


Applied Machine Learning

Input: Concepts, Instances, Attributes

Computer Science, Fall 2022

Instructor: Xuhong Zhang

Defining the Learning Task

- Improve on task T , with respect to performance metric P , based on experience E 
 - T : Categorize email messages as spam or legitimate
 - P : Percentage of email messages correctly classified
 - E : Database of emails, some with human-given labels
- T : Recognizing hand-written words
- P : Percentage of words correctly classified
- E : Database of human-labeled images of handwritten words

Concept Description

- The input:
 - Concepts, instances, and attributes.
 - *Ideally*, intelligible in that it can be understood, discussed, and disputed, and *operational* in that it can be applied to actual examples.

What is a Concept

- Classification learning

- The learning schema is presented with a set of classified examples from which it is expected to learn a way of classifying unseen examples.

- Association learning

- Any association among features is sought, not just ones that predict a particular class value.

Concept

- Clustering
 - Groups of examples that belong together are sought.
- Numeric Prediction
 - The outcome to be predicted is not a discrete class but a numeric quantity.

Classification: The weather data

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Classification: The Iris data

Sepal Length	Sepal Width	Petal Length	Petal Width	Type
5.1	3.5	1.4	0.2	Iris setosa
4.9	3.0	1.4	0.2	Iris setosa
4.7	3.2	1.3	0.2	Iris setosa
...				
7.0	3.2	4.7	1.4	Iris versicolor
6.4	3.2	4.5	1.5	Iris versicolor
6.9	3.1	4.9	1.5	Iris versicolor
5.5	2.3	4.0	1.3	Iris versicolor
...				
6.3	3.3	6.0	2.5	Iris virginica
5.8	2.7	5.1	1.9	Iris virginica
7.1	3.0	5.9	2.1	Iris virginica
6.3	2.9	5.6	1.8	Iris virginica

Classification

- Assumption: each example belongs to one, and only one, class.
 - There exists classification scenarios in which individual examples may belong to multiple classes.
 - These are called “multi-labeled instances”.
- How to solve this scenario?
 - Treat then as several different classification problems, one for each possible class
 - Use a regression (numeric) values to represent the classification probability.

Classification

- Is classification a supervised learning process or unsupervised learning process?

Association

- If there is *no specific* class, and the problem is to discover any structure in the data that is “interesting”.
 - Predict any attribute, not just the class and can predict more than one attribute's value at a time
 - Far more association rules
 - Association are often limited to those that apply to a certain minimum number of examples
 - Often limited to those that apply to a certain minimum number of examples (e.g. 80% of the dataset, and have greater than a threshold on accuracy, say 95% accurate).
 - Even then, there are usually lots of them, and they have to be examined manually to determine if they are meaningful.
 - Association rules usually involve only nonnumeric attributes: you won't normally look for association rules in the iris dataset.

Clustering

- When there is no specific class, clustering is used to group items that seem to fall naturally together.

Sepal Length	Sepal Width	Petal Length	Petal Width	Type
5.1	3.5	1.4	0.2	Iris setosa
4.9	3.0	1.4	0.2	Iris setosa
...				
7.0	3.2	4.7	1.4	Iris versicolor
6.4	3.2	4.5	1.5	Iris versicolor
...				
6.3	3.3	6.0	2.5	Iris virginica
5.8	2.7	5.1	1.9	Iris virginica

Clustering

- Is clustering supervised or unsupervised?
- How you can measure your clustering result?

Numeric Prediction

- Variant of classification learning where “class” is numeric
 - Also called “regression”
- Learning is supervised
- Measure success on test data

