MLA

04/05/15 EXAM PAPER 2014 Q1 DAVID WESTBROOKE

1 A. Use/Aim of PCA.
- Dimension reduction - included dimensions are orthogonal and ordered to account for as much variation as possible.
- Visualisation
- Lower Dimensional Summary.
- Helps identify relationships in data
- Potentially assists other clustering or classification algorithms (performance)

B Lagrange Multipliers

$P = \underline{a}^T \Sigma \underline{a} - \lambda(\underline{a}^T\underline{a} - 1)$

Solve $dP/da_k = 0$ for $k = 1 \dots m$

$dP/d\lambda = 0 \Rightarrow a^T a - 1 = 0 \Rightarrow$ constraint satisfied

$\Rightarrow \sum_{i=1}^{m} a_i \tilde{b}_{ik} - \lambda a_k = 0$

$\Rightarrow a^T \Sigma - \lambda a^T \Rightarrow a$ is an eigenvector of Σ

C i. Standardisation of data
- If not standardised, one component will account for most variance and figure more prominently in the solution, swamping out all other values. PCA will focus on that variable

ii. Number of PC's
- Could select 2 or 3 3-account for more variation (96.4%) but 2 dimension is easier to graph (visualise) and interpret.

iii. PC1: high score = high birth rate, high death rate, high ID and small LEM and LEF.
Low score = opposite
Interpreted as health care - independent of GNP

PC2: High score = High loading on GNP ⇒ wealth
High score for poor country, low score for rich country

C iv. (-0.3, -1.1, +0.5, 0.8, 0.9, -0.6)

$0.43(-0.3) + 0.37(-1.1) + 0.46(-0.5) - 0.47(0.8) - 0.48(0.9) + 0.08(-0.6) = -1.622$

$-0.3(-0.3) + 0.14(-1.1) + 0.06(-0.5) - 0.8(0.8) - 0.1(0.9) - 0.99(-0.6) = 0.282$

(-1.622, 0.282)

D PCA and MDS

- Same aim - Dimension reduction
- Create configuration of points (map) in lower dimension so that dissimilarities match original dissimilarities as much as possible
- Choice of loss function Sammon, Isoch or Stress
- Choose metric or non-metric scaling
- In classical setting - eigen decomposition of a function of a dissimilarity matrix
  PCA - eigen decomposition of covariance matrix
- Classical setting gives same map as PCA does

PCA and FA

- Same aim - dimensionality reduction
- PCA looks for linear combinations of the data matrix X that are uncorrelated and of high variance, while FA seeks unobserved linear combinations of the variables representing underlying fundamental quantities.
- PCA makes no assumption about the form of the cov matrix, FA assumes that data comes from a well-defined model in which specific assumptions hold
  eg. $\mathbb{E}[F] = 0$
- PCA: data $\Rightarrow$ PC's FA: Factors $\Rightarrow$ data.
- When specific variances are large they are absorbed into the PC's whereas FA model a special provision for them. When the specific variances are small, PCA and FA give similar results
- Choices: - orthogonal / oblique methods
  - Choice of rotation and Factor loadings
  - Scale invariant

MLA

2014 Q3 EXAM PAPER

3 A. Motivation for using Logistic Regression
- Parametric classification technique
- Easily handle continuous covariates without discretisation
- Binary response
- Interpretable
- Variable can be categorical not continuous ⇒ Categorical outcome
- Predictor variables in LDR do not need to be linearly related, normally distributed or have equal variance within each group

B. Explain output
- Intercept: β_0 expected mean value of y (learning certificate) when all $x=0$ If every is 0, person will not take learning certificate according to model. Not 0.

Each unit increase in DVRT increases by 0.06

Sex - more likely to do LC if girl.

Increase in Math by 0.034 by prestige score of person

C. $$\frac{\exp(\beta_0 + 0.06\,DVRT + 0.51\,Sex + 0.034\,Math)}{1+\exp(\beta_0 + 0.06\,DVRT + 0.51\,Sex + 0.034\,Math)} = P(LC)$$

female DVRT Score = 120 Math = 32

$$= \exp(-8.648 + 0.06(120) + 0.52(2) + 0.04(32)) \quad \frac{e^{(0.751)}}{1+\exp(0.751)}$$

$$= 0.67$$

$\Rightarrow$ will do LC.

