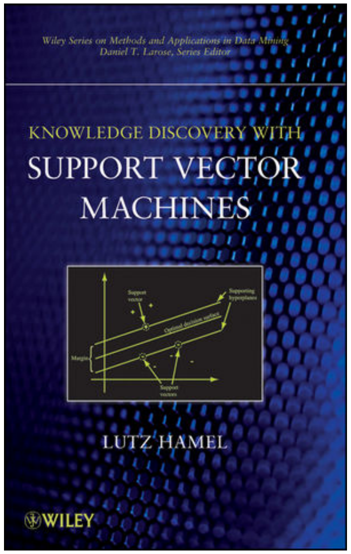
Welcome - CSC 581

Introduction to Machine Learning with Support Vector Machines

Dr. Lutz Hamel
hamel@cs.uri.edu
Tyler Hall, Rm 251



Reference Material:

We will be using my book *Knowledge Discovery with Support Vector Machines*, Hamel, Wiley, 2009.

Other books of interest:

Introductory Statistics with R, Peter Dalgaard, Springer, 2008.

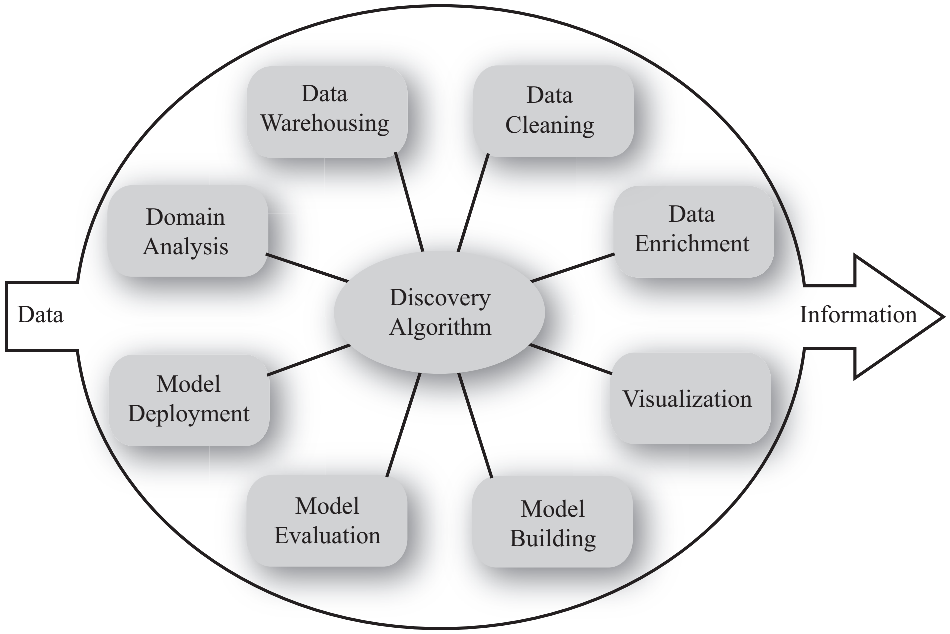
An Introduction to Support Vector Machines, Nello Cristianini and John Shawe-Taylor, Cambridge University Press, 2000.

The Nature of Statistical Learning Theory, Vladimir Vapnik, 2nd Edition, Springer, 2000.

Learning with Kernels, Berhard Schoelkopf and Alexander Smola, MIT Press, 2002.

Kernel Methods in Computational Biology, Berhard Schoelkopf, Koji Tsuda, and Jean-Philippe Vert (Editors), MIT Press, 2004.

Knowledge Discovery



- Semi-automated process of extracting useful information from collections of data.
- Computer-based tools for the discovery process but that guidance by an analyst is indispensable.
- Highly interdisciplinary.
- Data Mining - Knowledge Discovery in Databases (KDD)

Machine Learning

From our perspective, machine learning is at the core of knowledge discovery.

A qualitative definition of machine learning:

> Programs that get better with experience given a task and some performance measure.

Machine Learning

A more quantitative definition of machine learning:

Given

- A data universe X.
- A sample set S where $S \subset X$.
- Some target function (labeling process) $f : X \rightarrow \{\mathit{true}, \mathit{false}\}$.
- A labeled training set D, where $D = \{(x, y) \mid x \in S$ and $y = f(x)\}$.

Compute a function $\hat{f} : X \rightarrow \{\mathit{true}, \mathit{false}\}$ using D such that,

$$\hat{f}(x) \cong f(x),$$

for all $x \in X$.

This definition of machine learning is referred to as *supervised learning* due to the fact that the algorithm needs a labeled dataset D.

Observation: We can view the function $\hat{f}$ as a *model* or *approximation* of the original function f. The model is computed only based on the observations in the training dataset D.

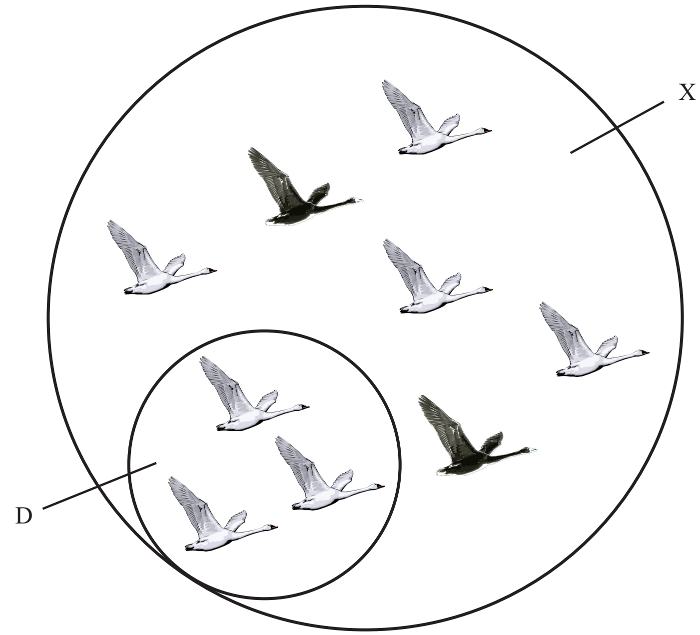
Inductive Reasoning

The fundamental assumption in machine learning is that the training set D is an accurate representation of the whole data universe X.

