COS 424: Interacting with Data

Lecturer: Rob Schapire & Dave Blei — Lecture #1
Scribe: Jonathan Chang — February 6, 2007

1 What is this course about? (Rob)

This course is about data! Data is everywhere. Everything is computerized now and vast amounts of data can be easily stored. Concomitant with vast amounts of data is the belief that this data will be useful. There are many practical issues regarding data which this course will **not** cover such as storing data, databases, transferring data, etc. This class **will** be concerned with how to get the most out of data and convert data into knowledge, information or predictions.

1.1 Examples

As examples of the many kinds of datasets that abound in the world, we considered specifically examples in which our ordinary activities are in some way recorded.

credit cards Every purchase you make is tracked. This information can be used to detect fraud or for marketing purposes. Your credit payment history is also used to help compute a credit score, which is essentially a prediction of how likely you are to make timely payments in the future.

web browsing Where you visit and what you search for may be tracked. Again, this may be used for marketing, but it may also be used to track terrorists, for example. It can also be used to help determine how to optimize a network configuration.

purchase history Sites such as netflix and amazon maintain records of the items you purchase or rent. They can use this to make predictions about what you will like.

border crossings

criminal records

online newspapers

security cameras Security cameras can in principal be used for tracking (in order to enforce fines, for example), or for finding terrorists or criminals via facial recognition software.

traffic patterns

cell phones Another way in which the location of a person can be tracked is via his or her cell phone.

telephone call records Every time a call is made the caller and callee are recorded. Again, this data can be used to detect fraud and for marketing purposes. It can also be used for security. Further, one can build a graph of calls for visualization purposes. The graph can be broken up into communities, facilitating sociological research into “social networking.”

medical records

bioinformatics, astronomy, physics These scientific disciplines are generating vast amounts of data (e.g. gene sequences, stars in the universe, etc.). Organizing and using this data has become a principal driving force in these fields.

fMRI fMRI allows the blood flow within the brain to be imaged. By analyzing this data with regard to what the subject is thinking about, it may let us guess what people are thinking.

pollution

articles Any time you write an article or a paper, it is indexed in multiple databases. With this data, one could organize articles by topics or even track the evolution of topics over time.

1.2 Summary

In summary, there are all kinds of data (text, images, transaction records, etc.). Data can vary in terms of the form of the data, the size of the dataset, and the size and complexity of each record. Governments, businesses, scientists, engineers and computer scientists are all interested in this data.

1.3 Tasks

There are many things one might want to do with data, including the following generic tasks which arise in many settings:

- make predictions or classifications. For instance, we might want to classify customers as to whether or not they are about to switch companies.

- cluster or organize data. For instance, we might want to cluster articles by topic, or cluster movie-goers by the type of movie they typically enjoy. (This is different from classification because you don't know the classes ahead of time.)

- find "simple" descriptions of complex objects. For instance, we might want to find a simple description of faces which would make it easier to compare two faces.

- identify what is typical and what is an outlier. This is closely related to the last item. For instance, we might want to identify purchases that are typical or unusual (and thus possibly fraudulent) for a given customer.

- uncover underlying truths. For instance, we might want to find the underlying causes of some disease. In general, this might not always be so important so long as we can make good predictions; we often don't care *why* the predictions are good.

The course will look at several general-purpose and modern approaches focusing on practical techniques. The problems presented in this course are by no means solved. We never know if a certain technique will work on a new data set. The goal of the course is to examine the underlying assumptions of each of the approaches and to help develop an intuition as to which techniques will work where.

1.4 Perspectives

For more than a century, statisticians have been working on these problems (although for much of that time, they were limited to dealing with small data sets for lack of computers). Pattern recognition arose in the 60s and was primarily concerned with images. Machine learning started in the 80s and is still an ongoing, active field of research. Machine learning was a natural outgrowth of Artificial Intelligence (AI), as learning is a principal feature of "intelligence." Another related field is data mining which arose in the 90s in order to deal with the vast amounts of data that were being collected with the goal of discovering "interesting patterns."