D. Including interaction add) possible products of the covariates to the model
- Allows for the effect of one covariate to be altered depending on value of another covariate

- Interaction would be reasonable if the effect of DVRT had upon LC would differ depending on or not an individual was a girl or boy
- For example the score of DVRT of a person may be considered very important if the person is male but not if they are female

E Difference between Logistic Regression and LDA

- In LDA the decision boundary between class k and class l given by

  $\log \frac{P(k|x)}{P(l|x)} = \log \frac{\pi_k}{\pi_l} + \log \frac{f(x|k)}{f(x|l)} > 0$

- In logistic regression model assumption is: $\log \frac{P(k|x)}{P(l|x)} = \beta_0 + \beta^T x$

- Model have same form

- Difference lie in way the linear coefficients are estimated

  Log R - parameters found through MLE - Binary outcome

  LDA - no restrictions to which group it belongs

MA

EXAM PAPER 2014 Q2 CLUSTERING

2 A. Hierarchical
- Start with each point as a cluster on its own
- Clusters are merged by combining closest 2 at each depth
- Closeness defined by dissimilarity matrix
- Continue until a single cluster is formed containing all data points
- Go back and determine when merging should be stopped: $h + 3s_h$ or relatively large jump in next merge distance as seen through dendrogram

Iterative
- Predefined number of clusters - specify number of clusters you want.
- Assign data points to a specific cluster
- Iterate and re-assign membership until convergence
- Convergence - repeat of previous clustering of data (check that within minimal internal dissimilarity)
- Example includes: K means, partitioning around medoid

B Maximum Dissimilarity matrix
- Compare all values in one row for one country against another row from a different country.
- Compare UK(BR) with USA(BR) etc for DR, ID and LEM.
- Distance = max of these four distances

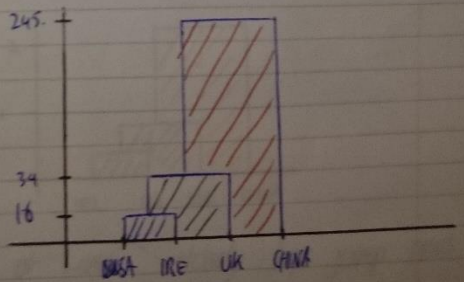
	UK	USA	IRELAND	CHINA
UK	0	3·4	2·4	23·6
USA	3·4	0	1·6	22·9
IRELAND	2·4	1·6	0	24·5
CHINA	23·6	22·9	24·5	0

C Dendrogram with complete linkage
complete linkage $d(A,B) = \max_{x \in A, y \in B} d(x,y)$

Merge USA + Ireland at 1·6
Merge UK and group at 3·4
Merge China and group at 24·5

Choose $\bar{h} + 3s_h$

We choose 2 groups as there is a) a massive jump in the merging height from two groups to one group



D Rand Index.
- Indicates level of agreement between two different clustering method with possibly different numbers of clusters by considering how frequently the method agree on when ~~points of~~ pairs of data points should be in same cluster.

$$\text{Rand Index} = \frac{\binom{n}{2} + 2\sum_{i=1}^{q}\sum_{j=1}^{q}\binom{n_{ij}}{2} - \left[\sum_{i=1}^{q}\binom{n_{i.}}{2} + \sum_{j=1}^{q}\binom{n_{.j}}{2}\right]}{\binom{n}{2}}$$

$$= \frac{\binom{31}{2} + 2\left[\binom{5}{2}+\binom{6}{2}+\binom{3}{2}+\binom{4}{2}+\binom{8}{2}+\binom{5}{2}\right] - \binom{11}{2} - \binom{7}{2} - \binom{13}{2} - \binom{16}{2} - \binom{13}{2}}{\binom{31}{2}} = 0.49$$

E Dissimilarity from data points Containg binary, numerical and categorical data.
- We can work out a dissimilarity for the measure variables (e.g euclidean), Simple matching for binary and proportion of times in agreement in categorical.
- Take a weighted average of these numbers, with weights being proportion of ~~two in agreement~~ variable of that type of data.

