COMP 102: Computers and Computing
Lecture 22: Machine Learning

Instructor: Joelle Pineau (jpineau@cs.mcgill.ca)

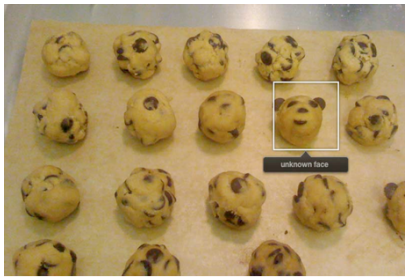
Class web page: www.cs.mcgill.ca/~jpineau/comp102

Machine learning at work

- Handwritten digit recognition: >99% accuracy (on a large dataset)

Machine learning at work

- Image segmentation
- Object / Face recognition



Other applications of machine learning

- Spam filtering
- Fraud detection
- Weather prediction
- Customer segmentation
- Categorization of news articles by topic
- …

Example: a data set

- Wisconsin Breast tumor data (from UC-Irvine Machine Learning repository).
- Thirty real-valued variables per case.
- Two variables that can be predicted:
  - Output (R=recurrence, N=non-recurrence)
  - Time (until recurrence for R, healthy for N)

tumor size	texture	perimeter	...	outcome	time
18.02	27.6	117.5		N	31
17.99	10.38	122.8		N	61
20.29	14.34	135.1		R	27
...					

A general problem

- Given a set of **labeled examples** $<x_1, x_2, x_3, \ldots, x_n, y>$ where x_i are **input variables** and y is the **desired output**.
- We want to learn a function $h : X_1 \times X_2 \times \ldots \times X_n \rightarrow Y$ which maps the input variables onto the output domain.
- We can then use that function to label new examples.

  E.g. Tell whether a person will have a recurrence, based on their features.

Terminology

- Columns are called **input variables** or **features** or **attributes**.
- The columns we are trying to predict (outcome and time) are called **output variables** or **targets**.
- A row in the table is called a **training example** or **instance**.
- The whole table is called a **data set**.

tumor size	texture	perimeter	...	outcome	time
18.02	27.6	117.5		N	31
17.99	10.38	122.8		N	61
20.29	14.34	135.1		R	27
...					

More formally

- A training example i has the form $<x_{i,1}, x_{i,2}, \ldots x_{i,n}, y_i>$, where n is the number of attributes.
- Notation $\mathbf{x}_i$ denotes the column vector with elements $x_{1,i}, \ldots, x_{n,i}$.
- The training set consists of m training examples.
- Let $X = X_1 \times X_2 \times \ldots \times X_n$ denote the space of input values.
- Let $Y = Y_1 \times Y_2$ denote the space of output values.

Supervised learning problem

- Given a dataset $D = X \times Y$, find a function: $h : X \rightarrow Y$ such that $h(x)$ is a good predictor for the value of y.
- Formally, h is called the **hypothesis**.
- Output Y can have many types:
  - If $Y = \Re$, this problem is called **regression**.
  - If Y is a finite discrete set, the problem is called **classification**.
  - If Y has 2 elements, the problem is called **binary classification** or **concept learning**.

Prediction problems

- The problem of predicting tumour recurrence is called:

  classification
- The problem of predicting the time of recurrence is called:

  regression
- Treat them as two separate supervised learning problems.

tumor size	texture	perimeter	...	outcome	time
18.02	27.6	117.5		N	31
17.99	10.38	122.8		N	61
20.29	14.34	135.1		R	27
...					

Hard question

- What form should the function *H* take?
- Many possibilities!
  - Rule-based learning
  - Decision-trees
  - Neural networks
  - Many other…

Rule-based learning for classification

- Observe a number of labeled examples.
- Pick a few rules that correctly describe your examples.

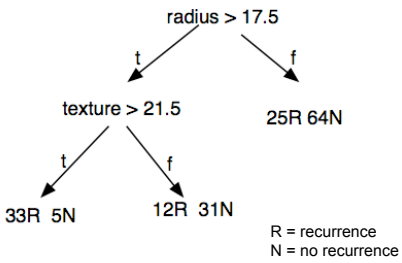
IF	THEN most likely class is
radius>17.5 AND texture>21.5	R
radius>17.5 AND texture≤21.5	N
radius≤17.5	N

- Apply these rules when you need to label a new case.

