44/45

Exam II, Fall 2009
Foundation of Data Mining Course
CSEE Department, University of Maryland Baltimore County

Note: Closed book exam.
Time: 1 hour and 15 minutes

Sandipan Dey

1. **Orthogonal Transformations: [4+3+3+10+10=30]**

a. What kind of basis functions are used in Fourier transformation? How are they different from the ones used in case of PCA?

Fourier transform can be represented by

$$F_w = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} f_t e^{-\frac{2\pi i}{n} wt} \quad \text{(unitary version)}$$

The basis function is: $e^{-\frac{2\pi i}{n} wt}$ and the ~~function~~ Fourier coefficients are expressed as a linear combination of the basis function ~~(orthogonal)~~ (orthogonal)

In case of PCA, our problem is to find

~~$\Sigma \lambda_i A$ max~~ arg max $(\mathrm{var}(w^T x))$,

subject to $ww^T = 1$, the solution of the optimization problem becomes WCW^T, where W represents set of orthonormal eigenvectors of the covariance matrix C.

Here the set of eigenvectors W ~~form~~ represent the orthonormal basis functions, since any data point can be represented as linear combination of the set of eigenvectors $\sum_{i=1}^{n} \lambda_i W_i$, with W_i being linearly independent set of vectors (Basis of R^n)

b. Name one major difference between the Fourier and Wavelet transformations.

fourier transfrom ~~does not take care of the~~ only expresses (by means of Fourier coefficients) ~~only represents~~ a frequency domain representation of the data, but it does not ~~take care~~ represent the time domain variations. It is not good for capturing spikes or ~~Wavelets transform~~ sharp changes in data. It does not take care of localizations.

Wavelet transform represents both time and frequency variations simultaneously. By means of Mother wavelets ~~as~~ ~~capture~~ as a generating function and ~~a~~ small building blocks the entire ~~wave~~ signal can be generated. Also, it ~~is better~~ performs better when capturing sharp changes in data

Wavelet's capture forest and trees in the data.
Also, takes care of the localizations of data, by having
Mother wavelets as building blocks and using scaling and
translation to shrink/expand and shift the blocks respective

c. If G denotes the estimator of a population parameter H then write down the expression for the bias of the estimator.

Bias of an estimator is given by the following :

$B_\theta(G) = E_\theta[G-\theta] = E_\theta[G] - E_\theta[\theta]$ what is θ?

$= E_\theta[G] - \theta$ (θ being a constant)

-1

$E_\theta(G) = \theta \Rightarrow B_\theta(G) = 0$

$\Rightarrow$ G is an unbiased estimator of θ.

d. Let $\bar{x}, j \in \{0,1\}^l$, $f_1 : \{0,1\}^l \to R$, and $f_2 : \{0,1\}^l \to R$ where R denotes the real numbers. Consider the following transformations:

$$f_1(\bar{x}) = \sum_j w_{1,j} \Psi_j(\bar{x}) ; \qquad f_2(\bar{x}) = \sum_j w_{2,j} \Psi_j(\bar{x})$$

where $w_{1,j}$-s and $w_{2,j}$-s are real valued coefficients. $\Psi_j(\bar{x})$-s define a set of orthonormal functions. In other words

$$\sum_x \Psi_j(\bar{x})\Psi_i(\bar{x}) = 0 \text{ when } j \neq i \text{ and } \sum_x \Psi_j(\bar{x})\Psi_i(\bar{x}) = 1 \text{ when } i = j.$$

Prove that

$$\sum_x f_1(\bar{x}) f_2(\bar{x}) = \sum_j w_{1,j} w_{2,j}$$

$$L.H.S. = \sum_{\bar{x}} f_1(\bar{x}) f_2(\bar{x})$$

$$= \sum_{\bar{x}} \sum_{j} w_{1,j} \varphi_j(\bar{x}) \sum_{j} w_{2,j} \varphi_j(\bar{x})$$

$$= \sum_{\bar{x}} \sum_{i,j} w_{1,j} \varphi_j(\bar{x}) w_{2,i} \varphi_i(\bar{x})$$

$$= \sum_{\bar{x}} \sum_{i,j} w_{1,j} w_{2,i} \varphi_j(\bar{x}) \varphi_i(\bar{x})$$

$$= \sum_{i,j} w_{1,j} w_{2,i} \sum_{\bar{x}} \varphi_j(\bar{x}) \varphi_i(\bar{x})$$

