MACHINE LEARNING IN AVIATION

PEDRO LARRAÑAGA

Computational Intelligence Group
Artificial Intelligence Department
Technical University of Madrid, Spain

Toulouse, July 11, 2013
Outline

1. Introduction
2. Supervised Classification
3. Clustering
4. Conclusions
Outline

1. Introduction
2. Supervised Classification
3. Clustering
4. Conclusions
Introduction

Data in aviation

- **Data in our world**: smartphones, social networks, atmospheric readings, financial data, medical records, bioinformatics, airplane instrumentation, etc.
- Estimation of **2.5 exabytes of data** generated on the planet **every day**
- **Data in aviation**:
  - Flight tracking
  - Weather conditions
  - Airport information
  - Airline information
  - Market information
  - Passenger information
  - Air safety reports
  - Aircraft data
Introduction

Data in USA cross-country commercial flights

Sensor data from a cross-country flight

- 20 terabytes of information per engine every hour
- Two-engine Boeing 737
- Six-hour, cross-country flight from New York to Los Angeles
- Number of commercial flights in the sky in the United States on any given day

= 2,499,841,200 TB
Introduction

Big Data. The five V’s: volume, variety, velocity, viability and value

Diagram 1:
Big data

DAVENPORT, TH (2013). *At the Big Data Crossroads: Turning Towards a Smarter Travel Experience*. Amadeus IT Group

P. Larrañaga
Machine Learning in Aviation
## Aviation industry

**Characteristics:** Large scale and unstructured mixture of data in a variety of data formats

- Radar
- Weather (spatial-temporal)
- Terrain (spatial)
- Infrastructure
- Text

**Benefits:**

- Help airlines increase sales
- Customer loyalty
- Improve fuel management and efficiency
- Manage fleets more effectively
- Improve customer service
Pattern recognition (PR) a step in knowledge discovery in databases (KDD)
Outline

1. Introduction

2. Supervised Classification

3. Clustering

4. Conclusions
Supervised: From labelled data to classification models

Predictor variables (attributes) and one labelled (class) variable:

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>...</th>
<th>$X_n$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(x^{(1)}, c^{(1)})$</td>
<td>$x_1^{(1)}$</td>
<td>...</td>
<td>$x_n^{(1)}$</td>
<td>$c^{(1)}$</td>
</tr>
<tr>
<td>$(x^{(2)}, c^{(2)})$</td>
<td>$x_1^{(2)}$</td>
<td>...</td>
<td>$x_n^{(2)}$</td>
<td>$c^{(2)}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$(x^{(N)}, c^{(N)})$</td>
<td>$x_1^{(N)}$</td>
<td>...</td>
<td>$x_n^{(N)}$</td>
<td>$c^{(N)}$</td>
</tr>
<tr>
<td>$x^{(N+1)}$</td>
<td>$x_1^{(N+1)}$</td>
<td>...</td>
<td>$x_n^{(N+1)}$</td>
<td>...</td>
</tr>
</tbody>
</table>

Construct a classification model that predicts with high accuracy the class of a new instance only characterized by the predictor variables.
Supervised classification

Linear decision boundary
Supervised classification

Non linear decision boundary
Supervised classification

Disease diagnosis

“How long have you been multitasking?”
Supervised classification

Prediction of the secondary structure of proteins
Supervised classification

Fraudulent use of credit cards
Supervised classification

Spam email
### Handwritten character recognition

```
9 7 4 0 2 5 2 / 1 6 8 8 8
6 7 8 8 8 7 9 0 0 1 1
3 0 1 2 3 4 5 6 7 8 8
9 0 1 2 3 4 5 6 7 8 8
0 1 2 5 6 0 9 9 9 6 1
3 4 5 6 7 8 9 0 1 2 2
6 7 8 9 0 1 2 3 4 5 5
/ / / / 1 2 3 4 5 5 6
4 6 4 5 6 7 7 7 8 9 9
9 9 8 9 9 1 2 4 5 5 5
6 7 8 9 4 5 6 3 2 2
3 3 3 8 8 4 4 9 6 0 0
7 0 0 8 6 5 8 9 5 6 8
8 8 1 8 9 4 6 7 8 9 9
4 2 2 5 7 8 3 6 4 1 1
```
Supervised classification