F. Real life Clustering Application
- Bank clust transaction and individual profile-fraud people are similarly
- Determing terrorist network - same behaviour - same cluster.
- Online dating - similar interest / pair them
- Search engine - google suggested path

3/11/14 DEATH BY MLA Exam Paper 2014

A Usefulness of PCA
- Dimension reduction technique - so included dimensions are orthogonal and ordered to account for as much variation as possible
- Visualisation
- Lower dimensional summary
- Helps identifying relationships in the data by removing inherent noise
- Potentially assist other clustering or classification algorithms (performance)

B $P = \underline{a}^T \Sigma \underline{a} - \lambda(\underline{a}\underline{a} - 1)$

Solve $\frac{dP}{da_k} = 0$ for $k = 1 \dots m$

$\frac{dP}{d\lambda} = 0 \Rightarrow \underline{a}^T \underline{a} - 1 = 0 \Rightarrow$ constraint satisfied

$\Rightarrow \sum_{i=1}^{m} a_i b_{ik} - \lambda a_k = 0$

$\Rightarrow \underline{a}^T \Sigma = \lambda \underline{a}^T \Rightarrow a$ is an eigen vector of Σ

C i One component will account for most of variance of the model, it will dominate the principle component. Will swamp our other values PCI will focus on that variance

ii Could be 2 or 3 3 - accounts for more variation.
2 - easier to graph / visualise.

iii 1. More people born - more must die, die early, life expectancy goes down

PC1: high score - high birth rate, high death rate, high ID, small LEM and LEF
Low score is the opposite HEALTH CARE
independent of GNP

2. 2014 Paper.

PC2: High loading on GNP ⇒ wealth
High score for poor countries
Low score for rich countries

C iv. (−0.3, −1.1, −0.5, 0.8, 0.9, −0.6)
(−1.622, 0.282)
coord 1. coord 2

all those by PC1, than all those by PC2
= 2 points

0.43(−0.3) + (0.37)(−1.1) + ... + (0.08)(−0.6) = −1.622
−0.03(−0.3) + (0.14)(−1.1) + ... + (0.99)(−0.6) = 0.282

D Factor Analysis, multidimensional scaling

MDS
- Create configuration of points (map) in lower dimension so dissimilarities match original dissimilarities as closely as possible
- Choice of loss function (Stress, Sammon Stress)
whether it is metric or non metric
↳ Dissimilarity measure / Standardisation
In CLASSICAL SETTING - eigen decomposition of a function of the dissimilarity matrix
(PCA ⇒ covariance matrix)
Classical setting gives same map as PCA does

FACTOR ANALYSIS
Seek unobserved linear combinations of the variable representing underlying quantities, i.e. the factors
- Tries to model the covariance matrix with a reduced set of parameters
Underlying specific assumptions holding true
factors ind, Specific factor E[X] of 0,
- Choices: - Orthogonal / oblique methods
- Choice of rotation of factor loadings.
- Scale invariant.

1/12/14 MLA 2014 Exam Paper Q2

Q2 A Hierarchical v iterative clustering

Hierarchical
- each point starts as its own cluster
- clusters are merged by combining closest 2 at each depth
- closer defined by dissimilarity matrix - means for single and linkage for distance
- Continue until a single cluster is formed consisting of all data points
- Go back and determine when merging should of stopped eg. rule average height + 3 s.d of heights or relatively large jump in next merging distance etc. - as seen through a dendrogram

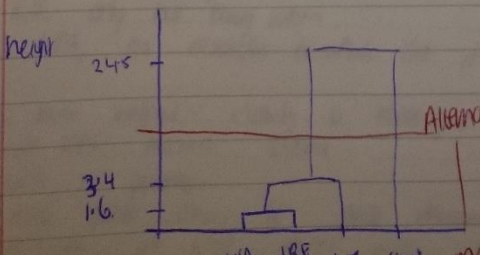
Iterative
- Pre defined number of clusters
- Assign data points to a specific cluster
- iterate points between clusters, until convergence
- Convergence - ~~repeat~~ repeat of a previous clustering of the data [choose that with minimum internal dissimilarities]
- Examples include k means, partitioning around medoids

