Everything You Wanted to Know About Data Mining but Were Afraid to Ask

By Alexander Furnas

A guide to what data mining is, how it works, and why it's important.

Big data is everywhere we look these days. Businesses are falling all over themselves to hire 'data scientists,' privacy advocates are concerned about personal data and control, and technologists and entrepreneurs scramble to find new ways to collect, control and monetize data. We know that data is powerful and valuable. But how?

This article is an attempt to explain how data mining works and why you should care about it. Because when we think about how our data is being used, it is crucial to understand the power of this practice. Without data mining, when you give someone access to information about you, all they know is what you have told them. With data mining, they know what you have told them and can guess a great deal more. Put another way, data mining allows companies and governments to use the

information you provide to reveal more than you think.

To most of us data mining goes something like this: tons of data is collected, then quant wizards work their arcane magic, and then they know *all of this amazing stuff*. But, how? And what types of things can they know? Here is the truth: despite the fact that the specific technical functioning of data mining algorithms is quite complex -- they are a black box unless you are a professional statistician or computer scientist -- the uses and capabilities of these approaches are, in fact, quite comprehensible and intuitive.

For the most part, data mining tells us about very large and complex data sets, the kinds of information that would be readily apparent about small and simple things. For example, it can tell us that "one of these things is not like the other" a la Sesame Street or it can show us categories and then sort things into pre-determined categories. But what's simple with 5 datapoints is not so simple with 5 billion datapoints.

And these days, there's always more data. We gather far more of it then we can digest. Nearly every transaction or interaction leaves a data signature that someone somewhere is capturing and storing. This is, of course, true on the internet; but, ubiquitous computing and digitization has made it increasingly true about our lives away from our computers (do we still have those?). The sheer scale of this data has far exceeded human sense-making capabilities. At these scales patterns are often too subtle and relationships too complex or multi-dimensional to observe by simply looking at the data. Data mining is a means of automating part this process to detect interpretable patterns; it helps us see the forest without getting lost in the trees.

Discovering information from data takes two major forms: description and prediction. At the scale we are talking about, it is hard to know what the data shows. Data mining is used to simplify and summarize the data in a manner that we can understand, and then allow us to infer things about specific cases based on the patterns we have observed. Of course, specific applications of data mining methods are limited by the data and computing power available, and are tailored for specific needs and goals. However, there are several main types of pattern detection that are commonly used. These general forms illustrate what data mining can do.

Anomaly detection : in a large data set it is possible to get a picture of what the data tends to look like in a typical case. Statistics can be used to determine if something is notably different from this pattern. For instance, the IRS could model typical tax returns and use anomaly detection to identify specific returns that differ from this for review and audit.

Association learning: This is the type of data mining that drives the Amazon recommendation system. For instance, this might reveal that customers who bought a cocktail shaker and a cocktail recipe book also often buy martini glasses. These types of findings are often used for targeting coupons/deals or advertising. Similarly, this form of data mining (albeit a quite complex version) is behind Netflix movie recommendations.

Cluster detection: one type of pattern recognition that is particularly useful is recognizing distinct clusters or sub-categories within the data. Without data mining, an analyst would have to look at the data and decide on a set of categories which they believe captures the relevant distinctions between apparent groups in the data. This would risk missing important categories. With data mining it is

possible to let the data itself determine the groups. This is one of the black-box type of algorithms that are hard to understand. But in a simple example - again with purchasing behavior - we can imagine that the purchasing habits of different hobbyists would look quite different from each other: gardeners, fishermen and model airplane enthusiasts would all be quite distinct. Machine learning algorithms can detect all of the different subgroups within a dataset that differ significantly from each other.

Classification: If an existing structure is already known, data mining can be used to classify new cases into these pre-determined categories. Learning from a large set of pre-classified examples, algorithms can detect persistent systemic differences between items in each group and apply these rules to new classification problems. Spam filters are a great example of this - large sets of emails that have been identified as spam have enabled filters to notice differences in word usage between legitimate and spam messages, and classify incoming messages according to these rules with a high degree of accuracy.

Regression: Data mining can be used to construct predictive models based on many variables. Facebook, for example, might be interested in predicting future engagement for a user based on past behavior. Factors like the amount of personal information shared, number of photos tagged, friend requests initiated or accepted, comments, likes etc. could all be included in such a model. Over time, this model could be honed to include or weight things differently as Facebook compares how the predictions differ from observed behavior. Ultimately these findings could be used to guide design in order to encourage more of the behaviors that seem to lead to increased engagement over time.

The patterns detected and structures revealed by the descriptive data mining are then often applied to predict other aspects of the data. Amazon offers a useful example of how descriptive findings are used for prediction. The (hypothetical) association between cocktail shaker and martini glass purchases, for instance, could be used, along with many other similar associations, as part of a model predicting the likelihood that a particular user will make a particular purchase. This model could match all such associations with a user's purchasing history, and predict which products they are most likely to purchase. Amazon can then serve ads based on what that user is most likely to buy.

Data mining, in this way, can grant immense inferential power. If an algorithm can correctly classify a case into known category based on limited data, it is possible to estimate a wide-range of other information about that case based on the properties of all the other cases in that category. This may sound dry, but it is how most successful Internet companies make their money and from where they draw their power.

Image: Reuters.

This article available online at:

http://www.theatlantic.com/technology/archive/2012/04/everything-you-wanted-to-know-about-data-mining-but-were-afraid-to-ask/255388/