Weather forecasting

[Weather map images]
Less is more. Dimensionality reduction with principal component analysis (PCA)

- Directions of max variance of the data
- 3-D data, but distributed mostly on a 2-D surface ⇒ Find it
Supervised classification

Less is more. Feature subset selection (FSS)

- Requirement: **Scoring function to measure the quality of a subset**
- **Filter approach**: intrinsic characteristics of the data
- **Wrapper approach**: knowledge about the classifier

---

**Filter FSS**
- **Search**
  - Feature subset
  - Information content
  - Objective function
- **Final feature subset**
  - Classification algorithm

**Wrapper FSS**
- **Search**
  - Feature subset
  - Predictive accuracy
- **Classification algorithm**
  - Final feature subset
  - Classification algorithm
Less is more. Feature subset selection (FSS)

- FSS as a **combinatorial optimization** problem: search strategies (forward, backward, genetic algorithms, ....)
- Cardinality of the **search space**: $2^n$
### Measuring the performance. Confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>$C$ True class</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$+$</td>
<td>+</td>
<td>$-$</td>
</tr>
<tr>
<td>$a$</td>
<td>$b$</td>
<td></td>
</tr>
<tr>
<td>$c$</td>
<td>$d$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$C_M$ Predicted class</th>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>$b$</td>
<td></td>
</tr>
<tr>
<td>$c$</td>
<td>$d$</td>
<td></td>
</tr>
</tbody>
</table>

### Figures of merit

- **Accuracy:** \( \frac{a+d}{a+b+c+d} \)
- **Error rate:** \( \frac{c+b}{a+b+c+d} \)
- **Rate of true positives (sensitivity):** \( \frac{a}{a+c} \)
- **Rate of true negatives (specificity):** \( \frac{d}{b+d} \)
Supervised classification

ROC curve. Area under the ROC curve (AUC)
Supervised classification

Estimation methods. No honest

\[ \hat{p}_M = \frac{1}{N} \sum_{i=1}^{N} \delta(c^{(i)} = c_M^{(i)}) \]
Supervised classification

Estimation methods. Train and test

\[ \hat{p}_M = \frac{1}{N - N_1} \sum_{i=1}^{N-N_1} \delta(c^{N_1+i} = c_M^{(N_1+i)}) \]
Supervised classification

Estimation methods. $k$-fold cross validation

$$\hat{p}_M = \frac{1}{k} \sum_{i=1}^{k} \hat{p}_i$$
Supervised classification

Classifiers. $k$-NEAREST NEIGHBORS

$k = 3$

$k = 6$
Supervised classification

Classifiers. CLASSIFICATION TREE

Equivalent set of rules:

R1: If (Outlook=Sunny) and (Humidity=High) then PLAYTENNIS=No
R2: If (Outlook=Sunny) and (Humidity=Normal) then PLAYTENNIS=Yes
R3: If (Outlook=Overcast) then PLAYTENNIS=Yes
R4: If (Outlook=Rain) and (Wind=Strong) then PLAYTENNIS=No
R5: If (Outlook=Rain) and (Wind=Weak) then PLAYTENNIS=Yes
Classifiers. **Naïve Bayes**

