Learning the structure of your data using clustering/mixture modelling

Daniel F. Schmidt and Enes Makalic

Centre for Molecular, Environmental, Genetic & Analytic (MEGA) Epidemiology
School of Population Health
University of Melbourne

Work in Progress
2nd August 2012

Content

Overview of Talk

- Clustering
- Mixture Modelling
- Fitting and interpreting mixture models
  - Choosing the number of classes
  - Understanding the structure of the classes
- Some examples
  - Analysis of Pima indians dataset

What is Clustering ?

- We have n items, each with q associated attributes, formed into a matrix

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{pmatrix} = \begin{pmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,q} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n,1} & y_{n,2} & \cdots & y_{n,q} \end{pmatrix}$$

- Group together, or “cluster”, similar items
- A form of unsupervised learning
- Sometimes called intrinsic classification
  $\Rightarrow$ Class labels are learned from the data

K-means Clustering (1)

- Perhaps most commonly used clustering technique
- Models data as having K "centroids" defined by mean vectors

$$\mathbf{M} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_K \end{pmatrix} = \begin{pmatrix} \mu_{1,1} & \cdots & \mu_{1,q} \\ \vdots & \ddots & \vdots \\ \mu_{K,1} & \cdots & \mu_{K,q} \end{pmatrix}$$

- Assigns items to class with most similar mean vector

K-means Clustering (1)

- Perhaps most commonly used clustering technique
- Models data as having K "centroids" defined by mean vectors

$$\mathbf{M} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_K \end{pmatrix} = \begin{pmatrix} \mu_{1,1} & \cdots & \mu_{1,q} \\ \vdots & \ddots & \vdots \\ \mu_{K,1} & \cdots & \mu_{K,q} \end{pmatrix}$$

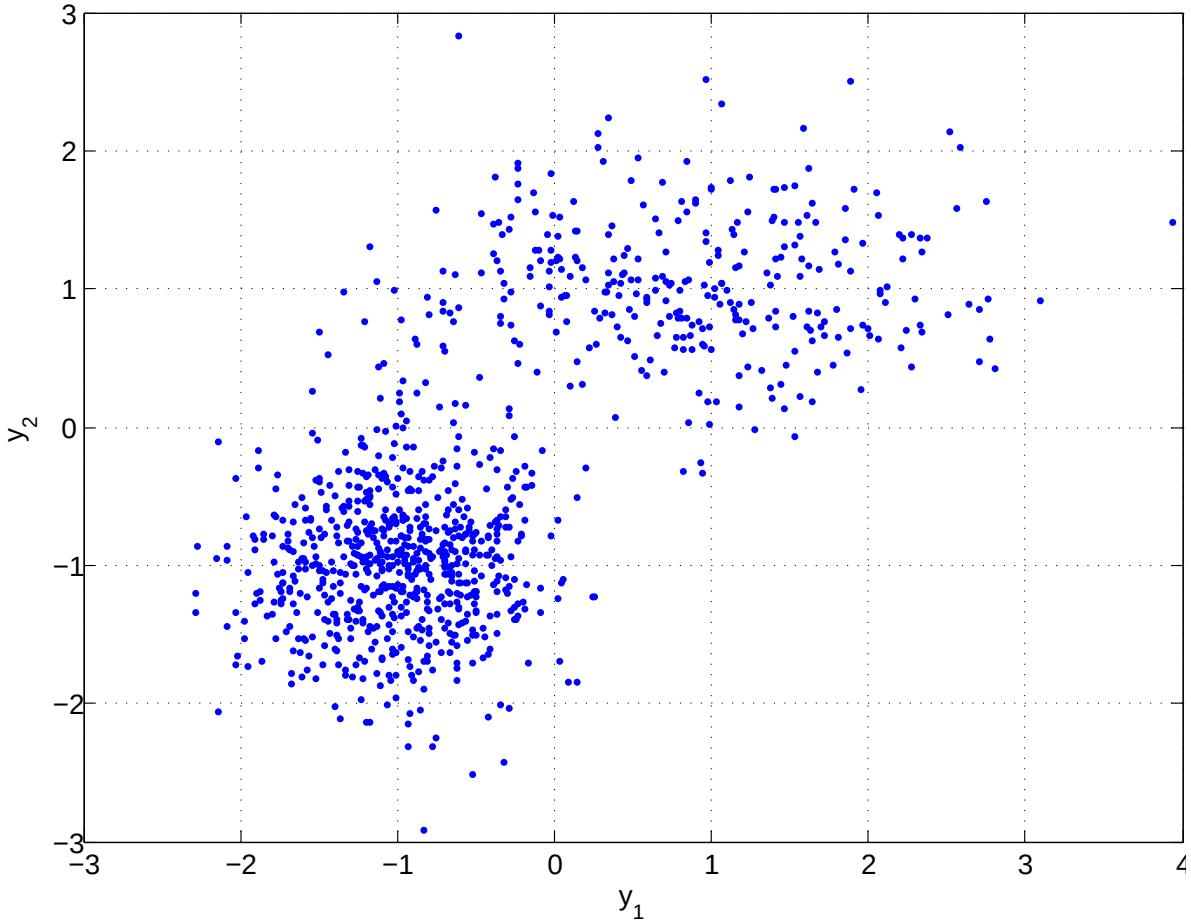
- Assigns items to class with most similar mean vector
- Similarity between item i and centroid k is

$$d_k(i) = \left(\sum_{j=1}^{q} (y_{i,j} - \mu_{k,j})^2 \right)^{\frac{1}{2}}$$

$\Rightarrow$ Euclidean distance between the vectors.

