Classification & Clustering

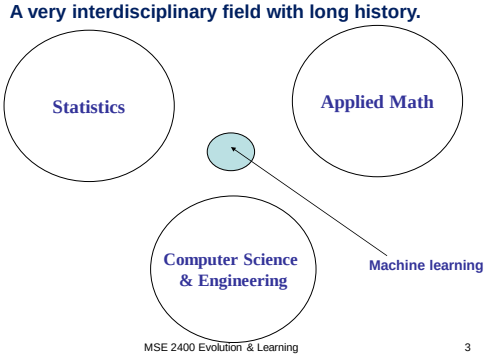
MSE 2400 EaLiCaRA
Dr. Tom Way

Recall: Machine Learning

- a branch of artificial intelligence, is about the construction and study of systems that can learn from data.
- The ability of a computer to improve its own performance through the use of software that employs artificial intelligence techniques to mimic the ways by which humans seem to learn, such as repetition and experience.

Machine Learning

A very interdisciplinary field with long history.



Classification & Clustering

The general idea…

- Machine learning techniques to group inputs into (hopefully) distinct categories…
- …and then to use those categories to identify group membership of new input in the future

An example application

- An emergency room in a hospital measures 17 variables (e.g., blood pressure, age, etc) of newly admitted patients.
- A decision is needed: whether to put a new patient in an intensive-care unit.
- Due to the high cost of ICU, those patients who may survive less than a month are given higher priority.
- Problem: to predict high-risk patients and discriminate them from low-risk patients.

Another application

- A credit card company receives thousands of applications for new cards. Each application contains information about an applicant,
  - age
  - Marital status
  - annual salary
  - outstanding debts
  - credit rating
  - etc.
- Problem: to decide whether an application should approved, or to classify applications into two categories, approved and not approved.

Definition of ML Classifier

❑ Definition of Machine Learning from dictionary.com

"The ability of a machine to improve its performance based on previous results."

❑ So, machine learning document classification is "the ability of a machine to improve its document classification performance based on previous results of document classification".

You as an ML Classifier

- Topic 1 words:

  baseball, owners, sports, selig, ball, bill, indians, isringhausen, mets, minors, players, specter, stadium, power, send, new, bud, comes, compassion, game, headaches, lite, nfl, powerful, strawberry, urges, home, ambassadors, building, calendar, commish, costs, day, dolan, drive, hits, league, little, match, payments, pitch, play, player, red, stadiums, umpire, wife, youth, field, leads

- Topic 2 words:

  merger, business, bank, buy, announces, new, acquisition, finance, companies, com, company, disclosure, emm, news, us, acquire, chemical, inc, results, shares, takeover, corporation, european, financial, investment, market, quarter, two, acquires, bancorp, bids, communications, first, mln, purchase, record, stake, west, sale, bid, bn, brief, briefs, capital, control, europe, inculab

Use the previous slide's topics & related words to classify the following titles

1. CYBEX-Trotter merger creates fitness equipment powerhouse
2. WSU RECRUIT CHOOSES BASEBALL INSTEAD OF FOOTBALL
3. FCC chief says merger may help pre-empt Internet regulation
4. Vision of baseball stadium growing
5. Regency Realty Corporation Completes Acquisition Of Branch properties
6. Red Sox to punish All-Star scalpers
7. Canadian high-tech firm poised to make $415-million acquisition
8. Futures-selling hits the Footsie for six
9. A'S NOTEBOOK; Another Young Arm Called Up
10. All-American SportPark Reaches Agreement for Release of Corporate Guarantees

Titles & Their Classifications

1. (2) CYBEX-Trotter merger creates fitness equipment powerhouse
2. (1) WSU RECRUIT CHOOSES BASEBALL INSTEAD OF FOOTBALL
3. (2) FCC chief says merger may help pre-empt Internet regulation
4. (1) Vision of baseball stadium growing
5. (2) Regency Realty Corporation Completes Acquisition Of Branch properties
6. (1) Red Sox to punish All-Star scalpers
7. (2) Canadian high-tech firm poised to make $415-million acquisition
8. (2) Futures-selling hits the Footsie for six
9. (1) A'S NOTEBOOK; Another Young Arm Called Up
10. (1) All-American SportPark Reaches Agreement for Release of Corporate Guarantees

A little math

❑ Canadian high-tech firm poised to make $415-million acquisition

1. Estimate the probablity of a word in a topic by dividing the number of times the word appeared in the topic's training set by the total number of word occurrences in the topic's training set.
2. For each topic,T, sum the probability of finding each word of the title in a title that is classified as T.
3. The title is classified as the topic with the largest sum.

```
Title's evidence of being in Topic 2=0.01152
Title's evidence of being in Topic 1=0.00932
    Canadian 1 0: high 0 0: tech 2 0: firm 1 0
    poised 0 0: make 0 0: million 4 4:
    acquisition 10 0
# of words in Topic2 = 1563
# of words in Topic1 = 429
```


Machine learning and our focus

- Like human learning from past experiences.
- A computer does not have "experiences".
- A computer system learns from data, which represent some "past experiences" of an application domain.
- Our focus: learn a target function that can be used to predict the values of a discrete class attribute, e.g., approve or not-approved, and high-risk or low risk.
- The task is commonly called: Supervised learning, classification, or inductive learning.

