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 3
   1.1 Visual Representation I .................... 4
   1.2 Visual Representation II ................... 5

2 Orthogonal Projection 6

3 Rule-Based Fuzzy Classifier 7
   3.1 Fuzzy Decision Making ................... 8

4 The Admission-Data Set 9
   4.1 Visual Representation .................... 10
   4.2 Questions .............................. 11

5 Linear Discriminant Analysis 12
   5.1 Example ................................... 13
   5.2 Decision Surface .......................... 14
   5.3 Problems ................................. 15
6 Fuzzy Clustering

6.1 Cluster Centres and Decision Surface ........................................... 17
6.2 Problems ......................................................................................... 18
6.3 Normalised Data ............................................................................... 19
6.4 Two Fuzzy Classes: “Reject” and “Admit” ........................................ 20
6.5 Remarks ............................................................................................. 21
6.6 Contour Plot ..................................................................................... 22
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:

\[ C_1: \text{Iris setosa}; \quad C_2: \text{Iris versicolor}; \quad C_3: \text{Iris virginica.} \]

For the following four variables 150 measurements were taken:

- Sepal length \( \text{sl} \)
- Sepal width \( \text{sw} \)
- Petal length \( \text{pl} \)
- Petal width \( \text{pw} \)
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

\[ R_1 : \text{IF } pw \text{ is } \text{AND } pl \text{ is } \text{THEN iris sestosa.} \]

\[ R_2 : \text{IF } pw \text{ is } \text{AND } pl \text{ is } \text{THEN iris versicolor.} \]

\[ R_3 : \text{IF } pw \text{ is } \text{AND } pl \text{ is } \text{THEN iris virginica.} \]
3.1. Fuzzy Decision Making

▷ Degree of confidence that data vector $x$ belongs to class $C_i$:

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

▷ Allocatory rule:

$$C^* = \arg \max_i \beta_i(x).$$
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
\[ x = (x_1 = \text{GPA}, x_2 = \text{GMAT}) \]

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 \( x \) to the closest population, i.e. to the class \( C_i \) for which

\[
-\frac{1}{2} d^2_{pooled}(x, c_i) + \ln p_i
\]

is largest [3]. Where \( p_i \) is the prior probability of \( C_i \) and the distance of \( x \) to the sample mean vector \( c_i \) is calculated as

\[
d^2_{pooled}(x, c_i) = (x - c_i)^T \Sigma^{-1}_{pooled}(x - c_i)
\]

and matrix \( \Sigma \) is the pooled estimate of the covariance matrix:

\[
\Sigma_{pooled} = \frac{1}{\frac{1}{d_1} + \frac{1}{d_2} + \cdots + \frac{1}{d_c}} \left( (d_1 - 1) \Sigma_1 + (d_2 - 1) \Sigma_2 + \cdots + (d_c - 1) \Sigma_c \right)
\]

and

- \( d_i \) : sample size,
- \( \Sigma_i \) : sample covariance matrix for population \( C_i \).
Section 5: Linear Discriminant Analysis

5.1. Example

Let a candidate have the following scores:

\[ x_1 = 3.21 \quad \text{(GPA)} \quad x_2 = 497 \quad \text{(GMAT)}. \]

Using a statistical software package:

\[
\begin{align*}
  d_1 &= 31 & d_2 &= 28 & d_3 &= 26 \\
  c_1 &= \begin{bmatrix} 3.40 \\ 561.23 \end{bmatrix} & c_2 &= \begin{bmatrix} 2.48 \\ 447.07 \end{bmatrix} & c_3 &= \begin{bmatrix} 2.99 \\ 446.23 \end{bmatrix} \\
  c &= \begin{bmatrix} 2.97 \\ 488.45 \end{bmatrix} & \Sigma_{\text{pooled}} &= \begin{bmatrix} 0.0361 & -2.0188 \\ -2.0188 & 3655.9011 \end{bmatrix}
\end{align*}
\]

For \( x = [3.21, 497]^T \), the sample distances to population means are:

\[ d^2_{\text{pooled}}(x, c_1) = 2.58 \quad d^2_{\text{pooled}}(x, c_2) = 17.10 \quad d^2_{\text{pooled}}(x, c_3) = 2.47 \]

Since the distance to class mean \( c_3 \) is smallest, the Business School applicant is assigned to \( C_3 \), is considered a “borderline case”.

\begin{table}
\end{table}
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 $U$ which can serve as a model for a classifier. With $u_{ij} \in U$, the final cluster centres are obtained as

$$c_i = \frac{\sum_{j=1}^{85} (u_{ij})^w m_j}{\sum_{j=1}^{85} (u_{ij})^w}, \quad i = 1, 2, \ldots, c,$$

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 $x = [x_1 = \text{GPA}, x_2 = \text{GMAT}]^T$, the membership in each class is calculated as

$$\mu_{C_i}(x) = \frac{1}{\sum_{k=1}^{c} \left( \frac{d(x, c_i)}{d(x, c_k)} \right)^{\frac{2}{w-1}}}$$
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) \( M = \{ \mathbf{m}_j \} \)

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

With \( A \) being the unity matrix, the Euclidean norm is

\[
d_A^2(\mathbf{m}_j, \mathbf{c}_1) = \| \mathbf{m}_j - \mathbf{c}_i \|^2 = (\mathbf{m}_j - \mathbf{c}_i)^T A (\mathbf{m}_j - \mathbf{c}_i) .
\]

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

\[ w = 1.25, \ c = 3, \ 17 \ \text{iterations}. \]

**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

$$c_1 = (0.9, 0.8), \quad c_2 = (0.7, 0.6).$$

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

For the test candidate with scores, $x_j = (3.21, 497)$, the degrees of membership in the classes for $w = 1.25$ are

$$\mu_{C_1}(x) = 0.73 \quad \checkmark \quad \mu_{C_2}(x) = 0.27$$

and for $w = 2$,

$$\mu_{C_1}(x) = 0.67 \quad \checkmark \quad \mu_{C_2}(x) = 0.33.$$  

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,

\[ \text{The less you know, the more you make.} \]
References


    See http://www.csc.umist.ac.uk/people/wolkenhauer.htm.