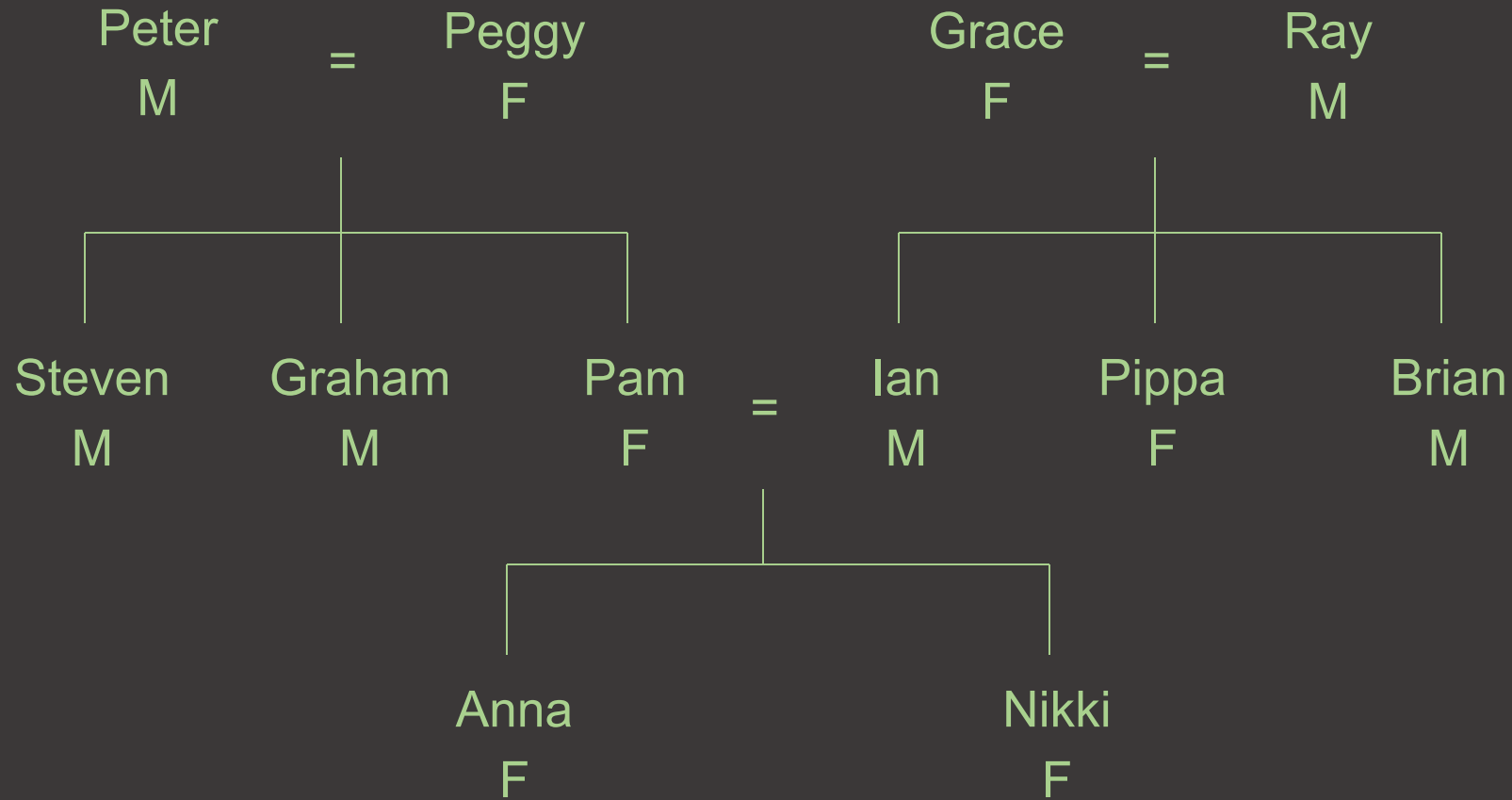
Numeric Prediction

Outlook	Temperature	Humidity	Windy	Play-time
Sunny	Hot	High	False	5
Sunny	Hot	High	True	0
Overcast	Hot	High	False	55
Rainy	Mild	High	False	40
Rainy	Cool	Normal	False	23

What's in an example?

