Further Reading: Course Prerequisites

R or Python

If you don't know *any* programming language, that may be something you need to go away and do another course on first, sorry.

If you don't know either R or Python, but are comfortable in other languages, my tip is to look at some of the side-by-side examples in the manual, by clicking the grey "r" and "python" buttons, e.g. http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-munging/merging-data.html or http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/algo-params/max_depth.html.

Don't worry about understanding them fully at this point, just see which looks most confusing and choose the other one. There are Coursera courses on both R and Python, as well as a host of books and online tutorials. You don't need advanced skills, just the basic syntax.

If using Python, I recommend you install and become familiar with pandas. E.g. http://pandas.pydata.org/pandas-docs/stable/10min.html Familiarity with pandas will also make the R examples much clearer.

R and Python

If you already know R or Python+Pandas, and want to take this chance to get more familiar with the other, I found this presentation quite useful:

https://www.slideshare.net/ajayohri/python-for-r-users

Basic Stats

These cartoons should help remind you of the difference between mean, median and mode:

https://mathwithbaddrawings.com/2016/07/13/why-not-to-trust-statistics/

There are some pretty pictures of distributions here:

http://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm

but most important is to understand the normal distribution and standard deviation:

https://en.wikipedia.org/wiki/Standard_deviation

More stats visualizations, including a nice visual introduction to linear regression, here: https://students.brown.edu/seeing-theory

And yet more cartoons: this time good advice intermixed with xkcd cartoons on stats:

http://livefreeordichotomize.com/2016/12/15/hill-for-the-data-scientist-an-xkcd-story

Confusion Matrix

This nicely describes the basic yes/no confusion matrix:

http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

As shown in the video, it is also used when you have more than just two classes.

Bias/Variance

I find figure 1, here, helpful:

http://scott.fortmann-roe.com/docs/BiasVariance.html

Another good one here:

https://elitedatascience.com/bias-variance-tradeoff

The figure here shows clearly how a biased estimator can be better than a perfectly unbiased estimator.

https://en.wikipedia.org/wiki/Bias_of_an_estimator#Bias.2C_variance_and_mean_squared_error

Conventions

In this course (and mirroring the H2O R and Python APIs) we will use "x" for the things to learn from, and "y" to name the thing we are learning.

Other names for "x" are the features, the input variables, or the independent variables. See https://en.wikipedia.org/wiki/Feature_(machine_learning)

Other names for "y" are the label, the output variable, or the dependent variable (because it *depends* on the independent variables). See

https://en.wikipedia.org/wiki/Dependent_and_independent_variables

Mark as completed