MIT | EXperimental Learning

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Big Data and Social Analytics certificate course

MODULE 5 UNIT 1
Video 3 Transcript

MIT BDA Module 5 Unit 1 Video 3 Transcript

Speaker key

YM: Yves-Alexandre de Montjoye

HY: Hapyak

YM: So, in the previous video we learned how to use bandicoot to visualize mobile phone metadata and to extract behavioral indicators from them. In this video we're going to learn how we can combine bandicoot's behavioral indicators with machine learning algorithm to predict characteristic of the user like gender, age or personality.

00:00:31

So, before we jump in, let's do a quick recap of what machine learning is and how it works. A lot of machine learning algorithm are what we call classifiers. We train an algorithm to look at a certain number of variables to characterize, for example, a house. In machine learning these are called features and, in this example, could include information like the elevation, height or price of the house.

We can then use the trained algorithm to predict whether a house is located in New York, in blue, or in San Francisco, in green. So if you want to know more about this example, please follow the link that appears on the screen.

HY: Machine learning example

YM: So, this particular example is a tree but as we will see in the exercise sessions, there's a lot of different classifiers, ranging from the simple logistic regression to more complex one like random forest or SVMs.

So let's take a deeper look at how machine learning works. In machine learning we train an algorithm on a labeled dataset like this one. This is a dataset where for a certain number of examples we know both the features, for example, the price and elevation of the house but also what the right label is, what we want to predict; whether the house is in New York, zero, or San Francisco, one.

In practice, we divide the dataset into two. The first part called the training/validation set, say 70% of the dataset, on which we're going to train the algorithm using a k-fold cross-validation. The second part called the test set, the remaining 30% of the dataset, on which we will evaluate the results of our algorithm and make sure that we did not overfit the data.

Once the algorithm is trained we can use it on new data to predict whether a house is in New York or in San Francisco. And you can follow the same approach and use bandicoot indicators as features to predict individual characteristic of users, such as age, gender and others, based on the way they use their phone.

So what we did first was to start at small scale with personality prediction. We had students from the MIT Living Lab fill out a personality questionnaire answering questions like, do you see yourself as someone who is talkative, do you get nervous easily, do you tend to be quiet and other questions? The answers to these questions allowed us to measure their personality along five axes: neuroticism, extroversion, openness, agreeableness and conscientiousness.

We were quite happy with the result we published in 2013. What we showed is that we could predict whether people who are low, medium or high for each of the personality traits. You can see these on the screen now. For example, we could predict people's degree of neuroticism 1.7 times better than through a random guess. While a random guess would be correct 38% of the time, the algorithm is correct 63% of the time.

You can see what the accuracy was for the other personality trait. What this shows is that a properly trained machine learning algorithm can fairly accurately predict someone's personality from the way they use their phone.

That study with the MIT Living Lab was however done at small scale. In a follow-up study we did the same thing at large scale, predicting people's gender from the way they used their phone. Here we showed that we could accurately predict people's gender in two countries. Using bandicoot behavioral indicators and a small training set of 10,000 people, we could predict people's gender with an accuracy of 74.3% in a European country and 74.5% in a south Asian country.

What this mean is that you can predict information about people, such as gender, only by collecting data from a small set of five to 10,000 people. So we followed this approach in another study, building into bandicoot a new data representation, which we call weak matrices. Using this data representation, how much someone calls on a daily and hourly basis, and deep learning algorithm such as convolutional neural networks, we showed that we could predict people's age with 63.1% accuracy.

HY: In your line of work, what demographic information are you interested in predicting?

a. Age
b. Gender
c. Income
d. Employment status

Thank you

YM: So, in this video, we learned how we can train an algorithm using bandicoot behavioral indicators and then use this algorithm to predict information about your users, at large scale.