B Maximum Dissimilarity

	UK	USA	IRELAND	CHINA
UK	0	3.4	2.4	23.6
USA	3.4	0	1.6	22.9
IRELAND	2.4	1.6	0	24.5
CHINA	23.6	22.9	24.5	0

Take max difference between two groups using horizontal values

C.



merge USA and Ireland 1.6
merge UK and group 3.4
merge China and gp 24.5
h+3 s.d
value 7 24.5 rule of thumb!
all in one cluster
massive jump: 2 groups most appropriate

D Rand Index

Indicates level of agreement between 2 different clustering methods with possibly different numbers of clusters by considering how frequently the methods agree ~~either~~ on when pairs of data points should be in the same cluster.

$$= \frac{\binom{31}{2} + 2\left[\binom{5}{2}+\binom{9}{2}+\binom{2}{2}+\binom{8}{2}+\binom{5}{2}\binom{3}{2}\right] - \left[\binom{11}{2}\binom{7}{2}+\binom{16}{2}\right] - \binom{15}{2}}{\binom{31}{2}} = 0.49$$

E - Take each collection of variables of a certain type and calculate dissimilarity for that subset eg euclidean for numeric, simple matching for binary and proportion of times in agreement in categorical
- Take a weighted average of those numbers with weights being proportion of variables of that type or data

F - Steering of students in cluster
- Bank: cluster transactions and individuals profile ⇒ fraud people acting similarly
- Determining in terrorist networks - same behaviour - same clusters
- Online dating - similar interests/hobbies/partners
- Search engines - google suggested posts/products

4/12/14 MLA Exam Paper

3 (b) Assumption for LDA
- Knows how many groups there are in data
LDA and QDA assume data within a class are distributed/spread according to a multivariate normal distribution, which each class having its own mean vector

In LDA a common covariance whilst in QDA, each class can have its own covariance matrix

c Derivation from the notes

d $\left[\begin{pmatrix}0.4\\0.6\end{pmatrix} - \frac{\begin{pmatrix}1.1\\1.9\end{pmatrix}+\begin{pmatrix}0.1\\0.6\end{pmatrix}}{2}\right]^T \begin{pmatrix}0.6 & -0.2\\-0.2 & 0.4\end{pmatrix}\left(\begin{pmatrix}0.1\\0.6\end{pmatrix}-\begin{pmatrix}1.1\\1.9\end{pmatrix}\right)$

Group 1 if expression is positive

$\left[\begin{pmatrix}0.4\\0.6\end{pmatrix} - \frac{\begin{pmatrix}0.1\\0.6\end{pmatrix}}{2}\right]^T \begin{pmatrix}0.6 & -0.2\\-0.2 & 0.4\end{pmatrix}\begin{pmatrix}0.1\\0.6\end{pmatrix}$

$= (-0.4, -0.3)\frac{1}{2}\begin{pmatrix}0.6 & -0.2\\-0.2 & 0.4\end{pmatrix}\begin{pmatrix}0.1\\0.6\end{pmatrix} = (-0.29, 0.3)\begin{pmatrix}0.1\\0.6\end{pmatrix} = 0.3$

Positive ⇒ assign to group one

E Logistic Regression uses a logit link function to link the probability of class assignment to a linear function of data points
$\log \frac{P(x=1)}{P(x=0)} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots$

Parameters found through MLE

In LDA we have $\log \frac{P(x=1)}{P(x=0)} = \log\left(\frac{\pi_1}{\pi_0}\right) + \log \frac{f(x \mid x \in 1)}{f(x \mid x \in 0)}$ ← ratio of prior probs

Difference is how parameters are obtained
LR only has binary option
LDA, no restriction to which class it belongs to

F Is nearer neighbour - classify to nearest group
LDA uses Mahalanobis distance
LDA is parametric
KNN doesn't classify by probability of class membership
As more of NN interest, will be assigned to class which is more prevalent in data
LDA - separated by hyperplane