- Instance: specific type of example
 - Thing to be classified, associated, or clustered
 - Individual, independent example of target concept
 - Characterized by a predetermined set of attributes
- Input to learning schema: set of instances/dataset
 - Represented as a single relation/flat file

Relations: A family tree



Relations

Family tree represented as a table

Name	Gender	Parent1	parent2
Peter	Male	?	?
Peggy	Female	?	?
Steven	Male	Peter	Peggy
Graham	Male	Peter	Peggy
Pam	Female	Peter	Peggy
Ian	Male	Grace	Ray
Pippa	Female	Grace	Ray
Brian	Male	Grace	Ray
Anna	Female	Pam	Ian
Nikki	Female	Pam	Ian

The “sister-of” relation

First person	Second person	Sister of?
Peter	Peggy	No
Peter	Steven	No
...
Steven	Peter	No
Steven	Graham	No
Steven	Pam	Yes
...
Ian	Pippa	Yes
...
Anna	Nikki	Yes
...
Nikki	Anna	yes

First person	Second person	Sister of?
Steven	Pam	Yes
Graham	Pam	Yes
Ian	Pippa	Yes
Brian	Pippa	Yes
Anna	Nikki	Yes
Nikki	Anna	Yes
<i>All the rest</i>		No

Closed-world assumption



A full representation in one table

First person				Second person				Sister of?
Name	Gender	Parent1	Parent2	Name	Gender	Parent1	Parent2	
Steven	Male	Peter	Peggy	Pam	Female	Peter	Peggy	Yes
Graham	Male	Peter	Peggy	Pam	Female	Peter	Peggy	Yes
Ian	Male	Grace	Ray	Pippa	Female	Grace	Ray	Yes
Brian	Male	Grace	Ray	Pippa	Female	Grace	Ray	Yes
Anna	Female	Pam	Ian	Nikki	Female	Pam	Ian	Yes
Pippa	Female	Pam	Ian	Anna	Female	Pam	Ian	Yes
All the rest								No

Comparing to the table in the previous page, what's the difference?

**If second person's gender = female
and first person's parent = second person's parent
then sister-of = yes**

Relations

- So far, we have transformed the original relational problem into the form of instances, each of which is an individual, independent example of the concept that is to be learned.
- Cautious: the instances are not really independent—there are plenty of relationships among different rows of the table.
- But they are independent as far as the concept of sisterhood is concerned.

Generating a flat file

- This is an example of you can take a relationship between different nodes of a tree and recast it into a set of independent instances.

Denormalization

- Several relations are joined together to make one
- Possible with any finite set of finite relations
- Problematic: relations among more people would require a larger table. Relationships in which the maximum number of people is not specified pose a more serious problem.
 - Nuclear-family
- Problematic: may produce spurious regularities that reflect the structure of the database
 - “supplier” predicts “supplier address”

The “ancestor-of” relation

First person				Second person				Ancestor of?
Name	Gender	Parent1	Parent2	Name	Gender	Parent1	Parent2	
Peter	Male	?	?	Steven	Male	Peter	Peggy	Yes
Peter	Male	?	?	Pam	Female	Peter	Peggy	Yes
Peter	Male	?	?	Anna	Female	Pam	Ian	Yes
Peter	Male	?	?	Nikki	Female	Pam	Ian	Yes
Pam	Female	Peter	Peggy	Nikki	Female	Pam	Ian	Yes
Grace	Female	?	?	Ian	Male	Grace	Ray	Yes
Grace	Female	?	?	Nikki	Female	Pam	Ian	Yes
Other positive examples here								Yes
All the rest								No

- Many abstract computational problems involve relations that are not finite.

How you solve problems like this?

Recursion

- Infinite relations require recursion

```
If person1 is a parent of person2  
    then person1 is an ancestor of person2
```

```
If person1 is a parent of person2  
    and person2 is an ancestor of person3  
    then person1 is an ancestor of person3
```

- Appropriate techniques are known as “inductive logic programming” (ILP) methods
 - Example ILP method: Quinlan’s FOIL rule learner
 - Problems: (a) noise and (b) computational complexity

Multi-instance concepts

- Each individual example comprises a bag (aka *multi-set*) of instances
 - All instances are described by the same attributes
 - One or more instances within an example may be responsible for the example's classification
- Goal of learning is still to produce a concept description
- Important real-world applications
 - Prominent examples are drug activity prediction and image classification
 - A drug can be viewed as bag of different geometric arrangements of the drug molecule
 - An image can be represented as a bag of image components

What's in an attribute?

