Typical Machine Learning and Data Mining Problems

Document search:

Given counts of words in a document, determine what its topic is.

Group documents by topic without a pre-specified list of topics.

Many words in a document, many, many documents available on the web.

Cancer diagnosis:

Given data on expression levels of genes, classify the type of a tumor.

Discover categories of tumors having different characteristics.

Expression levels of many genes measured, usually for only a few patients.

Marketing:

Given data on age, income, etc., predict how much each customer spends.

Discover how the spending behaviours of different customers are related.

Fair amount of data on each customer, but messy (eg, missing values).

May have data on a very large number of customers (millions).

Supervised Learning Problems

In the ML literature, a *supervised learning* problem has these characteristics:

We are primarily interested in prediction.

We are interested in predicting only one thing.

The possible values of what we want to predict are specified, and we have some *training* cases for which its value is known.

The thing we want to predict is called the *target* or the *response variable*.

For a *classification* problem, we want to predict the class of an item — the topic of a document, the type of a tumor, whether a customer will purchase a product.

For a *regression* problem, we want to predict a numerical quantity — the amount a customer spends, the blood pressure of a patient, the melting point of an alloy.

To help us make our predictions, we have various *inputs* (also called *features* or *predictors*) — eg, gene expression levels for predicting tumor type, age and income for predicting amount spent. We use these inputs, but don't try to predict them.

Unsupervised Learning Problems

For an *unsupervised learning* problem, we do not focus on prediction of any particular thing, but rather try to find interesting aspects of the data.

One non-statistical formulation: We try to find *clusters* of similar items, or to *reduce the dimensionality* of the data.

Examples: We may find clusters of patients with similar symptoms, which we call "diseases". We may find that an overall "inflation rate" captures most of the information present in the price increases for many commodities.

One statistical formulation: We try to learn the *probability distribution* of all the quantities, often using *latent* (also called *hidden*) variables.

These formulations are related, since the latent variables may identify clusters or correspond to low-dimensional representations of the data.

Machine Learning and Data Mining Problems Versus Problems in Traditional Statistics

Motivations:

Prediction · Understanding · Causality

Much traditional statistics is motivated primarily by showing that one factor causes another (eg, clinical trials). Understanding comes next, prediction last.

In machine learning and data mining, the order is usually reversed — prediction is most important.

Amount of Data:

Many machine learning problems have a large number of variables — maybe 100,000 or more. Data mining applications often involve very large numbers of cases — sometimes millions.

Complex, non-linear relationships:

Traditional statistical methods often assume linear relationships (perhaps after simple transformations), or simple distributions (eg, normal).

Attitudes in Machine Learning and Data Mining Versus Attitudes in Traditional Statistics

Despite these differences, there's a big overlap in problems addressed by machine learning and data mining and by traditional statistics. But attitudes differ...

Machine learning	**Traditional statistics**
No settled philosophy or widely accepted theoretical framework.	Classical (frequentist) and Bayesian philosophies compete.
Willing to use *ad hoc* methods if they seem to work well (though appearances may be misleading).	Reluctant to use methods without some theoretical justification (even if the justification is actually meaningless)
Emphasis on automatic methods with little or no human intervention.	Emphasis on use of human judgement assisted by plots and diagnostics.
Methods suitable for many problems.	Models based on scientific knowledge.
Heavy use of computing.	Originally designed for hand-calculation, but computing is now very important.

How Do “Machine Learning” and “Data Mining” Differ?

These terms are often used interchangeably, but. . .

Data mining is more often used for problems with very large amounts of data, where computational efficiency is more important than statistical sophistication — often business applications.

Machine learning is more often used for problems with a flavour of artificial intelligence — such as recognition of objects in visual scenes, or robot navigation.

The term “data mining” was previously used in a negative sense — to describe the misguided statistical procedure of looking for many, many relationships in the data until you finally find one, but one which is probably just due to chance.

One challenge of data mining is to avoid doing “data mining” in this sense!