Predictor variables are *conditionally independent* given $C$

$$P(c|x_1, \ldots, x_n) \propto P(C = c) \prod_{i=1}^{n} P(X_i = x_i | C = c)$$

$$\Rightarrow \quad c^* = \arg \max_c P(C = c) \prod_{i=1}^{n} P(X_i = x_i | C = c)$$
Classifiers. **LOGISTIC REGRESSION**

\[
\pi_j = P(C = 1|x_j) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{j1} + \cdots + \beta_n x_{jn})}}
\]

\[
1 - \pi_j = P(C = 0|x_j) = \frac{1}{1 + e^{(\beta_0 + \beta_1 x_{j1} + \cdots + \beta_n x_{jn})}}
\]
Supervised classification

Classifiers. **Neural Network**
Supervised classification

Classifiers. SUPPORT VECTOR MACHINE

Separation may be easier in higher dimensions

complex in low dimensions simple in higher dimensions
Supervised classification

Classifiers. **METACLASSIFIERS**

![Diagram of Supervised Clustering](image)

- Input
- Classifiers: $c_1, c_2, \ldots, c_N$
- Outputs: $o_1, o_2, \ldots, o_N$
- Combine Classifications
- Output $\hat{o}$

**Supervised Classification**

- **Intro**
- **Supervised**
- **Clustering**
- **Conclusions**
Supervised classification

Classifiers. MetaClassifiers. Random Forest
Case study on diagnose aviation turbulence

- Predominant cause of accidents and injuries
- Costing airlines millions of euros per year in compensation: aircraft damage, and delays due to post-event inspections and repairs
- Attempts to avoid turbulent airspace cause: flight delays and en route deviations, increasing air traffic controller workload, disrupt schedules of air crews and passengers and use extra fuel

Supervised classification

Case study on diagnose aviation turbulence

- **Data sets** (March 10 to November 4, 2010)
  - **UAL** from 95 United Airlines, Boeing 757: 5,623,738 instances
  - **DAL** from 80 Delta Air Lines, Boeing 737: 6,595,922 instances
- **Predictor variables**: Numerical weather predictors, Doppler radar, Geostationary satellite images, Lightning detection network, Derived fields and features
- **Class variable**: Eddy dissipation rate (EDR) an aircraft-independent atmospheric turbulence metric
- **Feature subset selection**: wrapper forward/backward selection guided by the AUC
- **Supervised classification methods**: $k$-nearest neighbor ($k = 100$); logistic regression; and random forest (200 trees)
- **Validation**: 10-fold cross-validation repeated 32 times with AUC as performance metric
Supervised classification

Table 2: Forward/backward selection procedure results for RF and logistic regression