> **The Inductive Learning Hypothesis:** *Any function found to approximate the target function well over a sufficiently large training set* D *will also approximate the target function well over the data universe* X.

But...

Inductive reasoning has its pitfalls, as can be demonstrated with the classic black swan problem.



That is, if your training set D is not representative of the data universe X then your model, in this case "all swans are white", will most likely not be correct.

The Universe X

A convenient way to describe objects in a data universe X is by the use of a *feature table*.

	Legs	Wings	Fur	Feathers
cat	4	no	yes	no
crow	2	yes	no	yes
frog	4	no	no	no
bat	4	yes	yes	no
barstool	3	no	no	no

- Each labeled column in the table is called an *attribute*.
- Each labeled row is an object of the universe.
- This is only a subset of all possible objects that can be described with the attributes (sample set).

The Universe X

Let $mammal : \; X \rightarrow \{true, false\}$ be a target function, then we can convert our feature table into a training set by (a) dropping the names of the objects and (b) adding a column with the labels generated by $mammal$:

Legs	Wings	Fur	Feathers	Mammal
4	no	yes	no	*true*
2	yes	no	yes	*false*
4	no	no	no	*false*
4	yes	yes	no	*true*
3	no	no	no	*false*

A reasonable model $\hat{f}$ for $mammal$ based on this training set is,

$$\hat{f}(legs, wings, fur, feathers) \equiv \textbf{ if } fur = \text{ yes } \textbf{ then } true \textbf{ else } false.$$

Representations of $\hat{f}$

We want to compute the model $\hat{f}$. In order to accomplish this we need to pick a representation of the model. Typically we consider two types of representations:

- transparent representations (or transparent models):
  - If-then-else rules
  - Decision trees
- non-transparent representations (or non-transparent models):
  - The weights on the connections between the elements in an artificial neural network.
  - the linear combination of vectors in support vector machines.

Transparent models are representations that can be interpreted by humans unaided, non-transparent models cannot be interpreted unaided.

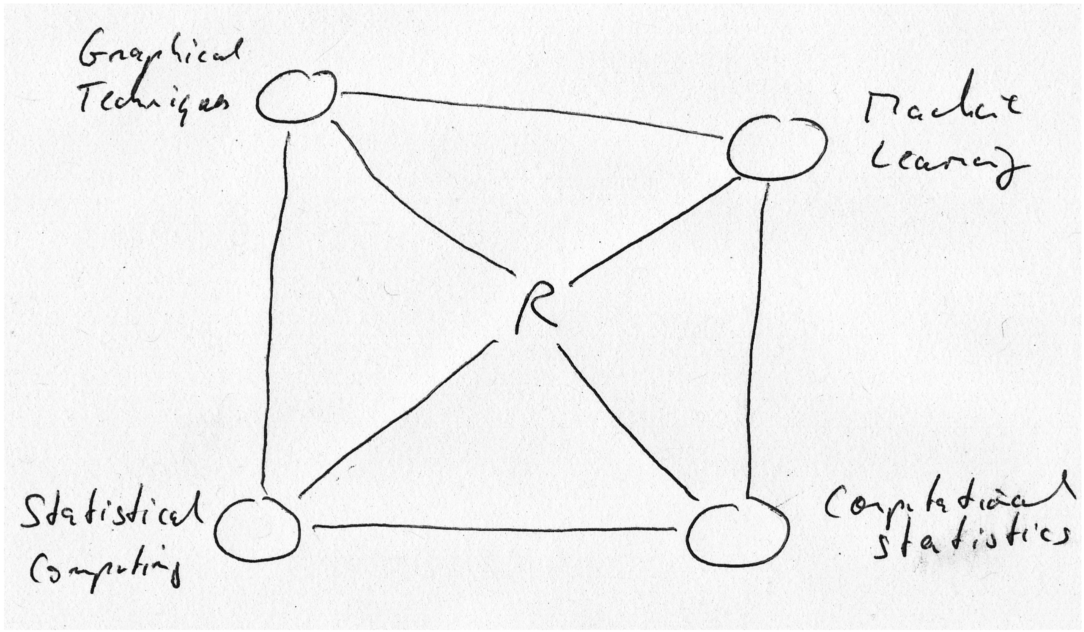
Why?

Why should we consider machine learning as a way to compute the model $\hat{f}$ rather than looking at other techniques such as linear models from statistics?

It turns out that it is really simply a matter of what kinds of assumptions you admit in your analysis/model building. Most statistical techniques rely on the fact that there is some normal distribution of the data/error. Machine learning techniques do not make these assumptions and therefore are able to provide more accurate models in situations where normality assumptions are not warranted.

When machine learning techniques are strictly applied to tabular data as part of data analyses we can consider machine learning as part of computational statistics. An area of statistics that explores models via computational experiments and resampling rather than normality assumptions.

Why?



Graphical Techniques: scatter plots, histograms

Statistical Computing: hypothesis testing, linear regression, generalized linear models

Computational Statistics: bootstrap, monte carlo

Machine Learning: computational model building

R: statistical computing environment supporting all of the above activities