The data and the goal

- Data: A set of data records (also called examples, instances or cases) described by
  - *k* attributes: A_1, A_2, … A_k.
  - a class: Each example is labelled with a pre-defined class.
- Goal: To learn a classification model from the data that can be used to predict the classes of new (future, or test) cases/instances.

An example: data (loan application)

Approved or not

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No

An example: the learning task

- Learn a classification model from the data
- Use the model to classify future loan applications into
  - Yes (approved) and
  - No (not approved)
- What is the class for following case/instance?

Age	Has_Job	Own_house	Credit-Rating	Class
young	false	false	good	?

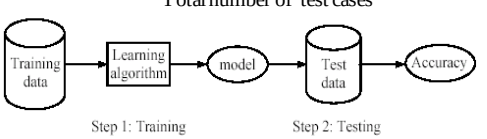
Supervised vs. unsupervised Learning

- Supervised learning (classification): same thing as learning from examples.
  - Supervision: The data (observations, measurements, etc.) are labeled with pre-defined classes. It is like a "teacher" gives the classes (supervision).
  - Test data are classified into these classes too.
- Unsupervised learning (clustering)
  - Class labels of the data are unknown
  - Given a set of data, the task is to establish the existence of classes or clusters in the data

Supervised learning process: two steps

Learning (training): Learn a model using the training data
Testing: Test the model using unseen test data to assess the model accuracy

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}},$$



What do we mean by learning?

- Given
  - a data set *D*,
  - a task *T*, and
  - a performance measure *M*,

  a computer system is said to **learn** from *D* to perform the task *T* if after learning the system's performance on *T* improves as measured by *M*.
- In other words, the learned model helps the system to perform *T* better as compared to no learning.

An example

- Data: Loan application data
- Task: Predict whether a loan should be approved or not.
- Performance measure: accuracy.

No learning: classify all future applications (test data) to the majority class (i.e., Yes):

Accuracy = 9/15 = 60%.

- We can do better than 60% with learning.

Fundamental assumption of learning

Assumption: The distribution of training examples is identical to the distribution of test examples (including future unseen examples).

- In practice, this assumption is often violated to certain degree.
- Strong violations will clearly result in poor classification accuracy.
- To achieve good accuracy on the test data, training examples must be sufficiently representative of the test data.

Two categories of machine learning

1. **Classification (supervised machine learning):**
   - With the class label known, learn the features of the classes to predict a future observation.
   - The learning performance can be evaluated by the prediction error rate.

2. **Clustering (unsupervised machine learning)**
   - Without knowing the class label, cluster the data according to their similarity and learn the features.
   - Normally the performance is difficult to evaluate and depends on the content of the problem.

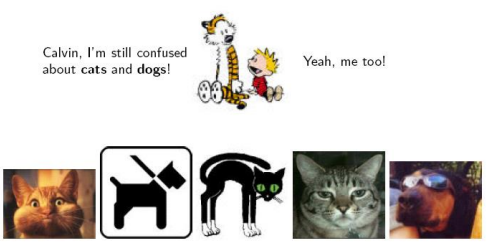
Machine learning classification

Supervised Learning

Florian Markowetz, *Classification by SVM, Practical DNA Microarray Analysis* 2003

Machine learning clustering

Unsupervised Learning



Florian Markowetz, *Classification by SVM, Practical DNA Microarray Analysis* 2003

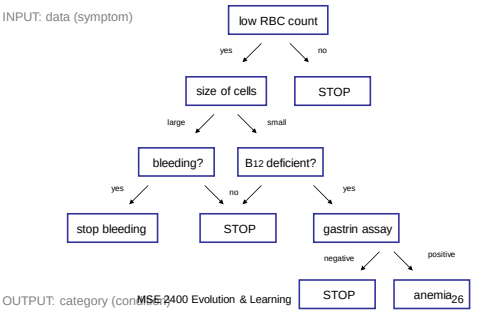
Examples of ML Classifiers

- Decision Trees
- Bayesian Networks
- Neural Networks
- Support Vector Machines
- K-nearest Neighbor
- Instance-based Classifiers

Examples of ML algorithms

- **Decision Trees** – Sort instances according to feature values, build into a hierarchy of tests based on entropy or other reasonable metric. Root node is the one that best divides the data. Branches represent the values that a node can assume (the "decision"). To build the tree, apply recursion to determine next best division...