(since $w_{1,j}$ and ~~…~~ $w_{2,i}$ are ~~…~~ scalars, not dependent on $\bar{x}$)

$$= \sum_{i,j} w_{1,j} w_{2,i} \left(\sum_{\substack{\bar{x} \\ i \neq j}} \varphi_j(\bar{x}) \varphi_i(\bar{x}) + \sum_{\substack{\bar{x} \\ i = j}} \varphi_j(\bar{x}) \varphi_i(\bar{x}) \right)$$

$$= \sum_{j} w_{1,j} w_{2,j} (0 + 1)$$

(By orthonormal condition)

(Since non-zero only for $j = i$)

$$= \sum_{j} w_{1,j} w_{2,j}$$

$= R.H.S.$ (Proved)

10

e. Consider the following basis function: $\Psi_{\bar{j}}(\bar{x}) = (-1)^{\bar{j}\cdot\bar{x}}$ for question 1 (d). Write down the value of all the four coefficients ($w_{00}, w_{01}, w_{10}, w_{11}$) for the following two bit function:

f(00)=0; f(01)=1; f(10)=1; f(11)=2

$\Psi_{\bar{j}}(\bar{x}) = (-1)^{\bar{j}\cdot\bar{x}}$ and we ~~found~~ know that

$w_{\bar{j}} = \frac{1}{2^n} \sum_{\bar{x}} f(\bar{x}) \Psi_{\bar{j}}(\bar{x})$, $f(\bar{x}) = \begin{bmatrix} f(0,0) \\ f(0,1) \\ f(1,0) \\ f(1,1) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 2 \end{bmatrix}$

here $n = 2$.

$\therefore w_{00} = \frac{1}{2^2} \sum_{\bar{x}} f(\bar{x}) (-1)^{00.\bar{x}}$, let $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$= \frac{1}{4} \sum_{\bar{x}} f(\bar{x}).(-1)^0 \quad (\because [0.0].\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0)$

$= \frac{1}{4} \sum_{\bar{x}} f(\bar{x}) = \frac{1}{4}(0+1+1+2) = 1$

$w_{01} = \frac{1}{4} \sum_{\bar{x}} f(\bar{x}).(-1)^{01.\bar{x}} \quad \cancel{(\because [0\ 1].\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_2)}$

$= \frac{1}{4}\left(f(00).(-1)^{[01][00]^T} + f(01).(-1)^{[01][01]^T} + f(10)(-1)^{[01][10]^T} + f(11)(-1)^{[01][11]^T} \right)$

$= \frac{1}{4}\left(0 + 1.(-1)^1 + 1.(-1)^0 + 2.(-1)^1\right)$

$= \frac{1}{4}(-\cancel{1} + \cancel{1} - 2) = -\frac{1}{2}$

$w_{10} = \frac{1}{4}\left(f(00).(-1)^{[10][00]^T} + f(01)(-1)^{[10][01]^T} + f(10)(-1)^{[10][10]^T} + f(11).(-1)^{[10][11]^T} \right)$

