50/54

Exam I, 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. **Basic Foundation Material: [2+4+4+6+4=20]**

   a. Let A be an n×n matrix where all the rows are linearly independent. What is the rank of A?

Row-Rank(A) = number of linearly independent vectors of the matrix A

Since, Rank(A) = Row-Rank(A),
we have, rank(A) = n (a full-rank matrix)

   b. Name two techniques to detect the statistical dependency between a pair of features. Write down the mathematical expressions.

Let the pair of features be X and Y (column vectors $m \times 1$)

① Covariance
$Cov(X, Y) = E[(X-\mu_X)^T(Y-\mu_Y)] = E[X^TY] - E[X]E[Y]$
$= E[\langle X, Y\rangle] - E[X]E[Y]$, where $\langle X, Y\rangle$ is the inner product

This represents the dependence between two features.

$\boxed{Cov(X,Y) = 0 \text{ means } X, Y \text{ are independent}}$

② Correlation coefficient
$Corr(X, Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}$, this represents the linear relationship between the two features, if any.

$\boxed{Corr(X,Y) = 0 \text{ they are uncorrelated}}$

   c. Write down the mathematical definition of Kullback-Leibler Distance. Is this a symmetric function?

KL-distance (not really a distance metric → not symmetric, does not respect triangle inequality, rather a divergence), between two distributions is given by the following:

Discrete version

$KL(p, q)$ (Two distributions having p.m.f. p and q respectively)
$= D(p\|q)$ (differential entropy)

$= \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$, (where X is the ~~set of~~ domain of the random variable)

$D(p\|q) = \int_{x \in X} p(x) \log \frac{p(x)}{q(x)}\, dx$, where p, q are p.d.f.s respectively

Continuous version

Not Symmetric: $D(p\|q) \neq D(q\|p)$ in general, it's an assymetric distance.

d. Let A be an m×n data matrix. How would you compute A' from A such that the columns of A' have mean zero and unit variance. Analytically argue that your answer is correct. What is the name of this normalization technique?

$A' = \frac{A - \mu(A)}{\sigma(A)}$, where $\mu(A)$ is ~~column~~ a row vector that represents the means of all columns and $\sigma(A)$ " , standard deviation (row) vector.

This is Z-Score normalization.

$A = [a_1 \; a_2 \cdots a_n]$. $\{a_1, a_2, \ldots, a_n$ are column vectors$\}$

$A' = \left[\frac{a_1 - \bar{a}_1}{\sigma_{a_1}} \cdots \frac{a_2 - \bar{a}_2}{\sigma_{a_2}} \cdots \frac{a_n - \bar{a}_n}{\sigma_{a_n}}\right]$.

e. If $(x_1, x_2, \ldots x_m)$ is a random sample from a Poisson distribution with finite mean μ, finite variance σ^2, and $\bar{x} = \frac{1}{m}\sum x_i$ then what is the distribution of $\bar{x}$?

By central limit theorem,

$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1)$ irrespective of the underlying distribution of the sample.

So, $\bar{x}$ (sample mean) will be normally distributed w.r.t. population mean and sample variance, (since, $E[\bar{x}] = \mu$, $Var[\bar{x}] = \frac{\sigma}{\sqrt{n}}$)

f. Write down the three main properties of a distance metric.

If d be a distance Metric

1. ~~#~~ Non-negative: $d(x,x) \geq 0, \ \forall x \in R^n$

2. Symmetry: $d(x,y) = d(y,x)$

~~3. Transitivity: $d(x,y) + d$~~

3. Triangle Inequality:

$$d(x,y) + d(y,z) > d(x,z)$$

2. Representation Construction: [4+6+4+6+2+4+4=30]

a. If Q be the eigenvector matrix and L be the eigenvalue matrix of the data matrix X then write down the eigenvalue decomposition of X.

$$\underset{n\times n}{X} = \underset{n\times n}{Q} \underset{n\times n}{L} \underset{n\times n}{Q^{-1}}$$

If the eigenvectors are linearly independent, i.e., Q is a full-rank matrix, then Q is orthogonal, $Q^{-1} = Q^T$ $(\because QQ^T = I)$, then it can be written as a orthogonal decomposition

$$X = QLQ^T$$

b. If $\bar{x}$ and $\bar{y}$ are two n-dimensional vectors generated by random variables X and Y respectively. Both X and Y have mean zero and unit variance. Prove that $\bar{x}$ and $\bar{y}$ are on average orthogonal to each other. Does it require any condition on n?

$$\left.\begin{array}{l} E[X] = E[Y] = 0 \\ V[X] = V[Y] = 1 \end{array}\right\} \text{(Given)}$$

$$\bar{X} = \begin{Bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{Bmatrix}_{n\times 1}, \quad \bar{Y} = \begin{Bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{Bmatrix}_{n\times 1}$$

$$\Rightarrow x^T y = \sum_{i=1}^{n} x_i y_i$$

$$E[\langle \bar{x}.\bar{y}\rangle] = E[\bar{x}^T \bar{y}]$$

$$= E\left[\sum_{i=1}^{n} x_i y_i\right]$$

$$= \sum_{i=1}^{n} E[x_i y_i] \quad \text{(by linearity of expectation)}$$

$$= \sum_{i=1}^{n} E[x_i] E[y_i]$$

(assuming independence)

