Fuzzy Classification

The 'Iris'- and 'Admission'- Data Sets

Olaf Wolkenhauer

Control Systems Centre

UMIST

o.wolkenhauer@umist.ac.uk

www.csc.umist.ac.uk/people/wolkenhauer.htm

Contents

1. The Iris-Data Set

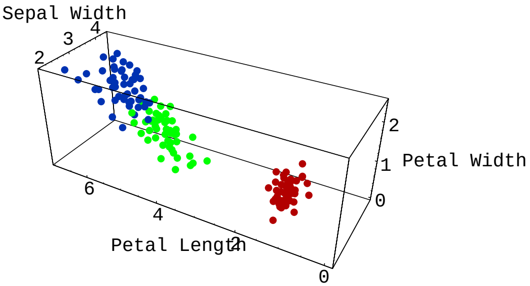
In his pioneering work on discriminant functions, Fisher presented data collected by Anderson on three species of iris flowers [3]. Let the classes be defined as :

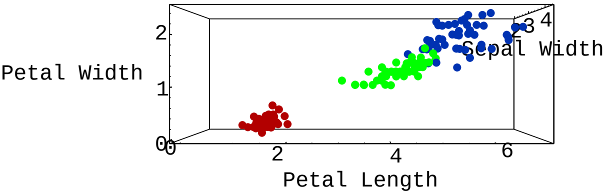
$$C_1: \text{ Iris sestosa}; \qquad C_2: \text{ Iris versicolor}; \qquad C_3: \text{ Iris virginica.}$$

For the following four variables 150 measurements were taken :

- Sepal length sl
- Sepal width sw
- Petal length pl
- Petal width pw

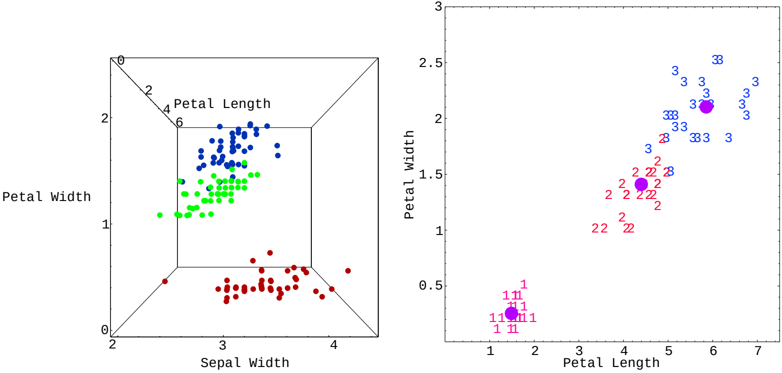
1.1. Visual Representation I





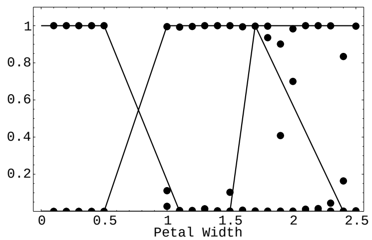
1.2. Visual Representation II

- ▷ Left: Full data set.
- ▷ Right: Training data set and fuzzy-c-means cluster centres for $w = 1.5$, $c = 3$, stopping criteria 0.01, 11 iterations.

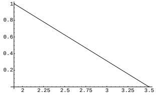


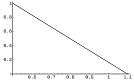
2. Orthogonal Projection

Orthogonal projection of cluster membership degrees and fitted piecewise-linear membership function.

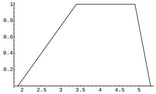


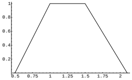
3. Rule-Based Fuzzy Classifier

R_1 : IF pw is 

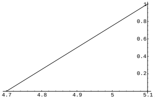
AND pl is 

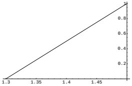
,
THEN iris sestosa.

R_2 : IF pw is 

, AND pl is 

,
THEN iris versicolor.

R_3 : IF pw is 

, AND pl is 

,
THEN iris virginica.

3.1. Fuzzy Decision Making

▷ Degree of *confidence* that data vector $\mathbf{x}$ belongs to class C_i :

$$\beta_i(\mathbf{x}) \doteq \mu_{A_{i1}}(x_1) \wedge \mu_{A_{i2}}(x_2) \wedge \cdots \wedge \mu_{A_{ir}}(x_r) \ .$$

