Data Mining and Exploration: Descriptive Modelling

Amos Storkey, School of Informatics

January 23, 2006

`http://www.inf.ed.ac.uk/teaching/courses/dme/`

`These lecture slides are based extensively on previous versions of the course written by Chris Williams.`

Descriptive Modelling

Descriptive models are a summary of the data

- Describing data by probability distributions
  - Parametric models
  - Mixture Models
  - Non-parametric models
  - Graphical models

Descriptive Modelling

Descriptive models are a summary of the data

- Clustering
  - Partition-based Clustering Algorithms
  - Hierarchical Clustering
  - Probabilistic Clustering using Mixture Models

Reading: HMS, chapter 9

Describing data by probability distributions

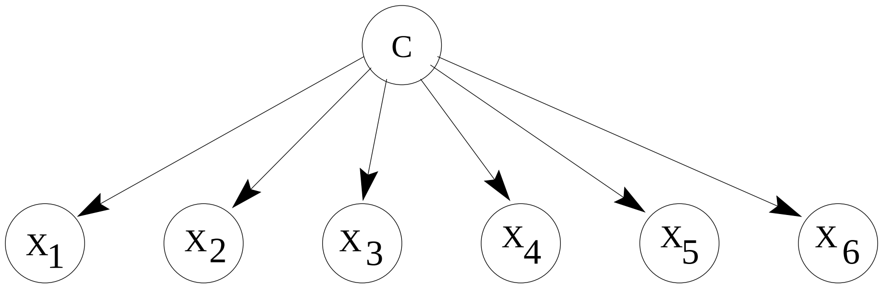
- Parametric models, e.g. single multivariate Gaussian
- Mixture models, e.g. mixture of Gaussians, mixture of Bernoullis
- Non-parametric models, e.g. kernel density estimation

$$\hat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\mathbf{x} - \mathbf{x}_i)$$

Does not provide a good *summary* of the data, expensive to compute on large datasets

Probability Distributions: Graphical Models

- Mixture of Independence Models



(also Naive Bayes model)
- Fitting a given graphical model to data
- Search over graphical structures

Clustering

Clustering is the partitioning of a data set into groups so that points in one group are similar to each other and are as different as possible from points in other groups

- Partition-based Clustering Algorithms
- Hierarchical Clustering
- Probabilistic Clustering using Mixture Models

Examples

- Split credit card owners into groups depending on what kinds of purchases they make
- In biology, can be used to derive plant and animal taxonomies
- Group documents on the web for information discovery

Defining a partition

- Clustering algorithm with k groups
- Mapping c from input example number to group to which it belongs
- In $\mathbb{R}^d$, assign to group j a cluster centre $\mathbf{m}_j$. Choose both c and the $\mathbf{m}_j$'s so as to minimize

$$\sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{m}_{c(i)}|^2$$

- Given c, optimization of the $\mathbf{m}_j$'s is easy; $\mathbf{m}_j$ is just the mean of the data vectors assigned to class j
- Optimiztion over c: cannot compute all possible groupings, use the k-means algorithm to find a local optimum

k-means algorithm

initialize centres $\mathbf{m}_1, \ldots, \mathbf{m}_k$
while (not terminated)
    for $i = 1, \ldots, n$
        calculate $|\mathbf{x}_i - \mathbf{m}_j|^2$ for all centres
        assign datapoint i to the closest centre
    end for
    recompute each $\mathbf{m}_j$ as the mean of the
    datapoints assigned to it
end while

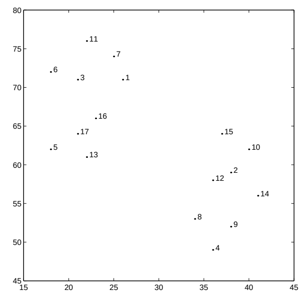
- This is a *batch* algorithm.
- There is also an *on-line* version, where the centres are updated after each datapoint is seen
- Also k-medoids; find a representative object for each cluster centre
- Choice of k?

Hierarchical clustering

for $i = 1, \ldots, n$ let $C_i = \{\mathbf{x}_i\}$
while there is more than one cluster left do
    let C_i and C_j be the clusters minimizing
    the distance $\mathcal{D}(C_i, C_j)$ between any two clusters
    $C_i = C_i \cup C_j$
    remove cluster C_j
end

- Results can be displayed as a *dendrogram*
- This is *agglomerative* clustering; divisive techniques are also possible

Hierarchical Clustering




```
                _____________|------> p08
               |             |______|-------> p04
               |                    |-------> p09
|--------------|                       _______|----> p02
|              |             |--------|       |----> p12
|              |-------------|        |-------> p14
|                            |________|-------> p10
|                                     |-------> p15
|                              ________|-----> p03
|              |--------------|        |-----> p06
|              |              |         _______|---> p01
|--------------|              |--------|       |---> p07
               |                       |--------> p11
               |______________|-------> p05
                              |_______|-------> p13
                                      |______|-----> p16
                                             |-----> p17
```


Distance functions for hierarchical clustering

- Single link (nearest neighbour)

$$D_{sl}(C_i, C_j) = \min_{\mathbf{x},\mathbf{y}} \{d(\mathbf{x}, \mathbf{y}) | \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

  The distance between the two closest points, one from each cluster. Can lead to “chaining”.