K-means Clustering (2)

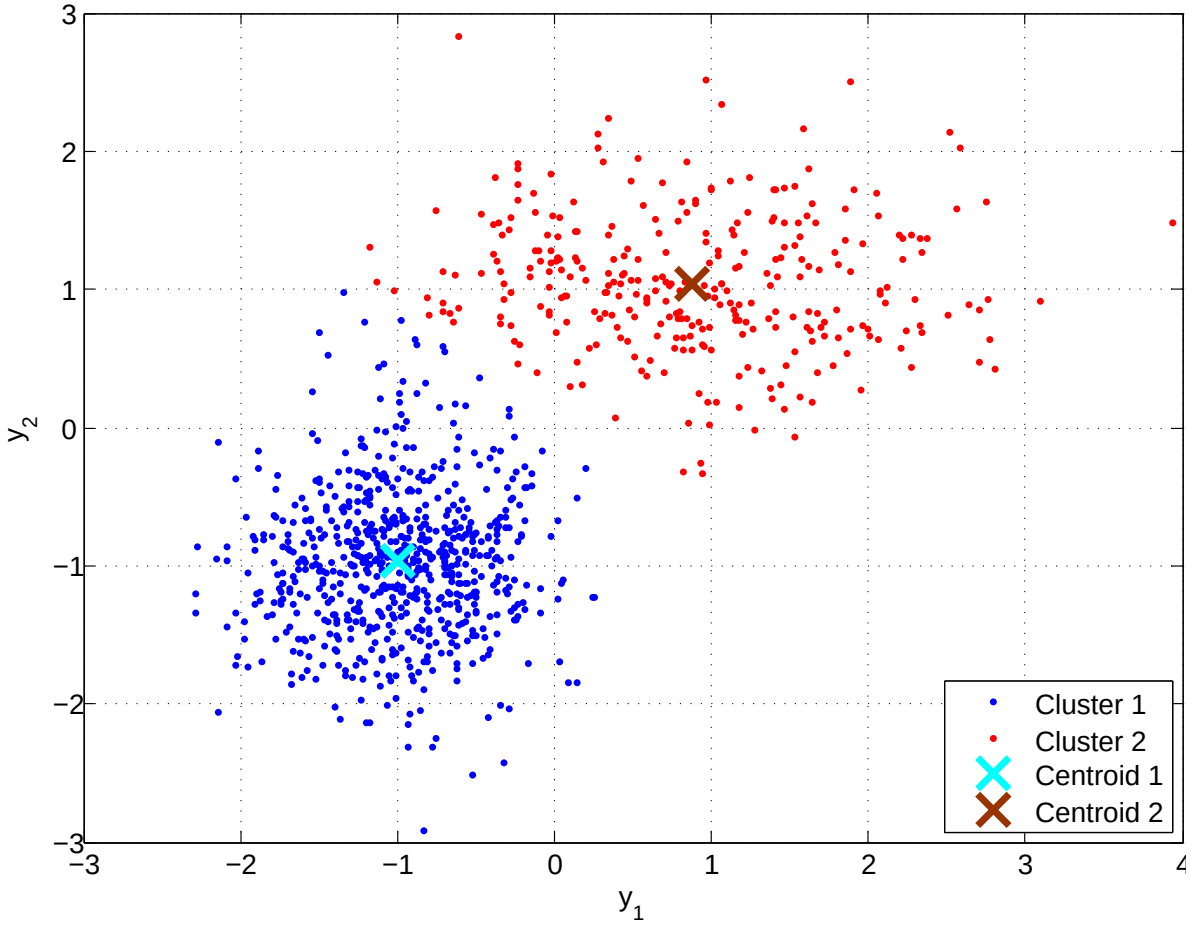
- Artificial data example



- Chosen so that the “clusters” are obvious for demonstration purposes

K-means Clustering (3)

- K-means clustering with $K = 2$



- The centroids are chosen so that the within-cluster sum-of-squares is minimised

Further Clustering

- Alternative similarity measures
  - Weighted Euclidean distance
  - “Cityblock” distance
  - Hamming distance (for pure binary data)
  - and many more ...

Further Clustering

- Alternative similarity measures
  - Weighted Euclidean distance
  - “Cityblock” distance
  - Hamming distance (for pure binary data)
  - and many more ...
- Some potential issues
  - “Hard” classification of items to clusters
  - Difficult to handle mixed attributes (continuous, discrete)
  - No explicit statistical interpretation
  - How to choose K using just the data ?
- Mixture modelling a flexible alternative

Content

Mixture Modelling (1)

- Models data as a mixture of probability distributions

$$p(y_{i,j}) = \sum_{k=1}^{K} \alpha_k p(y_{i,j}; \boldsymbol{\theta}_{k,j})$$

where

  - K is the number of classes
  - $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ are the mixing (population) weights
  - $\boldsymbol{\theta}_{k,j}$ are the parameters of the distributions

- Has an explicit probabilistic form
  $\Rightarrow$ allows for statistical interpretion

Mixture Modelling (2)

- How is this related to clustering ?
- Each class is a cluster
  - Class-specific probability distributions over each attribute
    - e.g., normal, inverse Gaussian, Poisson, etc.
  - Mixing weight is prevalance of items in the class

Mixture Modelling (2)