<table>
<thead>
<tr>
<th>Rank</th>
<th>Mean occ.</th>
<th>Predictor name</th>
<th>Rank</th>
<th>Mean occ.</th>
<th>Predictor name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>115</td>
<td>Dist. to NSSL echo top &gt; 10 kft</td>
<td>1</td>
<td>135</td>
<td>Model FRNTGTHR1</td>
</tr>
<tr>
<td>2</td>
<td>114</td>
<td>Model FRNTGTHR1</td>
<td>2</td>
<td>134</td>
<td>Diff. Alt. to 80-km max NTDA sev. top</td>
</tr>
<tr>
<td>3</td>
<td>104</td>
<td>Model RITW</td>
<td>3</td>
<td>127</td>
<td>Dist. to NSSL echo top &gt; 10 kft</td>
</tr>
<tr>
<td>4</td>
<td>89</td>
<td>Model ELLROD2</td>
<td>4</td>
<td>126</td>
<td>10-km max of NSSL echo top</td>
</tr>
<tr>
<td>5</td>
<td>88</td>
<td>Diff. Alt. to 80-km max NTDA sev. top</td>
<td>5</td>
<td>121</td>
<td>Model ELLROD2</td>
</tr>
<tr>
<td>6</td>
<td>85</td>
<td>Model MWT2</td>
<td>6</td>
<td>111</td>
<td>Model RITW</td>
</tr>
<tr>
<td>7</td>
<td>79</td>
<td>Model ELLROD1</td>
<td>7</td>
<td>107</td>
<td>Model BROWN2</td>
</tr>
<tr>
<td>8</td>
<td>78</td>
<td>160-km mean of Satellite Ch. 6</td>
<td>8</td>
<td>94</td>
<td>Diff. Alt. to 20-km max NTDA mod. top</td>
</tr>
<tr>
<td>9</td>
<td>69</td>
<td>Model F2DTW</td>
<td>9</td>
<td>92</td>
<td>Model BROWN1</td>
</tr>
<tr>
<td>10</td>
<td>68</td>
<td>Model MWT3</td>
<td>10</td>
<td>89</td>
<td>Model ELLROD1</td>
</tr>
<tr>
<td>11</td>
<td>68</td>
<td>40-km min of Satellite Ch. 6</td>
<td>11</td>
<td>88</td>
<td>Model MWT3</td>
</tr>
<tr>
<td>12</td>
<td>67</td>
<td>Model Atm. Pressure</td>
<td>12</td>
<td>87</td>
<td>Model EDRR1</td>
</tr>
<tr>
<td>13</td>
<td>66</td>
<td>Model BROWN2</td>
<td>13</td>
<td>85</td>
<td>Model DTF3</td>
</tr>
<tr>
<td>14</td>
<td>65</td>
<td>Satellite Ch. 4 minus Model temp.</td>
<td>14</td>
<td>83</td>
<td>20-km no. of good NTDA dBZ points</td>
</tr>
<tr>
<td>15</td>
<td>64</td>
<td>Model DUTTON</td>
<td>15</td>
<td>77</td>
<td>10-km no. of good NTDA dBZ points</td>
</tr>
<tr>
<td>16</td>
<td>63</td>
<td>Satellite Ch. 4 minus Satellite Ch. 3</td>
<td>16</td>
<td>76</td>
<td>10-km mean of NTDA composite EDR</td>
</tr>
<tr>
<td>17</td>
<td>58</td>
<td>Model NGM2</td>
<td>17</td>
<td>74</td>
<td>10-km max of NTDA composite EDR</td>
</tr>
<tr>
<td>18</td>
<td>56</td>
<td>160-km mean of Satellite Ch. 4</td>
<td>18</td>
<td>74</td>
<td>10-km min of Satellite Ch. 3</td>
</tr>
<tr>
<td>19</td>
<td>53</td>
<td>Model RICH</td>
<td>19</td>
<td>73</td>
<td>10-km mean of NSSL echo top</td>
</tr>
<tr>
<td>20</td>
<td>52</td>
<td>Diff. Model pres. to Mod. surf. pres.</td>
<td>20</td>
<td>69</td>
<td>Model IAWINDR1</td>
</tr>
</tbody>
</table>
Supervised classification

Case study on diagnose aviation turbulence

POD\textsubscript{y} = TPR; POD\textsubscript{n} = FPR; Random forest (blue) with AUC=0.924; k-NN (green) with AUC=0.915; logistic regression (green) with AUC=0.915; Graphical turbulence guidance (magenta) with AUC=0.816; storm distance (magenta) with AUC=0.743
### Multi-label classification

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2</td>
<td>1.4</td>
<td>4.7</td>
<td>7.5</td>
<td>3.7</td>
<td>1</td>
</tr>
<tr>
<td>2.8</td>
<td>6.3</td>
<td>1.6</td>
<td>4.7</td>
<td>2.7</td>
<td>0</td>
</tr>
<tr>
<td>7.7</td>
<td>6.2</td>
<td>4.1</td>
<td>3.3</td>
<td>7.7</td>
<td>1</td>
</tr>
<tr>
<td>9.2</td>
<td>0.4</td>
<td>2.8</td>
<td>0.5</td>
<td>3.9</td>
<td>0</td>
</tr>
<tr>
<td>5.5</td>
<td>5.3</td>
<td>4.9</td>
<td>0.6</td>
<td>6.6</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_3$</th>
<th>$C_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2</td>
<td>1.4</td>
<td>4.7</td>
<td>7.5</td>
<td>3.7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2.8</td>
<td>6.3</td>
<td>1.6</td>
<td>4.7</td>
<td>2.7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7.7</td>
<td>6.2</td>
<td>4.1</td>
<td>3.3</td>
<td>7.7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9.2</td>
<td>0.4</td>
<td>2.8</td>
<td>0.5</td>
<td>3.9</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5.5</td>
<td>5.3</td>
<td>4.9</td>
<td>0.6</td>
<td>6.6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Case study on the ASRS database