How can we choose good rules?

Decision tree example

- What does a node represent?
  - A partitioning of the input space.
- Internal nodes are tests on the values of different attributes
  - Don't need to be binary test.



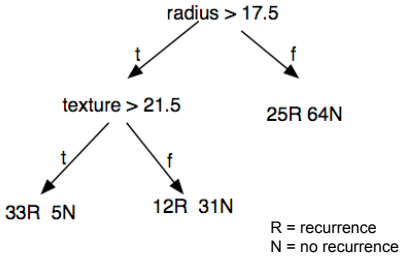
- Leaf nodes also include the set of training examples that satisfy the tests along the branch.
  - Each training example falls in precisely one leaf.
  - Each leaf typical contains more than one example.

Using decision trees for classification

- Suppose we get a new instance:

  radius=18, texture=12, …

  How do we classify it?

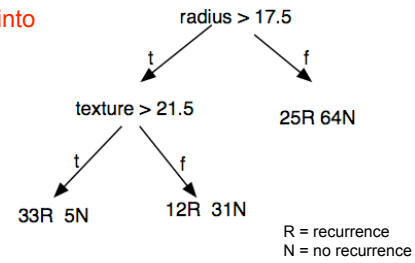


- Simple procedure:
  - At every node, test the corresponding attribute
  - Follow the appropriate branch of the tree
  - At a leaf, either predict the class of the majority of the examples for that leaf, or sample from the probabilities of the two classes.

Interpreting decision trees

- We can always convert a decision tree into an equivalent set of if-then rules.

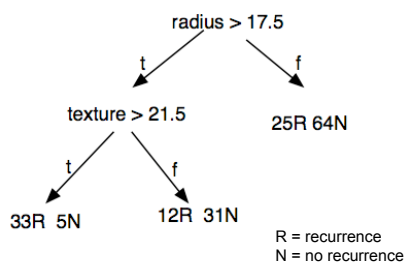
  (The reverse is not always possible.)



IF	THEN most likely class is
radius>17.5 AND texture>21.5	R
radius>17.5 AND texture≤21.5	N
radius≤17.5	N

Interpreting decision trees

- We can also calculate an estimated probability of recurrence.



IF	THEN P(R) is
radius>17.5 AND texture>21.5	$\frac{33}{33+5}$
radius>17.5 AND texture≤21.5	$\frac{12}{12+31}$
radius≤17.5	$\frac{25}{25+64}$

More Formally : Tests

- Each internal node contains a test that typically depends on one feature.
  - For discrete features, we typically branch on all possibilities
  - For real features, we typically branch on a threshold value
- Test returns a discrete outcome, e.g.
  - radius > 17.5
  - radius $\in [12, 18]$
  - grade is $\{A, B, C, D, F\}$
  - grade is $\geq B$
  - color is RED
  - $2*$radius$-3*$texture > 16
- Learning = choosing the tests at every node and the shape of the tree.
  - A finite set of candidate tests is usually chosen before learning the tree.

An algorithm for learning decision trees

Given a set of labeled training instances:

1. If all the training instances have the same class, create a leaf with that class label and exit.
2. Pick the best test to split the data on.
3. Split the training set according to the value of the outcome of the test.
4. Recursively repeat steps 2 and 3 on each subset of the training data.

How do we pick the **best test**?

What is a good Test?

- The test should provide **information** about the class label.

  E.g. You are given 100 examples: 40 R, 60 NR

  Consider two tests that would split the examples as follows: Which is best?

- Intuitively, we prefer an attribute that separates the training instances as well as possible. How do we quantify this (mathematically)?

Quick recap on information theory

- Consider two cases:
  - You are about to observe the outcome of a dice roll
  - You are about to observe the outcome of a coin flip

- Intuitively, in each situation, you have a different amount of uncertainty as to what outcome / message you will observe.