- How is this related to clustering ?
- Each class is a cluster
  - Class-specific probability distributions over each attribute
    - e.g., normal, inverse Gaussian, Poisson, etc.
  - Mixing weight is prevalance of items in the class
- Measure of similarity of item to class

$$p_k(\mathbf{y}_i) = \prod_{j=1}^{q} p(y_{i,j}; \boldsymbol{\theta}_{k,j})$$

$\Rightarrow$ probability of item's attributes under class distributions

Mixture Modelling (3)

- Membership of items to classes is soft

$$r_{i,k} = \frac{\alpha_k \, p_k(\mathbf{y}_i)}{\sum_{l=1}^{K} \alpha_l \, p_l(\mathbf{y}_i)}$$

- Posterior probability of belonging to class k
  - α_k is *a priori* probability item belongs to class k
  - $p_k(\mathbf{y}_i)$ is probability of data item $\mathbf{y}_i$ under class k

$\Rightarrow$ Assign to class with highest posterior probability

Mixture Modelling (3)

- Membership of items to classes is soft

$$r_{i,k} = \frac{\alpha_k \, p_k(\mathbf{y}_i)}{\sum_{l=1}^{K} \alpha_l \, p_l(\mathbf{y}_i)}$$

- Posterior probability of belonging to class k
  - α_k is *a priori* probability item belongs to class k
  - $p_k(\mathbf{y}_i)$ is probability of data item $\mathbf{y}_i$ under class k

  $\Rightarrow$ Assign to class with highest posterior probability

- Total number of samples in a class is then

$$n_k = \sum_{i=1}^{n} r_{i,k}$$

Mixture Modelling (4)

- Mixture modelling seamlessly handles missing values
  $\Rightarrow$ They are ignored when computing similarity $p_k(\mathbf{y}_i)$!
- Mixture models allow for imputation
  - Use non-missing attributes to estimate class memberships
  - Impute missing attributes using class memberships

Mixture Modelling (4)

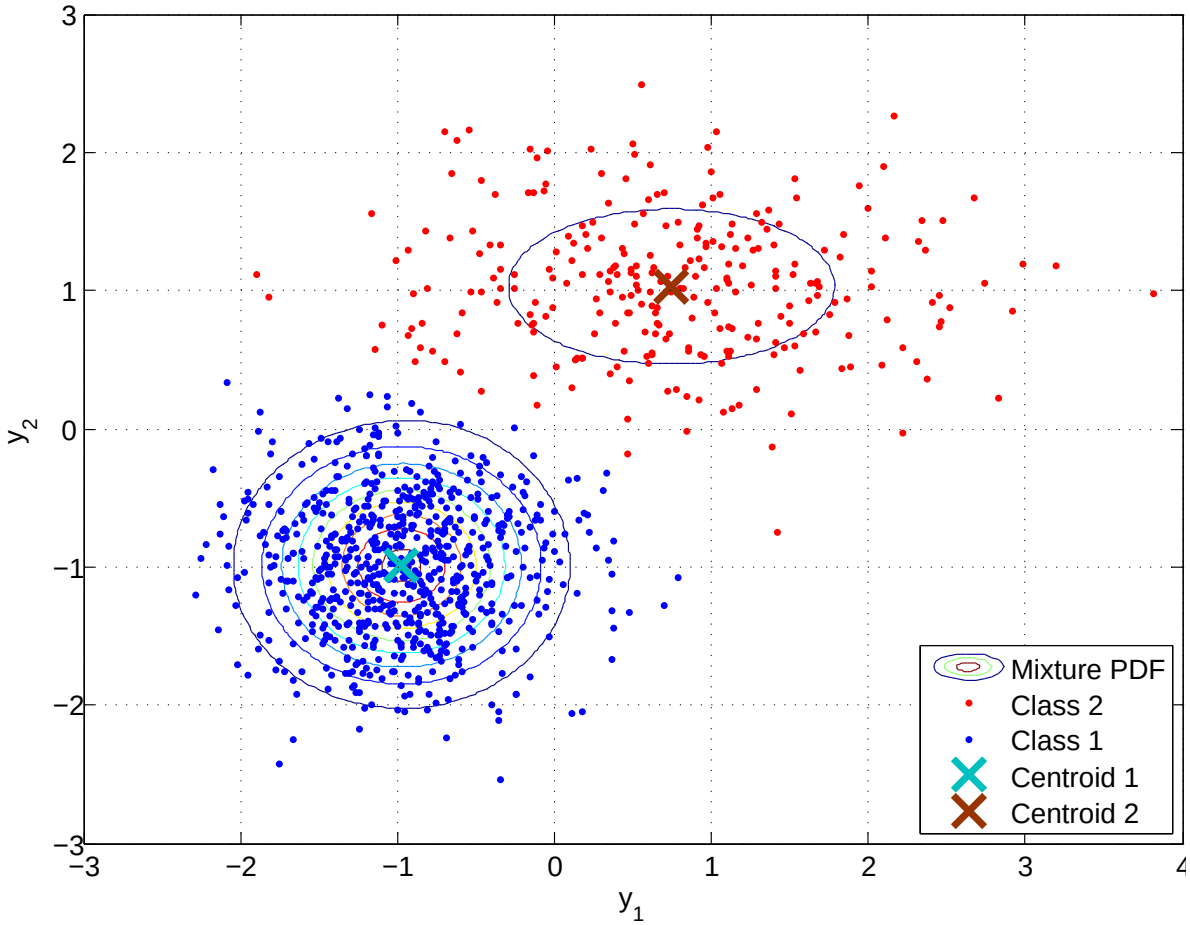
- Mixture modelling seamlessly handles missing values
  $\Rightarrow$ They are ignored when computing similarity $p_k(\mathbf{y}_i)$!
- Mixture models allow for imputation
  - Use non-missing attributes to estimate class memberships
  - Impute missing attributes using class memberships
- Can find probability density of missing attributes
  - Imagine for sample i that only attribute one is missing

$$p(y_{i,1}|y_{i,2},\ldots,y_{i,q}) = \sum_{k=1}^{K} r_{i,k}\; p(y_{i,1};\boldsymbol{\theta}_{k,1})$$