$$= \sum_{i=1}^{n} 0.0 = 0$$

$$E[X] = \frac{1}{n}\sum_{i=1}^{n} x_i = 0$$

$$E[Y] = \frac{1}{n}\sum_{i=1}^{n} y_i = 0$$

$$V(X) = E[X^2] - E^2[X] = E[X^2] = 1$$

$$V(Y) = E[Y^2] - E^2[Y] = E[Y^2] = 1$$

$$\cancel{E[XY] = \frac{1}{n}\sum_{i=1}^{n} x_i y_i}$$

~~Now, let's assume there is some linear relation between X & Y (correlation coefficient ±1), then $Y = mX$.~~

$$\cancel{E[\langle \bar{x}.\bar{y}\rangle] = \sum_{i=1}^{n} E[x_i . m x_i] = \sum_{i=1}^{n} m E[x_i^2]}$$

$$= m \sum_{i=1}^{n} E[x_i^2] = 0$$

as well.

c. What is singular value decomposition? Write down the decomposition and explain the different components.

SVD of a matrix A is: decomposition in the form

$\underset{m\times n}{U}\underset{m\times n}{\Sigma}\underset{n\times n}{V^T}$, where

$UU^T = VV^T = I$, ~~set of~~ orthogonal matrices.

U is left singular vector, eigenvectors of AA^T ← (captures variance in rows)

V " right " " " " A^TA. (captures variance in features)

Σ is a diagonal matrix containing the ~~diagonal~~ singular values, that are positive square roots of eigenvalues.

$$AA^T = (U\Sigma V^T)(V\Sigma^T U^T)$$

$$= U\Sigma \underbrace{V^T V}_{I} \Sigma^T U^T = U\Sigma^2 U^T$$

Similarly, $A^TA = (U\Sigma V^T)^T(U\Sigma V^T)$

$$= V\Sigma^T \underbrace{U^TU}_{I}\Sigma V^T$$

$$= V\Sigma^2V^T$$

d. Does random transformation using V, an n × n dimensional randomly generated matrix with i.i.d. mean zero entries, preserve Euclidean distance? Clearly prove your answer.

$\underset{\substack{m\times n \\ \text{Data}}}{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$ each x_i be an n dimensional row vector,

$x_i = [x_{i1}\ x_{i2} \cdots x_{in}]$

let V be the transformation matrix for $T: \underset{R^n}{X} \to \underset{R^n}{Y}$

$$\boxed{\underset{n\times n}{V}\underset{n\times n}{X} = \underset{n\times n}{Y}}$$

$$[a_1\ a_2 \cdots a_n]\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$= \underset{A}{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}} \underset{X}{\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{nn} \\ \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix}} = \underset{Y}{\begin{bmatrix} \sum_{k=1}^{n} a_{ik}x_{kj} \end{bmatrix}}_{n\times n}$$

$y_{ij} = \sum_{k=1}^{n} a_{ik}x_{kj}$

3

Now, before transformation,

$$d(x_1, x_2) = \sqrt{(x_{11}-x_{21})^2 + (x_{12}-x_{22})^2 + \cdots + (x_{1n}-x_{nn})^2}$$

After transformation,

$$d(Y_1, Y_2) = \sqrt{(y_{11}-y_{21})^2 + (y_{12}-y_{22})^2 + \cdots + (y_{1n}-y_{nn})^2}$$

$$= \sqrt{\left(\sum_{k=1}^{n} a_{1k}x_{k1} - \sum_{k=1}^{n} a_{2k}x_{k1}\right)^2 + \left(\sum_{k=1}^{n} a_{1k}x_{k2} - \sum_{k=1}^{n} a_{2k}x_{k2}\right)^2 + \cdots}$$

$\neq d(x_1, x_2)$ in general

since, ~~...~~

Since, $\left(\sum_{k=1}^{m} a_{1k}x_{k1} - \sum a_{2k}x_{k1}\right)^2$

$$= \left(\sum_{k=1}^{m} (a_{1k} - a_{2k})x_{k1}\right)^2 \neq (x_{11} - x_{21})^2 \text{ in } \underline{\text{general}}.$$

Hence, in general will not be preserved

e. What is Bayes Theorem? Write down the expression.

Bayes Theorem connects ~~p~~ conditional probability distribution with the ~~p~~ marginal distribution.

A, B be two events, $P(A)$, $P(B)$ be the <u>prior probabilities</u>, $P(B|A)$ be the conditional prob. of B given A, then the <u>posterior</u> prob. $P(A|B) = \frac{P(B|A)P(A)}{P(A)}$.

If B is divided into n partitions, $B = \bigcup_{i=1}^{n} B_i$, $\bigcap_{i=1}^{n} B_i = \phi$.

$$P(A|B_i) = \frac{P(B_i|A)\,P(A)}{\sum_{i=1}^{n} P(B_i|A)P(A)}, \text{ by countable additivity)}$$

f. Define l_p norm.

l_p norm can be defined by (on a feature variable X).

$$\left(\sum |x|^p\right)^{\frac{1}{p}}$$

g. Write down the relation between the eigenvalues and the sample variance in Principal Component Analysis.

$$\sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} \sigma_i^2 = \text{Trace of the covariance matrix } (A^T A)$$

sum of eigenvalues

sum of sample variance