Decision Trees, an example



Decision Trees: Assessment

- Advantages:
  - Classification of data based on limiting features is *intuitive*
  - Handles discrete/categorical features best
- Limitations:
  - Danger of "overfitting" the data
  - Not the best choice for accuracy

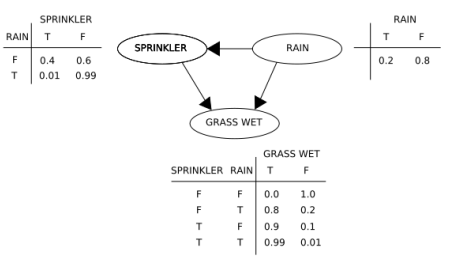
Examples of ML algorithms

- **Naïve Bayes** – This method computes the probability that a document is about a particular topic, T, using a) the words of the document to be classified and b) the estimated probability of each of these words as they appeared in the set of training documents for the topic, T – like the example previously given.

Bayesian Networks

- Graphical algorithm that encodes the joint probability distribution of a data set
- Captures probabilistic relationships *between* variables
- Based on probability that instances (data) belong in each category

Bayesian Networks, an example



Bayesian Networks: Assessment

- Advantages:
  - Takes into account *prior* information regarding relationships among features
  - Probabilities can be updated based on outcomes
  - Fast!…with respect to learning classification
  - Can handle incomplete sets of data
  - Avoids "overfitting" of data
- Limitations:
  - Not suitable for data sets with many features
  - Not the best choice for accuracy

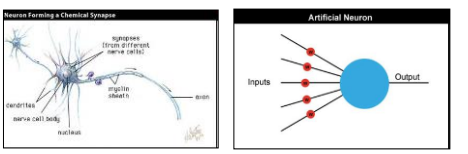
Examples of ML algorithms

- **Neural networks** – During training, a neural network looks at the patterns of features (e.g. words, phrases, or N-grams) that appear in a document of the training set and attempts to produce classifications for the document. If its attempt doesn't match the set of desired classifications, it adjusts the weights of the connections between neurons. It repeats this process until the attempted classifications match the desired classifications.

Neural Networks

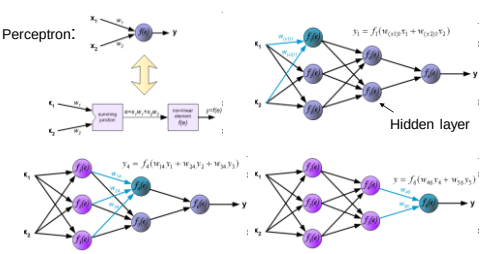
- Used for:
  - Classification
  - Noise reduction
  - Prediction
- Great because:
  - Able to learn
  - Able to generalize
- Kiran
  - Plaut's (1996) semantic neural network that could be lesioned and retrained – useful for predicting treatment outcomes
- Mikkulainen
  - Evolving neural network that could adapt to the gaming environment – useful learning application

Neural Networks: Biological Basis



Feed-forward Neural Network

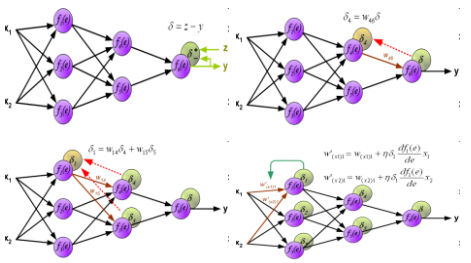
Perceptron:



Neural Networks: Training

- Presenting the network with sample data and modifying the weights to better approximate the desired function.
- Supervised Learning
  - Supply network with inputs and desired outputs
  - Initially, the weights are randomly set
  - Weights modified to reduce difference between actual and desired outputs
  - Backpropagation

Backpropagation



Examples of ML algorithms

- **Support Vector Machines (SVM)** – do supervised learning by analyzing data and recognizing patterns, used for classification tasks. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output.

Support Vector Machines

- Support vector machines were invented by V. Vapnik and his co-workers in 1970s in Russia and became known to the West in 1992.
- SVMs are linear classifiers that find a hyperplane to separate two class of data, positive and negative.
- Kernel functions are used for nonlinear separation.
- SVM not only has a rigorous theoretical foundation, but also performs classification more accurately than most other methods in applications, especially for high dimensional data.
- It is perhaps the best classifier for text classification.

Basic SVM concepts

- Let the set of training examples D be

  $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_r, y_r)\}$,

  where $\mathbf{x}_i = (x_1, x_2, \ldots, x_n)$ is an **input vector** in a real-valued space $X \subseteq R^n$ and y_i is its **class label** (output value), $y_i \in \{1, -1\}$.