  - Can now impute $y_{i,1}$ using mode, or mean, for example
    $\Rightarrow$ if all attributes missing, reverts to normal procedure

Artificial Example

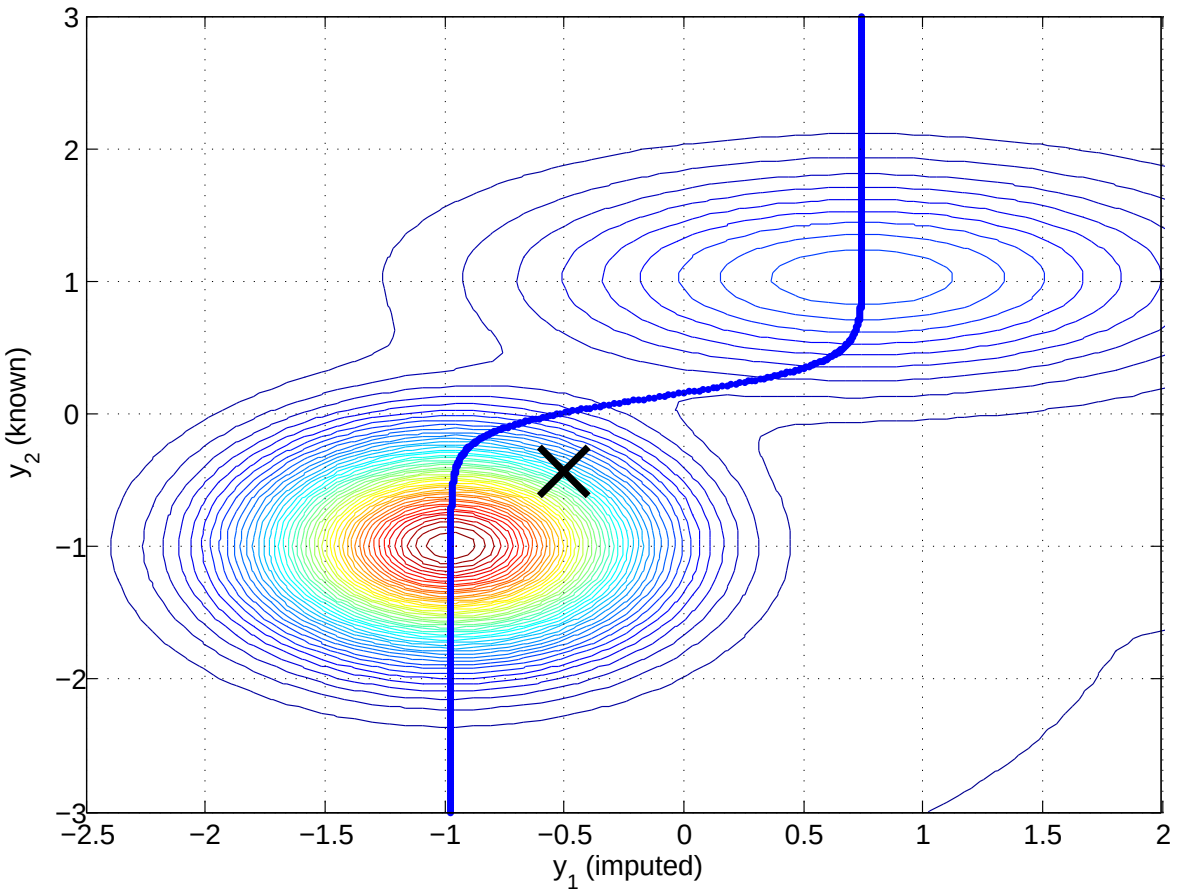
- Mixture modelling: automatically discovers $K = 2$



- Attributes y_1 and y_2 modelled as mixture of normals
  - Centroids are means of class-specific normal distributions
  - Estimated mixing weights: $\hat{\alpha} = (0.67, 0.32)$

Imputation Example

- Example of imputation
  - Attribute y_2 consider known
  - Impute attribute y_1 for values of y_2



Estimating Mixture Models (1)

- Given class memberships, the negative log-likelihood of data in class k is

$$-\sum_{i=1}^{n} r_{i,k} \sum_{j=1}^{q} \log p(y_{i,j}; \boldsymbol{\theta}_{k,j})$$

$\Rightarrow$ weighted negative log-likelihood

Estimating Mixture Models (1)

- Given class memberships, the negative log-likelihood of data in class k is

$$-\sum_{i=1}^{n} r_{i,k} \sum_{j=1}^{q} \log p(y_{i,j}; \boldsymbol{\theta}_{k,j})$$