▷ Allocatory rule :

$$C^* = \arg\max_i \beta_i(\mathbf{x}) \ .$$

4. The Admission-Data Set

The admission officer of a business school [3] has used an "index" of

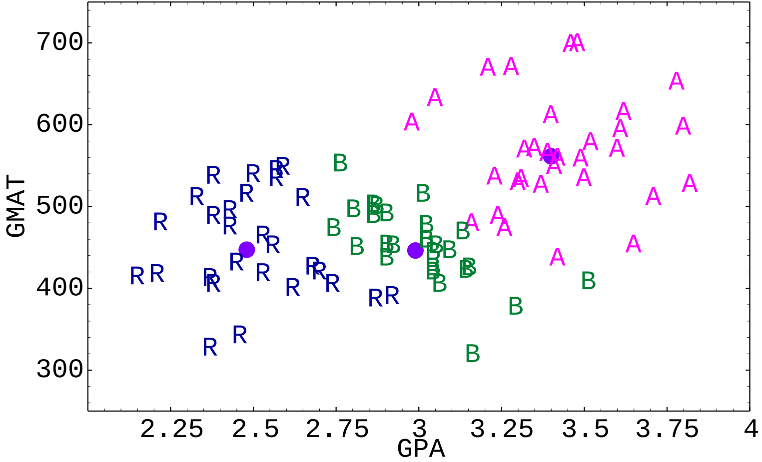
GPA:	Grade Point Average scores,
GMAT:	Graduate Management Aptitude Test score.

to help decide when applicants should be admitted to the school's graduate programs.

For 85 students the admission officer made a decision by classifying the applicants into three groups :

- R: Reject.
- A: Admit.
- B: Borderline.

4.1. Visual Representation



4.2. Questions

- We are given a set of *labelled* training data.
  ▷ How do we 'automatically' discriminate among students?
- What about *unlabelled* training data?
  ▷ Can we cluster data into 'natural' classes?
- For reasons of fairness, a "borderline" group is created.
  ▷ Does this remove unfairness?
- What are the problems with formal methods?

Let a (general) data point be denoted by

$$\mathbf{x} = (x_1 = \text{GPA}, x_2 = \text{GMAT})$$

Given the set of training vectors $\mathbf{m}_j =, j = 1, \ldots, 85$, we wish to group the data into $c = 3$ classes

▷ C_1 – admit; ▷ C_2 – do not admit; ▷ C_3 – borderline.

5. Linear Discriminant Analysis

Decision Rule: Assign $\mathbf{x}$ to the closest population, i.e. to the class C_i for which

$$-\frac{1}{2} d^2_{\mathbf{\Sigma}_{\text{pooled}}}(\mathbf{x}, \mathbf{c}_i) + \ln p_i$$

is largest [3]. Where p_i is the prior probability of C_i and the distance of $\mathbf{x}$ to the sample mean vector $\mathbf{c}_i$ is calculated as

$$d^2_{\text{pooled}}\big(\mathbf{x}, \mathbf{c}_1\big) = \big(\mathbf{x} - \mathbf{c}_i\big)^T \mathbf{\Sigma}^{-1}_{\text{pooled}}\big(\mathbf{x} - \mathbf{c}_i\big)$$

and matrix $\mathbf{\Sigma}$ is the *pooled* estimate of the covariance matrix :

$$\mathbf{\Sigma}_{\text{pooled}} = \tfrac{1}{d_1 + d_2 + \cdots + d_c}\big((d_1 - 1)\mathbf{\Sigma}_1 + (d_2 - 1)\mathbf{\Sigma}_2 + \cdots + (d_c - 1)\mathbf{\Sigma}_c\big)$$

and

d_i	:	sample size,
$\mathbf{\Sigma}_i$	:	sample covariance matrix for population C_i.

5.1. Example

Let a candidate have the following scores :

$$x_1 = 3.21 \quad \text{(GPA)} \qquad x_2 = 497 \quad \text{(GMAT)} .$$

Using a statistical software package :

$$d_1 = 31 \qquad d_2 = 28 \qquad d_3 = 26$$

$$\mathbf{c}_1 = \begin{bmatrix} 3.40 \\ 561.23 \end{bmatrix} \qquad \mathbf{c}_2 = \begin{bmatrix} 2.48 \\ 447.07 \end{bmatrix} \qquad \mathbf{c}_3 = \begin{bmatrix} 2.99 \\ 446.23 \end{bmatrix}$$

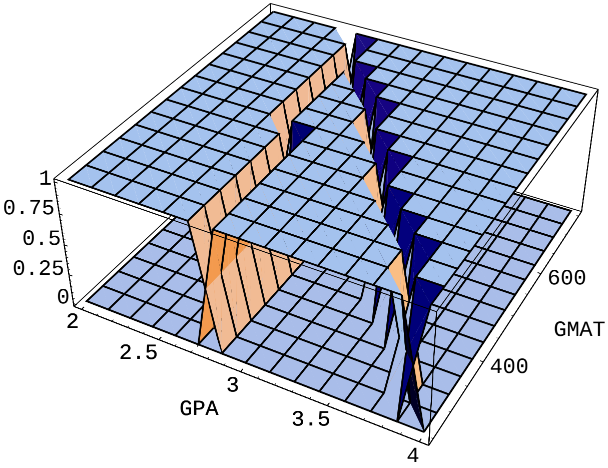
$$\mathbf{c} = \begin{bmatrix} 2.97 \\ 488.45 \end{bmatrix} \qquad \mathbf{\Sigma}_{\text{pooled}} = \begin{bmatrix} 0.0361 & -2.0188 \\ -2.0188 & 3655.9011 \end{bmatrix}$$