$= \frac{1}{4}\left(0 + 1.(-1)^0 + 1.(-1)^1 + 2.(-1)^1\right)$

$= \frac{1}{4}(\cancel{1} - \cancel{1} - 2) = -\frac{1}{2}$

$w_{11} = \frac{1}{4}\left(0 + 1.(-1)^1 + 1.(-1)^1 + 2.(-1)^2\right)$

$= \frac{1}{4}(0 - 1 - 1 + 2) = 0$

10

2.a

1. Choose $\bar{C}_i$, $\forall i = 1, 2, .., K$
2. $j \leftarrow \arg\min_{i \in \{1,2,..,K\}} d(x, \bar{C}_i)$, $\forall x \in D$, (d is distance metric, e.g. Euclidian)
   assign x to C_j
3. $\bar{C}_i' \leftarrow \frac{1}{|C_i|} \sum_{x_i \in C_i} x_i$ (compute changed cluster centers)
4. Go to Step 2, if $(\exists i) \mid C_i' \neq C_i$

$$f(\bar{x}) = \sum_j w_j \varphi_j(\bar{x}).$$

$f : X^n \to R$

$w_j \in R$.

$$w_j = \frac{1}{2^n} \sum_{\bar{x}} f(\bar{x}) \varphi_j(\bar{x})$$

$$\varphi_j(\bar{x}) = (-1)^{\bar{j} \cdot \bar{x}}$$

2. **Clustering: [4+3+3+3+2=15]**

a. How does k-means clustering work? Write down the main steps.

Steps   K-Means Clustering (Input parameter 'K': #clusters)

1. Randomly select K centroids for the K clusters from the data points, select $\bar{C}_1, \bar{C}_2, \ldots, \bar{C}_K$ (cluster centers)

2. For each data point find the distance (e.g. using Euclidian distance metric) from each of the clusters and find the minimum distance, place the point in the cluster with minimum distance, $\forall x \in D$,

$j \leftarrow \arg\min_{i \in \{1..K\}} d(x, \bar{C}_i)$, j will be the cluster index to which x is to be assigned (Maximize intra-cluster similarity)

3. Recompute the mean for each of the clusters and assign new centroids for each cluster, $(\bar{C}_1', \bar{C}_2', \ldots, \bar{C}_K')$

$\bar{C}_i' \leftarrow \frac{1}{m_i} \sum_{x_i \in C_i} x_i$

4. Go to step 1 until no change, i.e., convergence Stop if $\bar{C}_i = \bar{C}_i'$, $\forall i \in \{1, 2, \ldots, K\}$ and output clusters

b. Identify two major problems of the k-Means Clustering.

Problems of K-means clustering

① Since the centroids are calculated by the statistic 'mean', since mean is very sensitive to extreme points (outliers), the method is not very robust to outliers. The initial choice of cluster-centers must be done very carefully for this reason.

② Can't detect arbitrary-shaped clusters. clusters tend to be of nearly symmetric shape. Also, dependent on the ordering of the data points. Also, number of clusters are needed to be specified initially.

c. What is the time complexity of the k-Means Clustering?

Complexity of the k-means clustering
$= O(Imnk)$

Where I = number of iterations for convergence

m = " " data points

n = " " dimensions

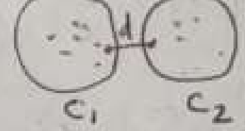
k = " " clusters

$$\left[\begin{array}{l} \text{Step 1: } O(k) \text{ once} \\ \left.\begin{array}{l}\text{Step 2} \\ \text{Step 3}\end{array}\right\} O(mnk) \times I \text{ times} \end{array}\right]$$

$O(mnkI + k)$
$= O(mnkI)$

d. What is Single-link Hierarchical clustering? In bottom-up Hierarchical clustering ~~...~~ technique, while merging the lower level ~~..~~ smaller clusters down the tree to higher level larger clusters in the tree, if the inter-cluster distances are computed like the following:

$$d = \min_{\substack{p \in C_i \\ p' \in C_j}} \|p - p'\|$$



using some norm (e.g. L^2), it's called single link hierarchical clustering

e. How is the divisive hierarchical clustering different from agglomerative hierarchical clustering?

As the name suggests, agglomerative hierarchical clustering starts with a <u>bottom-up</u> approach, i.e., it starts with every data point as a ~~...~~ separate cluster and goes on merging the clusters hierarchically to ~~minimize intra-cluster~~ maximize intra-cluster similarity.

To the contrary, divisive cluster follows a <u>top-down</u> approach, starting from the entire data set and then partitioning it in hierarchical manner.