  1: positive class and -1: negative class.
- SVM finds a linear function of the form (**w**: weight vector)

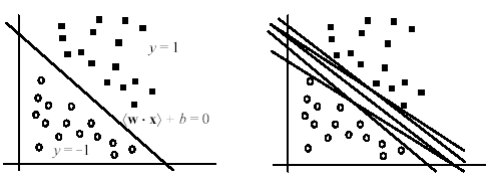
  $$f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b$$

  $$y_i = \begin{cases} 1 & if \ \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \geq 0 \\ -1 & if \ \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b < 0 \end{cases}$$

The SVM hyperplane

- The hyperplane that separates positive and negative training data is

  $\langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 0$
- It is also called the decision boundary (surface).
- So many possible hyperplanes, which one to choose?



Examples of ML algorithms

- **K-nearest neighbor** – Creates clusters of data in an interactive fashion… does not use training data… it works on the actual data.

k-Nearest Neighbor Classification (kNN)

- Unlike all the previous learning methods, kNN does not build model from the training data.
- To classify a test instance *d*, define *k*-neighborhood *P* as *k* nearest neighbors of *d*
- Count number *n* of training instances in *P* that belong to class c_j
- Estimate $Pr(c_j|d)$ as n/k
- No training is needed. Classification time is linear in training set size for each test case.

kNNAlgorithm

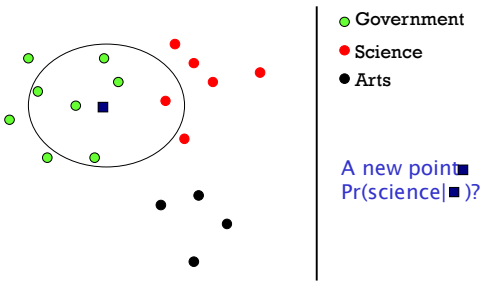
Algorithm kNN(D, d, k)

1 Compute the distance between d and every example in D;
2 Choose the k examples in D that are nearest to d, denote the set by P ($\subseteq D$);
3 Assign d the class that is the most frequent class in P (or the majority class);

k is usually chosen empirically via a validation set or cross-validation by trying a range of *k* values.

Distance function is crucial, but depends on applications.

Example: k=6 (6NN)



Discussions

- kNN can deal with complex and arbitrary decision boundaries.
- Despite its simplicity, researchers have shown that the classification accuracy of kNN can be quite strong and in many cases as accurate as those elaborated methods.
- kNN is slow at the classification time
- kNN does not produce an understandable model

Examples of ML algorithms

- **Instance based** – Saves documents of the training set and compares new documents to be classified with the saved documents. The document to be classified gets tagged with the highest scoring classifications. One way to do this is to implement a search engine using the documents of the training set as the document collection. A document to be classified becomes a query/search. A classification, C, is picked if a large number of its training set documents are at the top of the returned answer set.

Building our own classifiers

- In an upcoming lab, we will build our own classifiers using the Python programming language

How well do ML classifiers work?

- A good system will have an accuracy of above 80%.
- Strong evidence of how good these systems are is the number of companies in the market place with machine learning document classification systems.

Advantages

- Advantage over classification by humans: Once the system is trained, classification is done automatically with no or little human intervention – saving human resources.
- Advantage over classification by humans: Consistent classification.
- Advantage over rule based classification: Human resources are not needed to make rules.

Disadvantages

- Disadvantage: Not always obvious why it classified a document in a certain way and not obvious how to keep it from doing the same type of classification in the future (i.e. don't know how to modify it.)
- Disadvantage: Human resources must be used to manually classify documents for the training set. Furthermore, the number and type of document that should be in the training set isn't straightforward.

Uses of ML classifiers

- Automatically classify things.
- Suggest classifications that a human can pick from.
- Classify paragraphs or even sentences in a document.
- Find important information in a document. For example, rules of law in a case law document, or the facts of the case.

Challenges

- Labeling the documents of the training set.
- What is the best way to pick documents for the training set so the machine-learning algorithm produces a classifier with high accuracy?
- Which machine-learning algorithm works best on your classification problem?

The Future of Supervised Learning (1)

- Generation of synthetic data
  - A major problem with supervised learning is the necessity of having large amounts of training data to obtain a good result.
  - Why not create synthetic training data from real, labeled data?
  - Example: use a 3D model to generate multiple 2D images of some object (such as a face) under different conditions (such as lighting).
  - Labeling only needs to be done for the 3D model, not for every 2D model.

The Future of Supervised Learning (2)

- Future applications
  - Personal software assistants learning from past usage the evolving interests of their users in order to highlight relevant news (e.g., filtering scientific journals for articles of interest)
  - Houses learning from experience to optimize energy costs based on the particular usage patterns of their occupants
  - Analysis of medical records to assess which treatments are more effective for new diseases
  - Enable robots to better interact with humans