This course is largely a mixture of statistics, machine learning, and data mining. The goal will be to instill practical knowledge of how to get the most out of data. We will explore the kind of general tasks described above, covering a range of methods and algorithms, focusing on underlying assumptions and developing a practical sense of what to use when. There will be some theory, but the course will primarily be practical and hands on. The course will look at interacting with data *vis à vis* classification, clustering, regression, and dimensionality reduction.

In comparison, COS402 is a broad survey of AI of which machine learning is a component. This course will be much more focused on what to do with data. There will be some overlap, particularly with regard to classification learning, although the emphasis will be quite different. COS511 is a theoretical treatment of machine learning, while the treatment in this course will be grounded in practice.

2 Some Probability and Statistics (Dave)

2.1 Random variable

Probability is about *random variables.* A random variable is any "probabilistic" outcome. For example, the outcome of flipping a coin is a random variable, as is the height of someone chosen randomly from a population. We'll see that it's sometimes useful to think of quantities that are not strictly probabilistic as random variables. For instance, the temperature on 11/12/2013 is not truly random; the temperature on 03/04/1905, since it already happened, certainly is not random. And the number of times the word "streetlight" appears in a document is also not clearly random. Thus, probability can encode a degree of belief or uncertainty.

Random variables take on values in a *sample space.* They can be discrete or continuous. For instance, a coin flip has values $\{H, T\}$. A person's height has positive real values: $(0, \infty)$. Temperature takes real values $(-\infty, \infty)$. And the number of words in a document is a positive integer $\{1, 2, \ldots\}$.

We typically denote random variables with capital letters, and we denote a realization of the random variable with a lower case letter. For instance, X is a coin flip, x is the value (H or T) of that coin flip.

2.2 Discrete distribution

A discrete distribution assigns a probability to every element in the sample space. For example, if X is an (unfair) coin, then we might have

$$\begin{aligned} P(X = H) &= 0.7 \\ P(X = T) &= 0.3 \end{aligned}$$

The probabilities over the entire space must sum to one:

$$\sum_x P(X = x) = 1.$$

Probabilities of disjunctions are sums over part of the space. For instance, the probability that a die is bigger than 3 is

$$P(D > 3) = P(D = 4) + P(D = 5) + P(D = 6) = 1/2.$$

2.3 Joint distribution

Typically, we consider collections of random variables. The joint distribution is a distribution over the configuration of all the random variables in the ensemble. For example, imagine flipping 4 coins. The joint distribution is over the space of all possible outcomes of the four coins.

$$\begin{aligned} P(HHHH) &= 0.0625 \\ P(HHHT) &= 0.0625 \\ P(HHTH) &= 0.0625 \\ &\ldots \end{aligned}$$

You can think of it as a single random variable with 16 values.

2.4 Conditional distribution

A *conditional distribution* is the distribution over a random variable given some evidence. Thus, $P(X = x|Y = y)$ is the probability that $X = x$ when $Y = y$. For example,

$$\begin{aligned} P(\text{I listen to Steely Dan}) &= 0.5 \\ P(\text{I listen to Steely Dan}|\text{Toni is home}) &= 0.1 \\ P(\text{I listen to Steely Dan}|\text{Toni is not home}) &= 0.7 \end{aligned}$$

$P(X = x|Y = y)$ is a different distribution for each value of y so that

$$\sum_x P(X = x|Y = y) = 1$$

but it is not necessarily true that

$$\sum_y P(X = x|Y = y)$$

sums to one. Conditional probability is defined as:

$$P(X|Y) = \frac{P(X,Y)}{P(Y)},$$

which holds when $P(Y) > 0$.

Two important consequences of conditional probability are the *chain rule* which states that

$$P(X,Y) = P(X|Y)P(Y)$$

and *Bayes rule* which states that

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}.$$