For $\mathbf{x} = [3.21, 497]^T$, the sample distances to population means are

$$d^2_{\text{pooled}}(\mathbf{x}, \mathbf{c}_1) = 2.58 \quad d^2_{\text{pooled}}(\mathbf{x}, \mathbf{c}_2) = 17.10 \quad d^2_{\text{pooled}}(\mathbf{x}, \mathbf{c}_3) = 2.47 ✔$$

Since the distance to class mean $\mathbf{c}_3$ is smallest, the Business School applicant is assigned to C_3, is considered a "borderline case".

5.2. Decision Surface



5.3. Problems

- ✘ We do not know the prior probabilities p_i.
  - ▷ Assume $p_1 = p_2 = \cdots = p_c = 1/c$.
- ✘ What is a *population* of business students?
- ✘ Requires labelled training data.
- ✘ For borderline cases a new class is created.

The main advantage of a statistical framework is that one can prove *properties of the classifier analytically.*

6. Fuzzy Clustering

The fuzzy-c-means algorithm [2, 1] returns a partition matrix $\mathbf{U}$ which can serve as a model for a classifier. With $u_{ij} \in \mathbf{U}$, the final cluster centres are obtained as

$$\mathbf{c}_i = \frac{\sum_{j=1}^{85} (u_{ij})^w \ \mathbf{m}_j}{\sum_{j=1}^{85} (u_{ij})^w} \ , \quad i = 1, 2, \ldots, c \ .$$

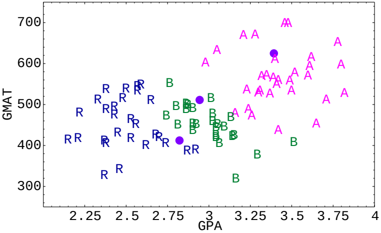
where c defines the number of clusters searched for and w is a weighting factor that determines the “fuzziness” of the clusters.

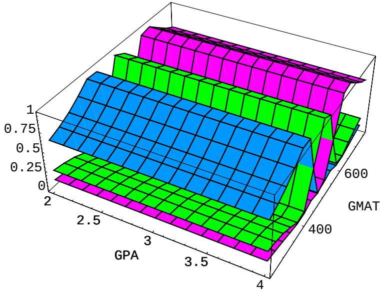
For any new applicant with scores $\mathbf{x} = [x_1 = \mathrm{GPA}, x_2 = \mathrm{GMAT}]^T$, the membership in each class is calculated as

$$\mu_{C_i}(\mathbf{x}) \doteq 1 \ / \ \sum_{k=1}^{c} \left(\frac{d(\mathbf{x}, \mathbf{c}_i)}{d(\mathbf{x}, \mathbf{c}_k)} \right)^{\frac{2}{w-1}}$$

6.1. Cluster Centres and Decision Surface

Weighting, Cluster Fuzziness	$w = 2$
Number of Classes	$c = 3$
Number of iterations	14





6.2. Problems

- ✘ Cluster centres are in the wrong place.
- ✘ The algorithm is sensitive w.r.t the scales of variables.
  ▷ Normalise or scale data.

For $c = 2$ and data set (matrix) $\mathbf{M} = \{\mathbf{m}_j\}$

$$u_{ij} = \frac{1}{\left(\frac{d(\mathbf{m}_j,\mathbf{c}_i)}{d(\mathbf{m}_j,\mathbf{c}_1)}\right)^{\frac{2}{w-1}} + \left(\frac{d(\mathbf{m}_j,\mathbf{c}_i)}{d(\mathbf{m}_j,\mathbf{c}_2)}\right)^{\frac{2}{w-1}}}$$

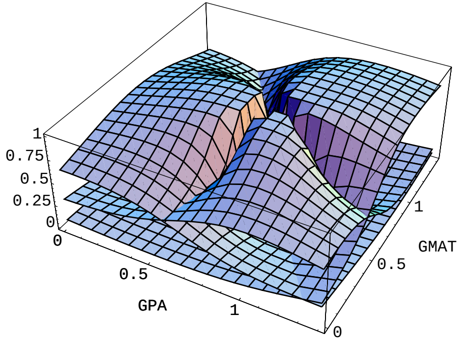
With $\mathbf{A}$ being the unity matrix, the Euclidean norm is

$$d^2_{\mathbf{A}}(\mathbf{m}_j, \mathbf{c}_1) = \|\mathbf{m}_j - \mathbf{c}_i\|^2 = (\mathbf{m}_j - \mathbf{c}_i)^T \mathbf{A}(\mathbf{m}_j - \mathbf{c}_i) \ .$$

The fuzzy-c-means algorithm uses distance measures iteratively which can lead to deceptive results if the scales of variables differ considerably.