  $\Rightarrow$ weighted negative log-likelihood

- Use expectation-maximisation (EM) algorithm:
  1. Estimate parameters, $\boldsymbol{\theta}_{k,j}$, $(k = 1, \ldots, K)$, $(j = 1, \ldots, q)$ using weighted negative log-likelihood
  2. Re-calculate class memberships based on new parameters
  3. If estimates have not stabilised, go to step (1)
- Initialise model with random class memberships

Estimating Mixture Models (2)

- Find K by minimising a goodness-of-fit criterion
- Difficult, non-convex optimisation problem
  $\Rightarrow$ Many local minima

Estimating Mixture Models (2)

- Find K by minimising a goodness-of-fit criterion
- Difficult, non-convex optimisation problem
  $\Rightarrow$ Many local minima
- Each iteration, do the following:
  - Remove classes with too few data points
  - Attempt to split all classes
  - Attempt to combine pairs of classes
  - Randomly assign data to classes, and re-estimate
- The mixture model with the smallest criterion score is retained, and the process is repeated

Estimating Mixture Models (3)

- Minimum Message Length goodness-of-fit criterion
  - Popular criterion for mixture modelling
- Based on the idea of compression

Estimating Mixture Models (3)

- Minimum Message Length goodness-of-fit criterion
  - Popular criterion for mixture modelling
- Based on the idea of compression
- Message length of data is our yardstick; comprised of
  1. Length of codeword needed to state model
  2. Length of codeword needed to state data, given model

  $\Rightarrow$ choose model which balances complexity against fit

Estimating Mixture Models (3)

- Minimum Message Length goodness-of-fit criterion
  - Popular criterion for mixture modelling
- Based on the idea of compression
- Message length of data is our yardstick; comprised of
  1. Length of codeword needed to state model
  2. Length of codeword needed to state data, given model

  $\Rightarrow$ choose model which balances complexity against fit
- Can be interpreted as negative log-posterior probability
  - Given two models, with message lengths I_1 and I_2,

$$\exp(-I_1 + I_2)$$

is approximate posterior-odds in favour of model 1

Interpreting Mixture Models (1)

- How different are the classes from each other ?
- Seperation matrix for attribute i

$$\mathbf{D}_i = \begin{pmatrix} 0 & d_i(1,2) & \dots & d_i(1,K) \\ d_i(2,1) & 0 & \dots & d_i(2,K) \\ \vdots & \vdots & \ddots & \vdots \\ d_i(K,1) & d_i(K,2) & \dots & 0 \end{pmatrix}$$

- $d_i(k_1, k_2)$ measures the seperation between class k_1 and k_2 for attribute i

Interpreting Mixture Models (1)

- How different are the classes from each other ?
- Seperation matrix for attribute i

$$\mathbf{D}_i = \begin{pmatrix} 0 & d_i(1,2) & \dots & d_i(1,K) \\ d_i(2,1) & 0 & \dots & d_i(2,K) \\ \vdots & \vdots & \ddots & \vdots \\ d_i(K,1) & d_i(K,2) & \dots & 0 \end{pmatrix}$$

- $d_i(k_1, k_2)$ measures the seperation between class k_1 and k_2 for attribute i
  - Seperation measured by Kullback–Leibler divergence
  - Analogue of distance between cluster centroids

Interpreting Mixture Models (1)

- How different are the classes from each other ?
- Seperation matrix for attribute i

$$\mathbf{D}_i = \begin{pmatrix} 0 & d_i(1,2) & \dots & d_i(1,K) \\ d_i(2,1) & 0 & \dots & d_i(2,K) \\ \vdots & \vdots & \ddots & \vdots \\ d_i(K,1) & d_i(K,2) & \dots & 0 \end{pmatrix}$$

- $d_i(k_1, k_2)$ measures the seperation between class k_1 and k_2 for attribute i
  - Seperation measured by Kullback–Leibler divergence
  - Analogue of distance between cluster centroids
- Can be used to determine which attributes define classes
- Seperation matrix for our artificial example:

$$\mathbf{D}_1 = \begin{pmatrix} 0 & 1.95 \\ 6.31 & 0 \end{pmatrix}, \quad \mathbf{D}_2 = \begin{pmatrix} 0 & 7.95 \\ 8.48 & 0 \end{pmatrix}$$

Interpreting Mixture Models (2)

- How different are the classes from the one-class model ?
- Significance matrix

$$\mathbf{S} = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,q} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ p_{K,1} & p_{K,2} & \cdots & p_{K,q} \end{pmatrix}$$

- $p_{k,i}$ is a "p-value" for class k, attribute i
  - Tests whether the model for attribute i in class k is significantly different from the one-class model
  - Likelihood ratio test based on χ^2 statistic
- A rough guide to see which attributes define a class

Content

Matlab Snob

- All examples done using Matlab Snob
- Currently being developed in Matlab
  - Flexible; easily allows addition of new distributions
  - Freely available
  - Stand-alone version in pipeline
- Currently implements following distributions:
  - Univariate Gaussian: for continuous data
  - Inverse Gaussian: for continous, positive data
  - Multinomial: for categorical data
  - Poisson: for non-negative integers