- Each instance is described by a fixed predefined set of features, its “attributes”
- But: number of attributes may vary in practice
 - Possible solution: “irrelevant value” flag
- Related problem: existence of an attribute may depend on value of another one
- Possible attribute types (“levels of measurement”):
 - Nominal, ordinal, interval and ratio

Nominal levels of measurements

- Values are distinct symbols
 - Values themselves serve only as labels or names
 - *Nominal* comes from the Latin word for name
- Example: attribute “outlook” from weather data
 - Values: “sunny”, “overcast”, and “rainy”
- No relation is implied among nominal values (no ordering or distance measure)

Ordinal levels of measurements

- Impose order on values
- But: no distance between values defined
- Example: attribute “temperature” in weather data
 - Values: “hot” > “mild” > “cool”
 - Note: addition and subtraction don’t make sense
- Example rule: temperature < hot \Rightarrow play = yes
- Distinction between nominal and ordinal not always clear (e.g., attribute “outlook”)

Interval quantities

- Interval quantities are not only ordered but measured in fixed and equal units
 - Example 1: attribute “temperature” expressed in degrees Fahrenheit
 - Example 2: attribute “year”
- Difference of two values makes sense
- Sum or product doesn't make sense

Attribute types used in practice

- Many data mining schemes accommodate just two levels of measurement: nominal and ordinal
- Others deal exclusively with ratio quantities
- Nominal attributes are also called “categorical”, “enumerated”, or “discrete”
 - But: “enumerated” and “discrete” imply order
- Special case: dichotomy (“boolean” attribute)
- Ordinal attributes are sometimes coded as “numeric” or “continuous”
 - But: “continuous” implies mathematical continuity

Metadata

- Information about the data that encodes background knowledge
- In theory this information can be used to restrict the search space of the learning algorithm
- Examples:
 - Dimensional considerations
(i.e., expressions must be dimensionally correct)
 - Circular orderings
(e.g., degrees in compass)
 - Partial orderings
(e.g., generalization/specialization relations)

Preparing the input

- Denormalization is not the only issue when data is prepared for learning
- Problem: different data sources (e.g., sales department, customer billing department, ...)
 - Differences: styles of record keeping, coding conventions, time periods, data aggregation, primary keys, types of errors
 - Data must be assembled, integrated, cleaned up
 - “Data warehouse”: consistent point of access
- External data may be required (“overlay data”)
- Critical: type and level of data aggregation

Missing values

- Missing values are frequently indicated by out-of-range entries for an attribute
 - There are different types of missing values: unknown, unrecorded, irrelevant
 - Reasons:
 - malfunctioning equipment
 - changes in experimental design
 - collation of different datasets
 - measurement not possible
- Missing value may have significance in itself (e.g., missing test in a medical examination)
 - “missing” may need to be coded as an additional, separate attribute value

Question: How you want to deal with missing values?

Inaccurate values

- Result: errors and omissions that affect the accuracy of data mining
- These errors may not affect the original purpose of the data (e.g., age of customer)
- Typographical errors in nominal attributes \Rightarrow values need to be checked for consistency
- Typographical and measurement errors in numeric attributes \Rightarrow outliers need to be identified
- Errors may be deliberate (e.g., wrong zip codes)
- Other problems: duplicates, stale data

How can you deal with inaccurate values?

Unbalanced data

- Unbalanced data is a well-known problem in classification problems
 - One class is often far more prevalent than the rest
 - Example: detecting a rare disease
- Main problem: simply predicting the majority class yields high accuracy but is not useful
 - Predicting that no patient has the rare disease gives high classification accuracy
- Unbalanced data requires techniques that can deal with unequal misclassification costs
 - Misclassifying an afflicted patient may be much more costly than misclassifying a healthy one

How can you deal with unbalanced values?

Getting to know your data

- Simple visualization tools are very useful
 - Nominal attributes: histograms (Is the distribution consistent with background knowledge?)
 - Numeric attributes: graphs (Any obvious outliers?)
- 2-D and 3-D plots show dependencies
- May need to consult domain experts
- Too much data to inspect manually? Take a sample!

Plot your data !!!!