Applied Machine Learning

Course number: W207

Prof. Alexander I. Iliev, Ph.D.

Applied Machine Learning

Lecture 5 …

- *The ARFF format*
- *Entropy, Information gain (cont.)*
- *Random Forests (RF)*
- *Ensemble method comparison (code): DT, RF, AdaBoost*
- *Regression*
    - *Linear*
    - *Logistic*

Applied Machine Learning

Lecture 5 …

- *The ARFF format*
- *Entropy, Information gain (cont.)*
- *Random Forests (RF)*
- *Ensemble method comparison (code): DT, RF, AdaBoost*
- *Regression*
  - *Linear*
  - *Logistic*

Preparing the input

- Preparing input for a data mining investigation usually consumes the bulk of the effort invested in the entire data mining process.

- Bitter experience shows that real data is often disappointingly low in quality, and careful checking—a process that has become known as data cleaning —pays off many times over.

- When beginning work on a data mining problem, it is first necessary to bring all the data together into a set of instances.

Preparing the input

- *ARFF* (Attribute-Relation File *Format*) file is an ASCII text file that describes a list of instances sharing a set of attributes

- *ARFF* files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software.

The ARFF data format

```
%
% ARFF file for weather data with some numeric features
%
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {true, false}
@attribute play? {yes, no}

@data
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
...
```

The ARFF data format

```
%
% ARFF file for weather data with some numeric features
%
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {true, false}
@attribute play? {yes, no}

@data
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
...
```

The ARFF data format

- Lines that begin with a % are comments

- @RELATION, @ATTRIBUTE and @DATA declarations are case insensitive

- @relation <relation-name> - <relation-name> is a string. The string must be quoted if the name includes spaces

- Attribute declarations take the form of an ordered sequence of **@attribute** statements

The ARFF data format

- Each attribute in the data set has its own **@attribute**

- Each **@attribute** statement uniquely defines the name of that attribute and it's data type

- The order the attributes are declared indicates the column position in the data section of the file:

Example: if an attribute is declared on the 2nd line, that attributes values must be found in the 2nd comma delimited column of the instances

The ARFF data format

- The format for the **@attribute** statement is:

@attribute <attribute-name> <datatype>

where if spaces are included in the attribute name then the entire name must be quoted

```
% ARFF file for weather data with some numeric features
%
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute "outside temperature" numeric
...
@data
sunny, 85, 85, false, no
sunny, 80, 90, true, no
```

The ARFF data format

- The <*datatype*> can be any of the four types:
  - numeric
  - <nominal-specification>
  - string
  - date [<date-format>]

  where, **numeric**, **string** and **date** are case insensitive

- Numeric attributes can be real or integer numbers

- Nominal values are defined by listing the possible values:

  { <nominal-name1>, <nominal-name2>, <nominal-name3>, ... }

Additional attribute types

- ARFF data format also supports *string* attributes:

  ```
  @attribute description string
  ```

  - Similar to nominal attributes but list of values is not pre-specified

- Additionally, it supports *date* attributes:

  ```
  @attribute today date
  ```

  - Uses the ISO-8601 combined date and time format:

    @ATTRIBUTE timestamp DATE "yyyy-MM-dd HH:mm:ss”

    @DATA "2018-02-15 10:15:18”

- Missing values are represented by a single question mark, as:

  @data 4.4,?,1.5,?,Iris-setosa

Relational attributes

- Relational attributes allow multi-instance problems to be represented in ARFF format
  - Each value of a relational attribute is a *separate* bag of instances, but each bag has the *same attributes*

```
@attribute bag relational
       @attribute outlook { sunny, overcast, rainy }
       @attribute temperature numeric
       @attribute humidity numeric
       @attribute windy { true, false }
@end bag
```

  - Nested attribute block gives the structure of the referenced instances
  - The @end bag indicates the end of the nested attribute block

Multi-instance ARFF

```
%
% Multiple instance ARFF file for the weather data
%
@relation weather

@attribute bag_ID { 1, 2, 3, 4, 5, 6, 7 }
@attribute bag relational
       @attribute outlook {sunny, overcast, rainy}
       @attribute temperature numeric
       @attribute humidity numeric
       @attribute windy {true, false}
       @attribute play? {yes, no}
@end bag

@data
1, “sunny, 85, 85, false\nsunny, 80, 90, true”, no
2, “overcast, 83, 86, false\nrainy, 70, 96, false”, yes
...
```

Sparse data

- In some applications most attribute values are zero and storage requirements can be reduced
  - E.g.: word counts in a text categorization problem
- ARFF supports sparse data storage

```
0, 26, 0,  0, 0 ,0, 63, 0, 0, 0, “class A”
0,  0, 0, 42, 0, 0,  0, 0, 0, 0, “class B”
```

```
{1 26, 6 63, 10 “class A”}
{3 42, 10 “class B”}
```

- This also works for nominal attributes (where the first value of the attribute corresponds to “zero”)
- Some learning algorithms work very efficiently with sparse data

Applied Machine Learning

Lecture 5 …

- *The ARFF format*
- *Entropy, Information gain (cont.)*
- *Random Forests (RF)*
- *Ensemble method comparison (code): DT, RF, AdaBoost*
- *Regression*
  - *Linear*
  - *Logistic*

Applied Machine Learning

Lecture 5 …

- *The ARFF format*
- *Entropy, Information gain (cont.)*
- *Random Forests (RF)*
- *Ensemble method comparison (code): DT, RF, AdaBoost*
- *Regression*
    - *Linear*
    - *Logistic*

Applied Machine Learning

Lecture 5 …

- *The ARFF format*
- *Entropy, Information gain (cont.)*
- *Random Forests (RF)*
- *Ensemble method comparison (code): DT, RF, AdaBoost*
- *Regression*
    - *Linear*
    - *Logistic*

Applied Machine Learning

Lecture 5 …

- *The ARFF format*
- *Entropy, Information gain (cont.)*
- *Random Forests (RF)*
- *Ensemble method comparison (code): DT, RF, AdaBoost*
- *Regression*
    - *Linear*
    - *Logistic*