6.3. Normalised Data

$w = 1.25$, $c = 3$, 17 iterations.

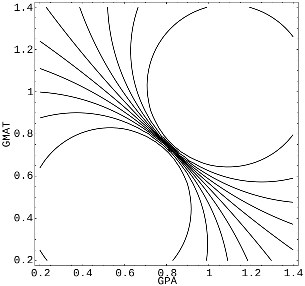


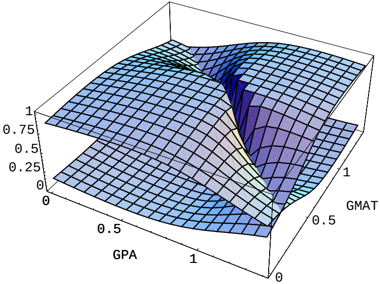
Problem

✘ What is the meaning of a fuzzy borderline-class?

6.4. Two Fuzzy Classes: "Reject" and "Admit"

$w = 1.25$, $c = 2$, normalised data.





The fuzzy-c-means algorithm, employing the Euclidean norm, searches for spherical clusters.

6.5. Remarks

For both $w = 1.25$ and $w = 2$, the cluster centres are

$$\mathbf{c}_1 = (0.9, 0.8)\ , \qquad \mathbf{c}_2 = (0.7, 0.6)\ .$$

Weighting $w = 1.25$ 8 iterations.
Weighting $w = 2$ 7 iterations.

For the test candidate with scores, $\mathbf{x}_j = (3.21, 497)$, the degrees of membership in the classes for $w = 1.25$ are

$$\mu_{C_1}(\mathbf{x}) = 0.73 \ ✔ \qquad \mu_{C_2}(\mathbf{x}) = 0.27$$

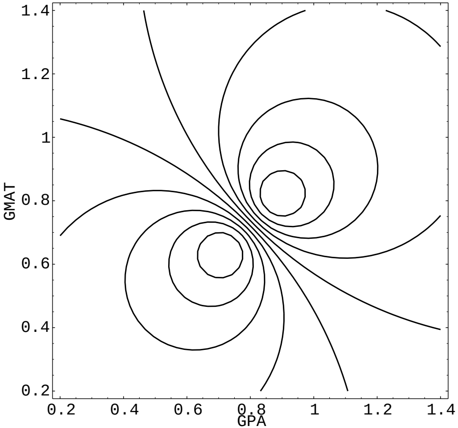
and for $w = 2$,

$$\mu_{C_1}(\mathbf{x}) = 0.67 \ ✔ \qquad \mu_{C_2}(\mathbf{x}) = 0.33\ .$$

The weighting factor w reflects the fuzziness in the decision making (student most probably would refer to w as the (un)fairness factor).

6.6. Contour Plot

Fuzzy c-means, normalised data, $w = 2$.



And finally...

▷ *Engineers and scientists will never make as much money as MBA's (Masters of Business Administration) and business executives.*

Now a rigorous mathematical proof that explains why this is true:

Postulate 1: Knowledge is power.

Postulate 2: Time is money.

As every engineer knows,

$$\frac{\text{Work}}{\text{Time}} = \text{Power}\ . \tag{1}$$

Since from postulate 1,

$$\text{Knowledge} = \text{Power} \tag{2}$$

and postulate 2,

$$\text{Time} = \text{Money} \tag{3}$$

inserting (2) and (3) into (1) we have

$$\frac{\text{Work}}{\text{Money}} = \text{Knowledge}. \tag{4}$$

Solving (4) for Money, we get

$$\frac{\text{Work}}{\text{Knowledge}} = \text{Money}\ .$$

▷ as Knowledge approaches zero, Money approaches infinity regardless of the Work done. Hence,

The less you know, the more you make.

References

[1] Babuska, R. : *Fuzzy Modelling for Control*. Kluwer, 1998. See http://lcewww.et.tudelft.nl/. 16

[2] Bezdek, J.C. : *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, 1981. 16

[3] Johnson, R.A. and Wichern, D.W. : *Applied Multivariate Statistical Analysis*. Prentice Hall, 4 ed. 1998. 3, 9, 12

[4] Wolkenhauer, O. : *Possibility Theory with Applications to Data Analysis*. Research Studies Press, 1998.

[5] Wolkenhauer, O. : *Data Engineering: Data, Systems and Uncertainty*. Book manuscript, 1999. See http://www.csc.umist.ac.uk/people/wolkenhauer.htm.