Information content

- Let E be an event that occurs with probability $P(E)$. If we are told that E has occurred with certainty, then we received $I(E)$ **bits of information**.

$$I(E) = \log_2 \frac{1}{P(E)}$$

- You can also think of information as the amount of "surprise" in the outcome (e.g., consider $P(E) = 1$, then $I(E) \approx 0$)

  E.g.
  - fair coin flip provides $\log_2 2 = 1$ bit of information
  - fair dice roll provides $\log_2 6 \approx 2.58$ bits of information

Entropy

- Suppose we have an information source S which emits symbols from an alphabet $\{s_1, \ldots s_k\}$ with probabilities $\{p_1, \ldots p_k\}$.
- Each emission is independent of the others. What is the **average amount of information** we expect from the output of S?

$$H(S) = \sum_i p_i I(s_i) = \sum_i p_i \log \frac{1}{p_i} = -\sum_i p_i \log p_i$$

- Call this the **entropy** of S.

Binary Classification

- Suppose we have examples from 2 classes *p* and *n*.
- What is the entropy of this data set?

$$H(D) = -\frac{p}{p+n} log_2 \frac{p}{p+n} - \frac{n}{p+n} log_2 \frac{n}{p+n}$$

- Back to our breast cancer example, what is the entropy?

  $H(D) = - (40/100) log_2(40/100) - (60/100) log_2(60/100)$

Entropy of an attribute/feature

- Compare the entropy of our two possible features:

- Calculate the conditional entropy of each feature:

  $H(D \mid Rad) = (60/100) [-(10/60) log_2(10/60) - (50/60) log_2(50/60)] +$
  $(40/100) [-(30/40) log_2(30/40) - (10/40) log_2(10/40)]$

  $H(D \mid Smoker) = (50/100) [-(20/50) log_2(20/50) - (30/50) log_2(30/50)] +$
  $(50/100) [-(20/50) log_2(20/50) - (30/50) log_2(30/50)]$

Information gain

- We want to choose features with low **conditional entropy** *H(D | x)*.
- *H(D | x) = 0* means that the data is perfectly separated.
- How much information do we gain when testing an attribute *x*?

$$IG(x) = H(D) - H(D|x)$$

An algorithm for learning decision trees

Given a set of labeled training instances:

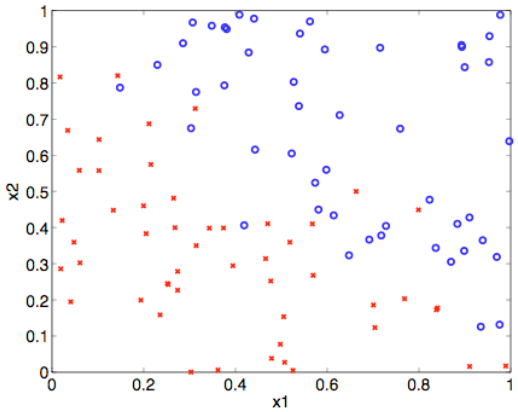
1. If all the training instances have the same class, create a leaf with that class label and exit.
2. Pick the best test to split the data on .
3. Split the training set according to the value of the outcome of the test.
4. Recursively repeat steps 2 and 3 on each subset of the training data.

How do we pick the **best test**?

Answer: We choose the test that has highest information gain.

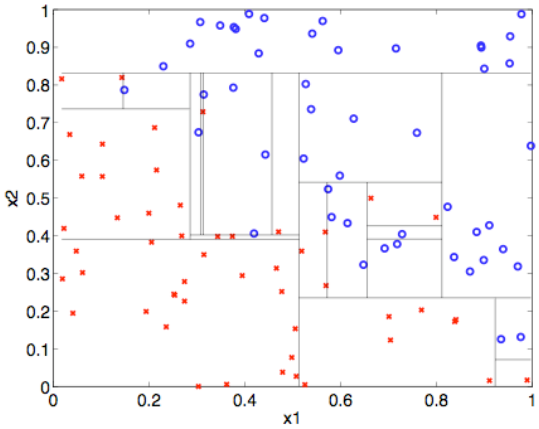
A complete (artificial) example

- An artificial binary classification problem with two real-valued input features:



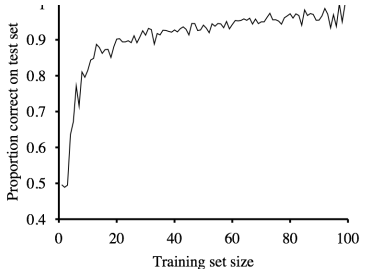
A complete (artificial) example

- The decision tree, graphically:



Assessing Performance

- Split data into a **training set** and a **test set**
- Apply learning algorithm to training set
- Measure error with the test set



From Russell&Norvig, AIMA

Regression trees

- Similar to classification trees, but for regression problems.
- We try to predict a numerical value.
- Tests can be the same as before.
- At the leaves, instead of predicting the majority of the examples, we define a linear function of some subset of numerical values (e.g. predict the mean).

Applications of Decision Trees

- Simple things like a How-To
- Failure Diagnosis
- And many more …

- More sophisticated example: Classification of financial data
  - Explanation of stock dynamics using publicly available information
  - Classify stocks as undervalued, overvalued or neutral

Application of Decision Trees

Input Features

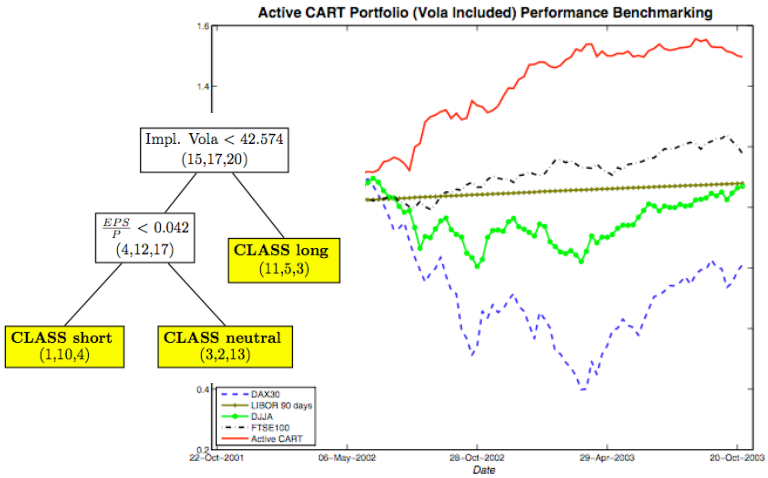
Indicator	Type	Frequency	Description
Momentum	Technical	1 day	$M_t = P_t - P_{t-T},\ T = 20$
Stochastic	Technical	1 day	$\frac{P_t - P_L}{P_H - P_L},\ P_H = max(P_t),\ P_L = min(P_t)$
MA	Technical	1 day	$MA(T) = \frac{\sum_{i=t-T}^{t} P_i}{T},\ T = 12$
MA St. Error	Technical	1 day	Standard deviation of MA
MACD	Technical	1 day	$(1 - \frac{n_1}{n_2})\{MA(n_1) - MA(n_2 - n_1)\}$ $n_1 = 12,\ n_2 = 26$
ROC	Technical	1 day	$\frac{P_t}{P_{t-T}},\ T = 10$
TRIX	Technical	1 day	Triple exponentially smoothed MA
BV	Fundamental	1 month	Book Value
CF	Fundamental	1 month	Cash Flow
Dividends paid	Fundamental	1 month	-
			Depreciation
EPS	Fundamental	1 month	Earnings Per Share
Sales	Fundamental	1 month	-
ImplVola	Fundamental	1 day	Implied volatility

Output

Short
Long
Neutral

http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2008-009.pdf

Application of Decision Trees



Take-home message

- Understand the goal of machine learning.
- Understand the main technical terms: classification, regression, dataset, input/output variables.
- Have a sense for how information-theory is used to build a decision tree.