- Complete link (furthest neighbour)

$$D_{cl}(C_i, C_j) = \max_{\mathbf{x},\mathbf{y}} \{d(\mathbf{x}, \mathbf{y}) | \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

- Centroid measure: distance between clusters is difference between centroids
- Others possible

Probabilistic Clustering

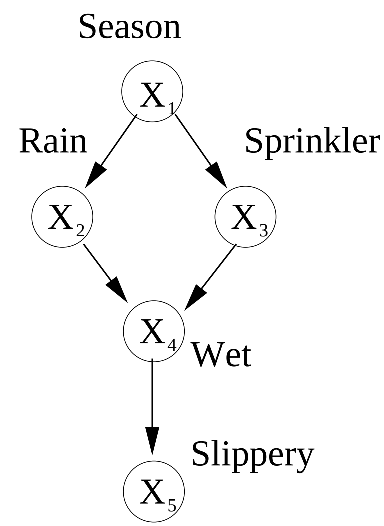
- Using finite mixture models, trained with EM
- Can be extended to deal with outlier by using an extra, broad distribution to “mop up” outliers
- Can be used to cluster non-vectorial data, e.g. mixtures of Markov models for sequences
- Methods for comparing choice of k
- Disadvantage: parametric assumption for each component
- Disadvantage: complexity of EM relative to e.g. k-means

Graphical Models: Causality

- J. Pearl, *Causality*, Cambridge UP (2000)
- To really understand causal structure, we need to predict effect of *interventions*
- Semantics of $do(X = 1)$ in a causal belief network, as opposed to *conditioning* on $X = 1$
- Example: smoking and lung cancer

Causal Bayesian Networks

A causal Bayesian network is a Bayesian network in which each arc is interpreted as a direct causal influence between a parent node and a child node, relative to the other nodes in the network. (Gregory Cooper, 1999, section 4)



Causation = behaviour under interventions

An Algebra of Doing

- Available: algebra of seeing (observation)
  e.g. what is the chance it rained if we *see* that the grass is wet?

$$P(rain|wet) = P(wet|rain)P(rain)/P(wet)$$

- Needed: algebra of doing
  e.g. what is the chance it rained if we *make* the grass wet?

$$P(rain|do(wet)) = P(rain)$$

Truncated factorization formula

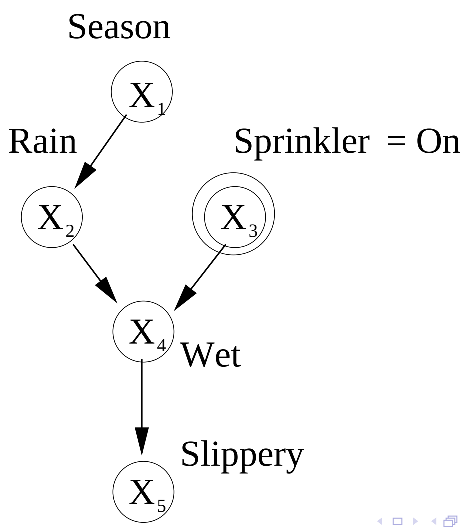
$$P(x_1, \ldots, x_n|\hat{x}_i') = \begin{cases} \prod_{j \neq i} P(x_j|pa_j) & \text{if } x_i = x_i' \\ 0 & \text{if } x_i \neq x_i' \end{cases}$$

$$P(x_1, \ldots, x_n|\hat{x}_i') = \begin{cases} \frac{P(x_1,\ldots,x_n)}{P(x_i'|pa_i)} & \text{if } x_i = x_i' \\ 0 & \text{if } x_i \neq x_i' \end{cases}$$

compare with conditioning

$$P(x_1, \ldots, x_n|x_i') = \begin{cases} \frac{P(x_1,\ldots,x_n)}{P(x_i')} & \text{if } x_i = x_i' \\ 0 & \text{if } x_i \neq x_i' \end{cases}$$

Intervention as surgery on graphs



Controlling confounding bias

We wish to evaluate the effect of X on Y; what other factors Z (known as covariates or confounders) do we need to adjust for?
Simpson's "paradox": an event C increases the probability of E in a population p, but decreases the probability of E in every subpopulation.
E.g. UC-Berkeley investigated for sex-bias (1975). Overall, higher rate of admission of males, but for every department there was a slight bias in favour of admitting females.
[Explanation: females applied to more competitive departments where admission rate was low]

- Another example: administering a drug gives rise to lower rates of recovery than giving a placebo for both males and females, but overall it can appear better
- What treatment would you give to a patient coming into your office? Apparent answer is "if know that patient is male or female, don't give drug, but if gender is unknown, do!". This answer is ridiculous!
- Correct answer to question will depend not only on observed probabilities, but also on assumed causal model. Diagrams below can have the same $P(C, E, F)$, but use of combined or gender-specific tables depends on diagram