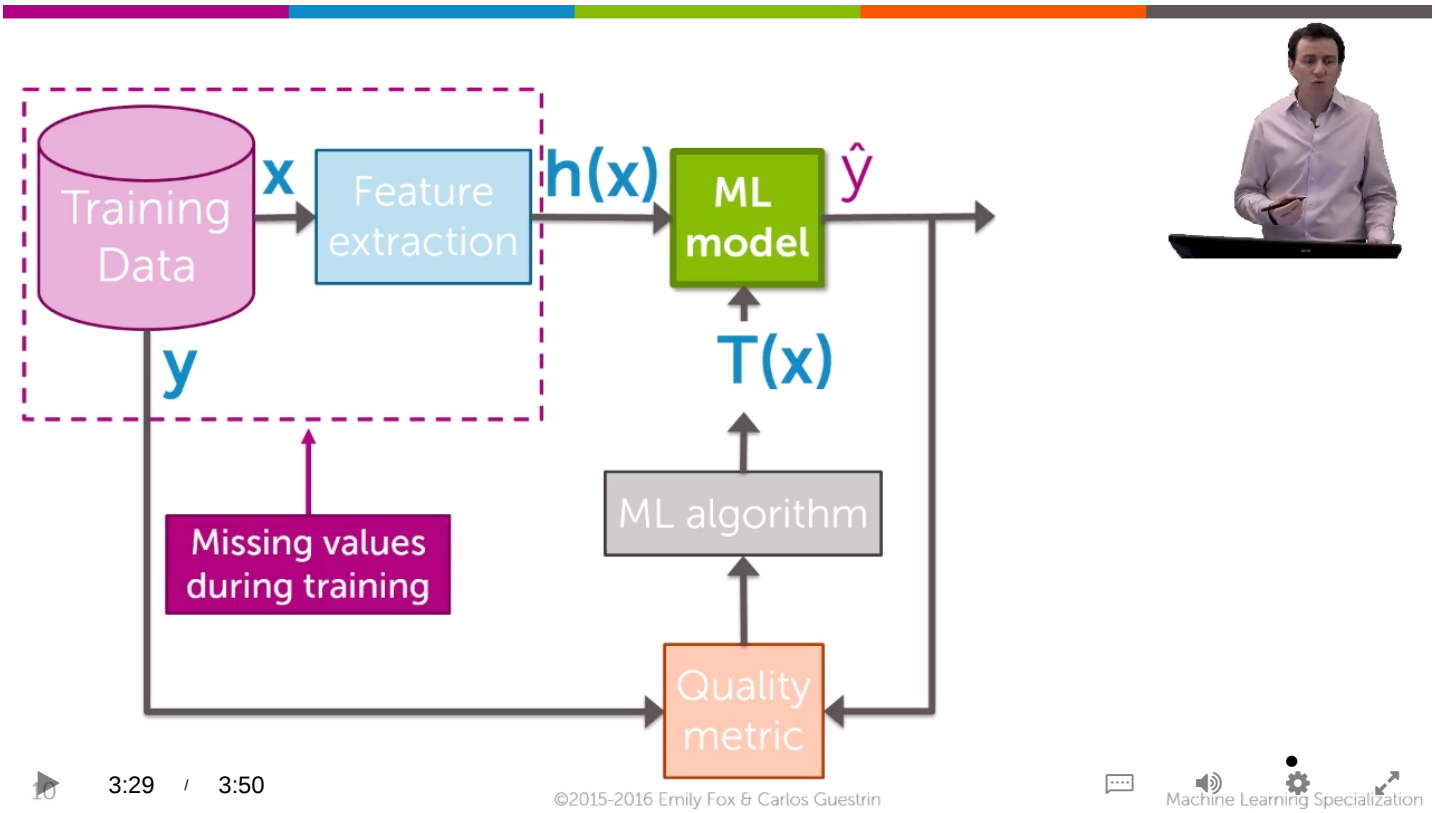
8



Challenge of missing data

Have a question? Discuss this lecture in the week forums.

Interactive Transcript

Search Transcript

English

0:00
[MUSIC] So far in this specialization data has always looked pretty beautiful. We sometimes end the look at little features in the area like taking raw text and turning into counts of words, the TFIDF, sometimes we created more advanced feature like polynomials, sines and cosines and so on. We did feature transformations, feature engineering but we always observe all of our data, so for every feature we always observe every possible value for every data point.

0:36
Now that, is rarely true in the real world. Real world data tends to be pretty messy, and often is fraught with missing data, and unobserved values. And this is a significant issue we should always be on the lookout for. In today's module, we're going to talk about some of the basic concepts and ideas of what you can do to try to address, missing data in a learning problem.

1:02
Approaches to dealing with missing data are better understood in the context of a particular learning algorithm. So for this module, we're going to pick decision trees as a way to kind of better see the impact of missing data, and some of the key approaches to dealing with it. Again, we're going to be dealing with loan data. So as input xi is coming out, another term of the loan, your credit history, and so on, we're going to push it through this crazy decision tree, set to make a decision, whether your loan is safe, or your loan is risky. Which is going to be the output y hat i, that we're trying to decide here.

1:36
As we discussed thus far, we've assumed that all the data was fully observed. So nothing was missing. So for every row of the data, for every feature we observed, for example the credit was excellent, fair, or poor. If the term of the loan was three years or five years, if the income was high or low. And we will observe the output of course, say for risky. Now, in reality you may have missing data. So missing data, for example in this highlighted row might say I know that for this particular loan application the credit was poor, the income was high. Turned out to be a risky loan, but nobody entered in there, whether the loan was a three year loan or five year loan.

2:15
And that may be true for multiple data points. And the question is, what can we do about this? What impact is on our learning algorithm, what happens? Well missing data can impact a learning algorithm both in the training phase, because I don't know how to train a model when you have this question marks we know what values those are. And it can have an impact on prediction time. Let's say I build a great decision tree, I put it out in the wild. Banks. Somebody an application in there but we don't know a particular entry. What predictions do we make?

2:49
So, let's be more specific, let's say that we have this tree that I learned from data and I have particular input where the credit was poor, the term was five years but the income was a question mark, I don't know the income of this person. So, I tried to go down the decision tree, I hit credit first, credit was poor. I hit income, that was a question mark. I didn't know it was unknown, what do we do next?

3:16
So where a learning problem, where you have some training data we tried some features, fed some machinery model, which then use a quality metric to learn a decision tree T of x. But, we're in a setting, where some of the data, might be missing a training time.

3:34
And, some of the data might be missing a prediction time. And what do we do? What we're going to do is modify the machinery model a little bit, the decision tree model a little bit, to be able to deal with this kind of missing data. Let's see how. [MUSIC]

Downloads

Lecture Video	mp4
Subtitles (English)	WebVTT
Transcript (English)	txt