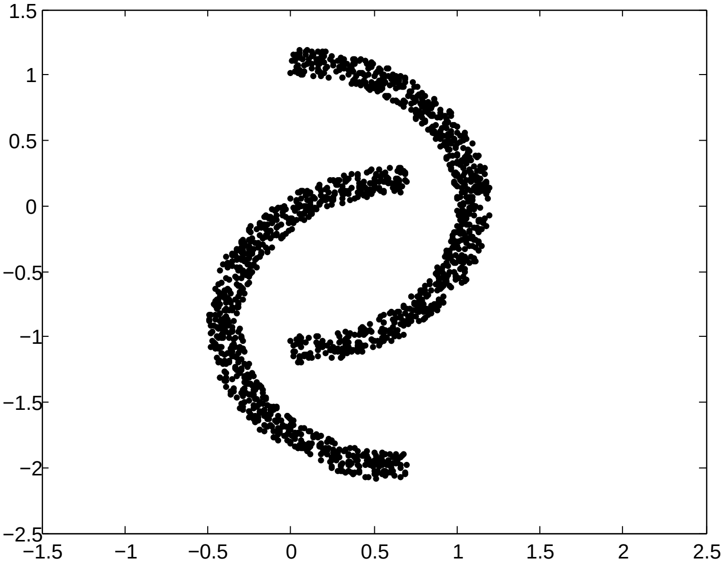
- Average linkage: the average distance between all pair of points from the two clusters, i.e.

$$\frac{1}{mp} \sum_{i=1}^{m} \sum_{j=1}^{p} \|\mathrm{x}_i - \mathrm{y}_j\|$$

(c) When we use the bottom up hierarchical clustering to realize the partition of data, which of the three cluster distance metrics described above would most likely result in clusters most similar to those given by K-means? (Suppose K is a power of 2 in this case). [5 pts]

Solution: The average linkage is very the best strategy here. Note that k-means and average linkage both uses the means as a base of deciding connectivity. Hence Average linkage is the best metric here.

(d) For the following data (two moons), which of these three distance metrics (if any) would successfully separate the two moons? [5 pts]



Solution:

- Single Linkage is the best algorithm to use here. Spectral clustering is the best algorithm to use in this scenario and spectral clustering is essentially uses same ideology.
- To try to understand this, if we use average of complete linkage, consider scenario where each moons are broken into two halfs. the top half of the left moon is very close to bottom half of right one and they both will be clustered into 1 category. Hence we need to use single linkage.