Pima Indians Diabetes Dataset

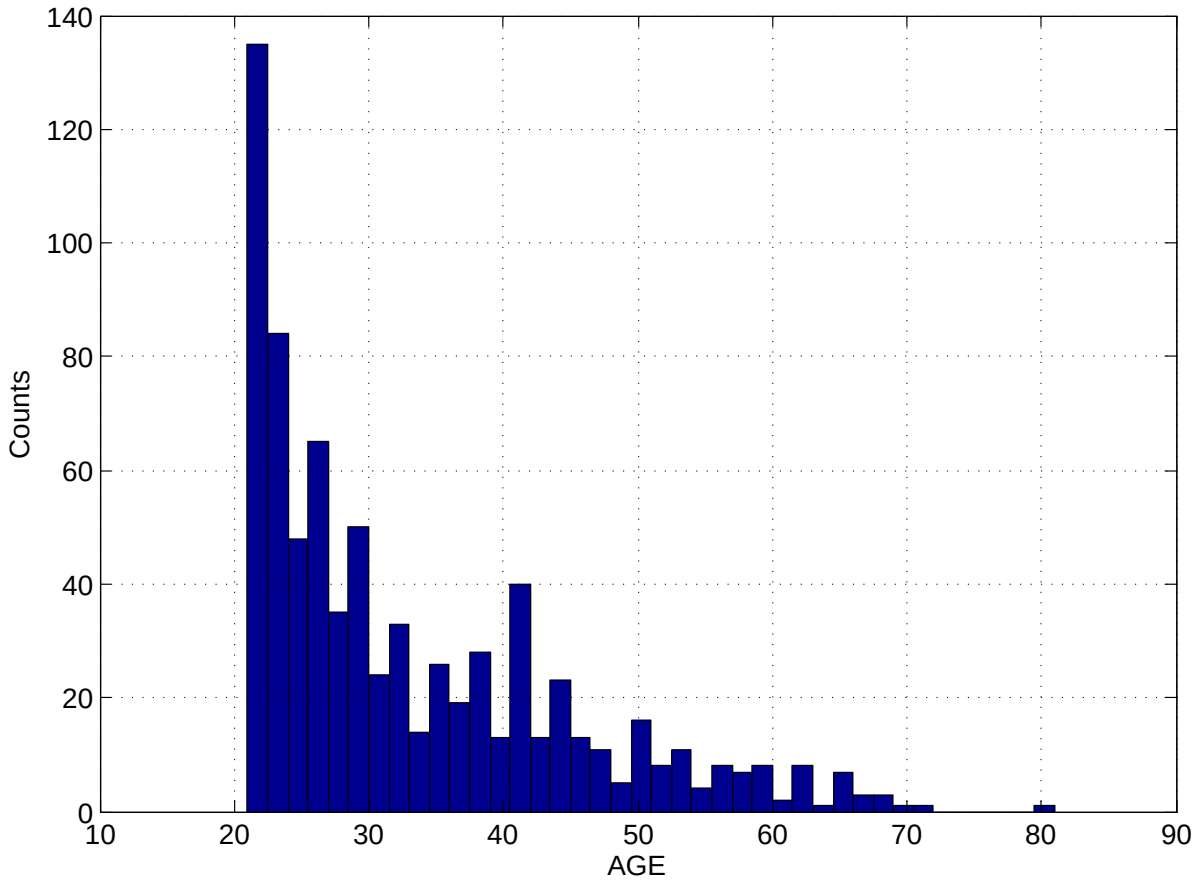
- Well known case-control dataset
  - 268 cases, 500 controls (1.86 controls per case)
  - 768 samples, with 8 exposures
  - 763 missing exposure measurements (12%)
- Outcome is diabetes in Pima indians (DIA)

TABLE : Pima Indians Exposures

Name	Mean	σ	Min	Max	% Missing
Number of Pregnancies (PREG)	4.5	3.2	1	17	14.4%
Plasma Glucose Concentration (PLAS)	121.6	30.5	44	199	0.6%
Diastolic Blood Pressure (BP)	72.4	12.4	24	122	4.5%
Triceps Skin Fold Thickness (SKIN)	29.1	10.5	7	99	29.5%
2-hour Serum Insulin (INS)	155.5	118.8	14	846	48.7%
Body Mass Index (BMI)	32.4	6.9	18.2	67.1	1.4%
Diabetes Pedigree Function (PED)	0.47	0.33	0.078	2.42	0%
Age (AGE)	33.2	11.7	21	81	0%