- The Aviation Safety Reporting System (ASRS) database spans 30 years and contains over 700,000 aviation safety reports in free text form.
- Primary concern is system health and safety: detection, diagnosis, prediction, mitigation, and prevention of ongoing and future system problems.
- ARSR reports are publicly available and are written by pilots, flight controllers, technicians, flight attendants, and others including passengers.
- Predictor variables, $X_1$ to $X_{25,729}$: 25,729 terms that occur in at least one document.
- Class variables to be predicted, $C_1$ to $C_{60}$: 60 problem types that appear during flights.
- A subset of 28,596 documents containing 22 classes was used.

Multi-label classification

Case study on the ASRS database
Multi-label classification

Case study on the ASRS database

Correlations between pairs of class variables
Multi-label classification

Case study on the ASRS database

AUC performances over the 60 univariate classification problems
A general dataset

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>...</th>
<th>$X_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x^{(1)}$</td>
<td>$x_1^{(1)}$</td>
<td>...</td>
<td>$x_n^{(1)}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>$x^{(N)}$</td>
<td>$x_1^{(N)}$</td>
<td>...</td>
<td>$x_n^{(N)}$</td>
</tr>
</tbody>
</table>

- **Grouping objects** in clusters
- **High similarity** within each cluster
- **High dissimilarity** between the different clusters
- **Main methods:**
  - Hierarchical clustering
  - Partitional clustering
  - Probabilistic clustering
Hierarchical Clustering

Hierarchical clustering is a method of cluster analysis which aims to build a hierarchy of clusters. It can be divided into two types: agglomerative (bottom-up) and divisive (top-down).

In this diagram, the hierarchical clustering process is illustrated. The points are grouped into clusters, and the dendrogram on the right shows the merging of clusters at different levels of similarity.

The dendrogram on the right represents the hierarchical clustering process. Each vertical line corresponds to a merge, and the height indicates the distance at which clusters were merged.
PARTITIONAL CLUSTERING: $k$-means
PROBABILISTIC CLUSTERING: finite mixture models with EM

(a) $L = 1$

(b) $L = 2$

(c) $L = 5$

(d) $L = 20$

P. Larrañaga  Machine Learning in Aviation
Clustering

Case study on detecting anomalous landing

- A key question on the aviation safety domain
- For a specific aircraft make and model at a specific airport
- Variables: sequence of switches that a pilot flips during the course of the landing phase of the flight
- Data set: N = 2,200 flights, and n = 1,500 possible switches
- Cluster algorithm: $k$-medoids
- Anomalous landing: a sequence that is far away from the cluster centroid

Outline

1. Introduction
2. Supervised Classification
3. Clustering
4. Conclusions
Conclusions

Machine learning in aviation

- Aviation industry generates large scale data
- Transform these data sets into knowledge
- Machine learning methods:
  - Supervised classification
  - Clustering
- Advances in the safety, security, and efficiency of civil aviation
References on Supervised Classification

- Kuncheva LI (2004). *Combining Pattern Classifiers*. John Wiley and Sons
References on Clustering

MACHINE LEARNING IN AVIATION

PEDRO LARRAÑAGA

Computational Intelligence Group
Artificial Intelligence Department
Technical University of Madrid, Spain

Toulouse, July 11, 2013