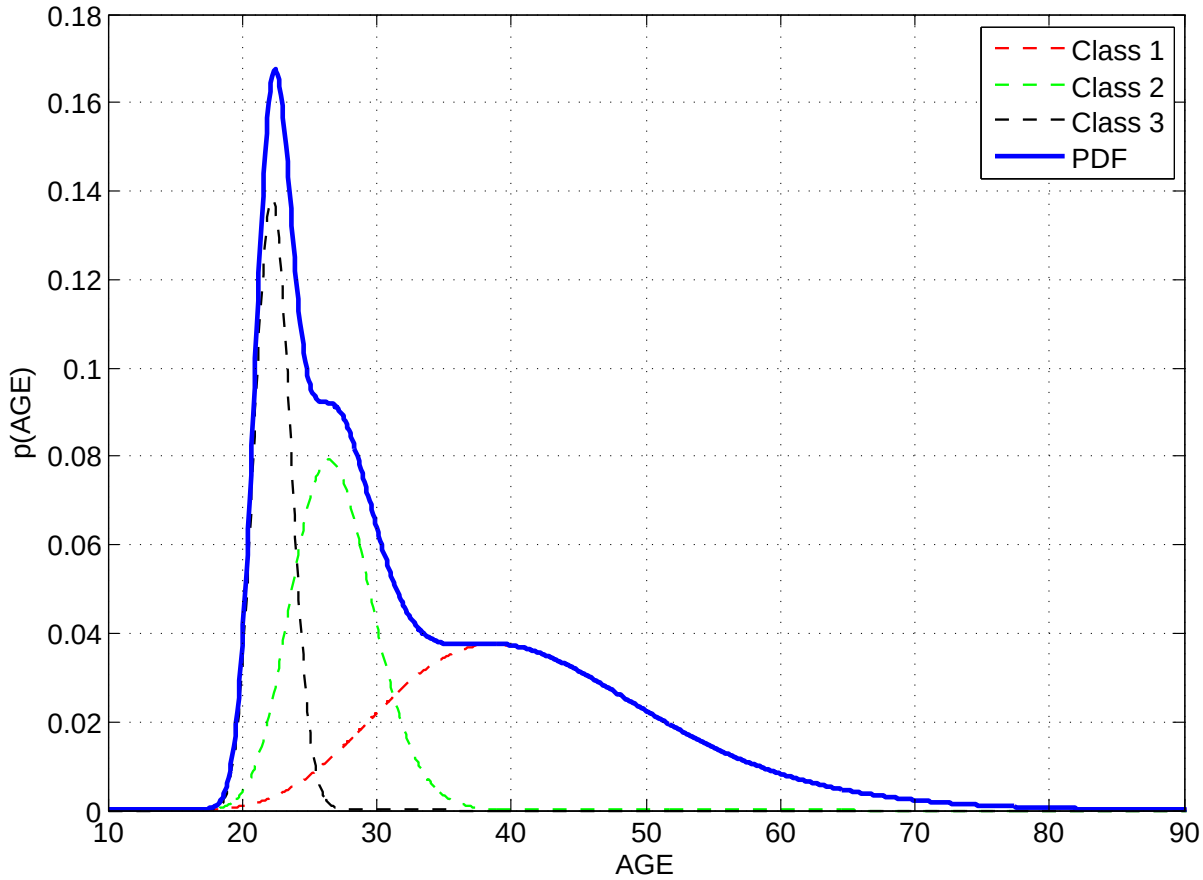
Example 1: Univariate Density Estimation (1)

- First consider 1-dimensional density estimation
  - Examine the AGE exposure
    $\Rightarrow$ clearly non-normal



Example 1: Univariate Density Estimation (2)

- Gaussian mixture: $\hat{K} = 4$, $I = 2{,}254.4$
- Inverse Gaussian mixture: $\hat{K} = 3$, $I = 2{,}237.9$
  - $\hat{\boldsymbol{\mu}} = (22.3, 26.9, 42.6)$, $\hat{\boldsymbol{\alpha}} = (0.23, 0.29, 0.47)$



Example 2: Multivariate Data Analysis (1)

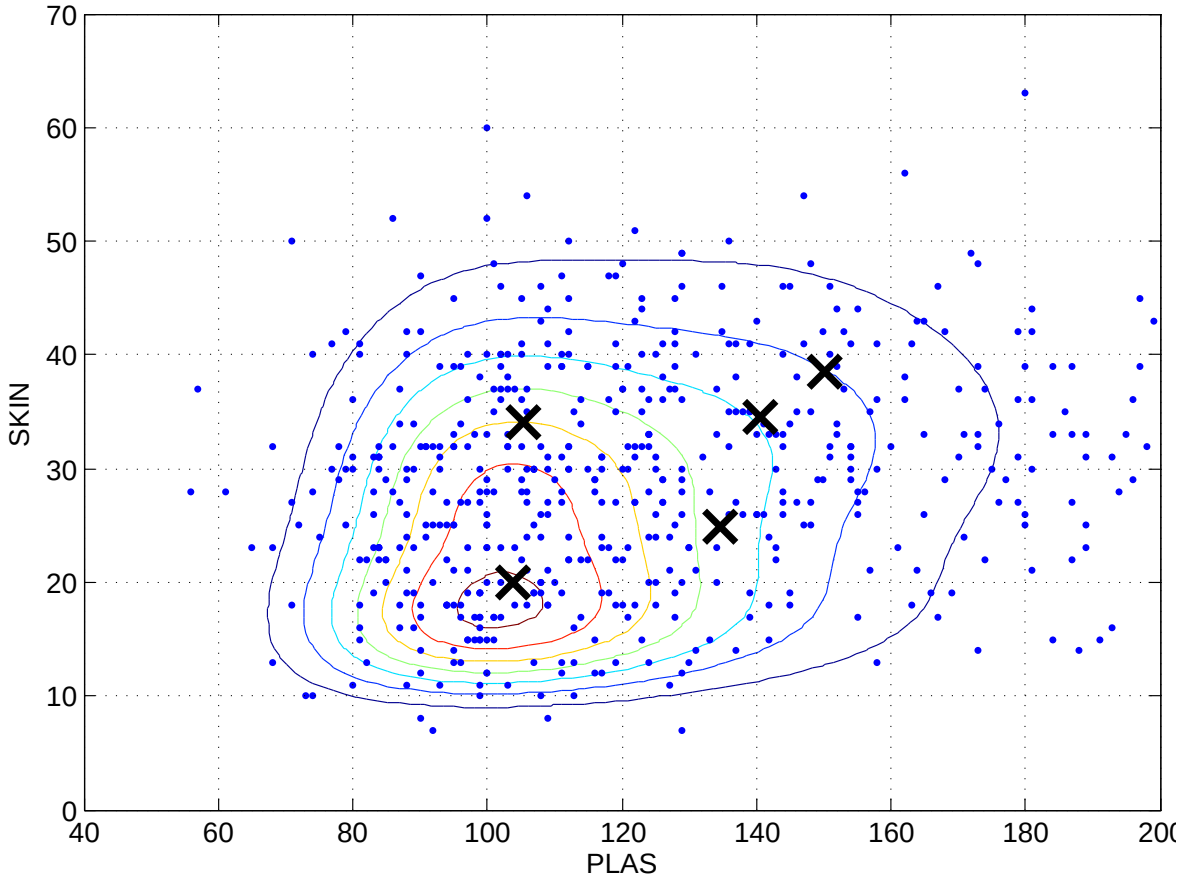
- Estimate mixture model for exposures and outcome
  - All exposures inverse Gaussian, outcome binomial
  - $I_4 = 18{,}719.1$, $I_5 = 18{,}713.0$, $I_6 = 18{,}714.7$, $I_7 = 18{,}732.7$

TABLE : Pima Indians Mixture Model (Means)

Class	$\hat{\alpha}_k$	PREG	PLAS	BP	SKIN	INS	BMI	PED	AGE	DIA
1	0.13	2.5	150	75	35	238	37	0.59	33	0.82
2	0.23	7.6	141	78	33	214	35	0.52	43	0.78
3	0.25	2.0	104	66	20	105	27	0.42	24	0.02
4	0.19	2.7	112	71	34	138	36	0.47	26	0.20
5	0.18	6.4	110	75	28	117	30	0.41	42	0.06

Example 2: Multivariate Data Analysis (2)

- Visualisation of mixture model for two attributes
  - PLAS vs SKIN



Example 2: Multivariate Data Analysis (3)

- Maximum seperation between classes
  $\Rightarrow$ help identify unimportant attributes

TABLE : Maximum Seperation between Classes

PREG	PLAS	BP	SKIN	INS	BMI	PED	AGE	DIA
8.53	2.02	0.68	2.86	1.49	2.73	0.18	10.60	2.68

- Suggests PED never defines a class and is superfluous

Example 2: Multivariate Data Analysis (4)

- Classes 2 and 5
  - Similar age (42 vs 43), different rates (0.78 vs 0.06)

TABLE : Seperation between classes 2 and 5

PREG	PLAS	BP	SKIN	INS	BMI	PED	AGE	DIA
0.34	1.44	0.07	0.60	1.52	0.72	0.13	0.1	2.77

Example 2: Multivariate Data Analysis (4)

- Classes 2 and 5
  - Similar age (42 vs 43), different rates (0.78 vs 0.06)

TABLE : Seperation between classes 2 and 5

PREG	PLAS	BP	SKIN	INS	BMI	PED	AGE	DIA
0.34	1.44	0.07	0.60	1.52	0.72	0.13	0.1	2.77

- Classes 3 and 4
  - Similar age (24 vs 26), different rates (0.02 vs 0.2)

TABLE : Seperation between class 3 and 4

PREG	PLAS	BP	SKIN	INS	BMI	PED	AGE	DIA
0.21	0.16	0.23	4.26	0.44	4.4	0.07	0.33	0.41

Example 2: Multivariate Data Analysis (5)

- Odds and odds-ratios for the classes
  - Odds-ratios relative to class with smallest odds

TABLE : Odds and Odds-Ratios

Class	1	2	3	4	5
Odds	4.711	3.327	0.022	0.243	0.066
Odds-Ratio	208	147	1	10.7	2.91

- Logistic regression interpretation
  - Logistic regression with one K-category exposure
  - Categories determined from structure of data

Example 3: Logistic regression and mixture models (1)

- Logistic regression and mixture modelling
  - Complementary
- Imputation of missing exposures
  - Use mixture model to impute missing exposures
  - Generates a probability density over exposures
    $\Rightarrow$ Sample missing data for multiple imputation
- Can now fit a logistic regression model

Example 3: Logistic regression and mixture models (2)

- Logistic regression classifier vs mixture model classifier
- Classifying with mixture models
  - Treat the outcome as missing
  - Use mixture model to “impute” missing outcomes
- Comparison to logistic regression
  - Imputed exposures using mixture model
  - Fit logistic regression, AUC = 0.86
  - Mixture model classifier, AUC = 0.85
    $\Rightarrow$ essentially same AUC, data reduced to only five groups

Future Work

- Release completed first version on Matlab exchange
- Handle more distributions
  - Multivariate normal
  - Overdispersed Poisson (for counts)
  - Gamma/Weibull distribution
  - Gaussian Processes
  - Mixtures of linear/logistic regressions
- Write stand-alone version

References

- Wallace, C. S., Boulton, D. M. “An information measure for classification”. *Computer Journal*, 1968, Vol. 11, pp. 185-194
- Wallace, C. S., Dowe, D. L. “MML mixture modelling of multi-state, Poisson, von Mises circular and Gaussian distributions”. *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics*, 1997, pp. 529-536
- Wallace, C. S. “Intrinsic Classification of Spatially Correlated Data”. *The Computer Journal*, 1998, Vol. 41, pp. 602-611
- Wallace, C. S., Dowe, D. L., “MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions”. *Statistics and Computing*, 2000, Vol. 10, pp. 73-83
- Wallace, C. S. “Statistical and Inductive Inference by Minimum Message Length”, *Springer*, 2005
- Schmidt, D. F., Makalic, E., “Minimum Message Length Inference and Mixture Modelling of Inverse Gaussian Distributions”, *under review*