Microsoft: DAT210x Programming with Python for Data Science

Bookmarks

- Start Here
- 1. The Big Picture
- 2. Data And Features
- 3. Exploring Data
- 4. Transforming Data
- **5. Data Modeling**

Lecture: Clustering
Quiz

Lab: Clustering
Lab

Lecture: Splitting Data
Quiz

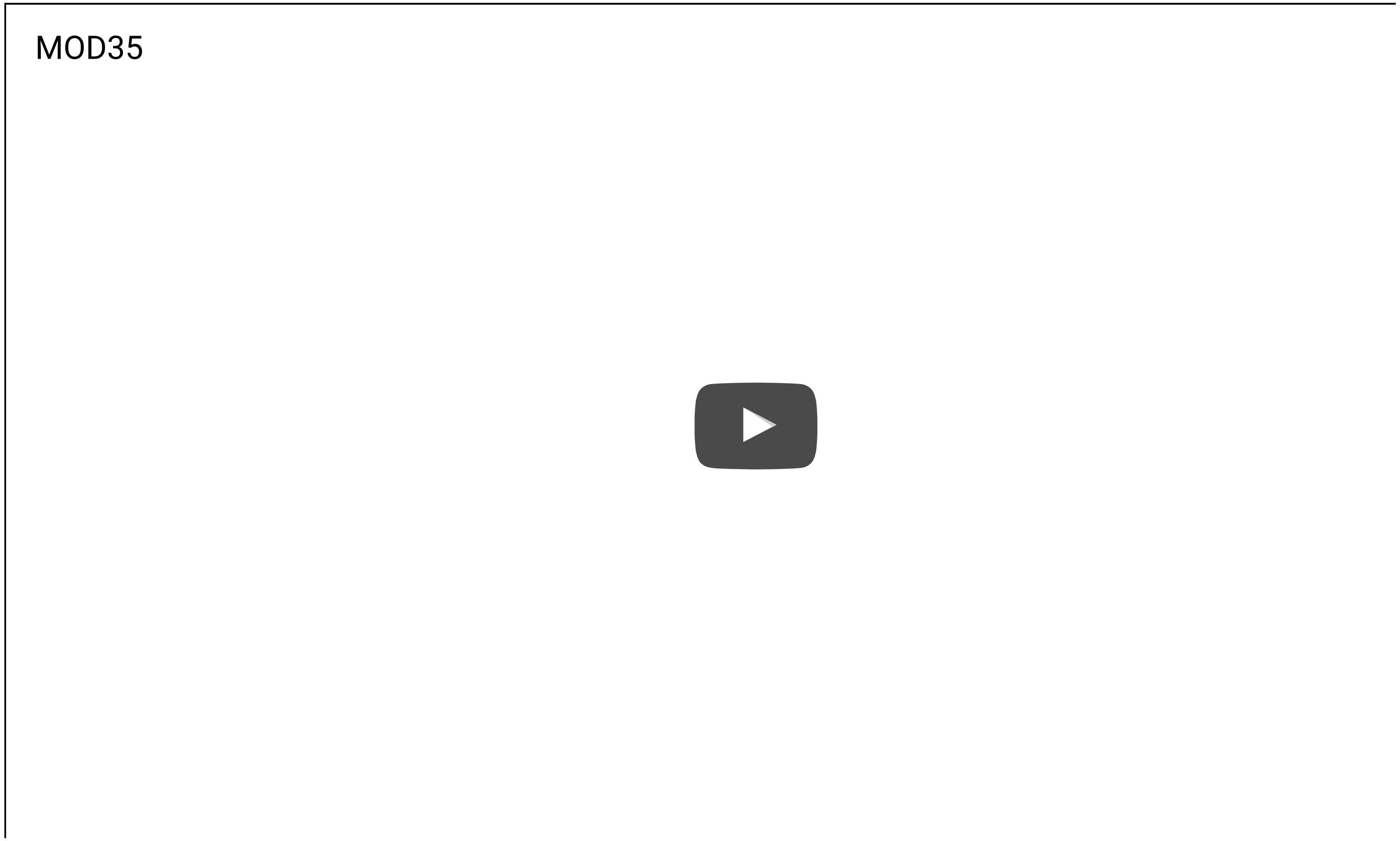
Lecture: K-Nearest Neighbors
Quiz

Lab: K-Nearest Neighbors

5. Data Modeling > Lecture: Regression > Video

Bookmark

Linear Regression



▶ 0:00 / 2:08 1.0x

Download video Download transcript .srt

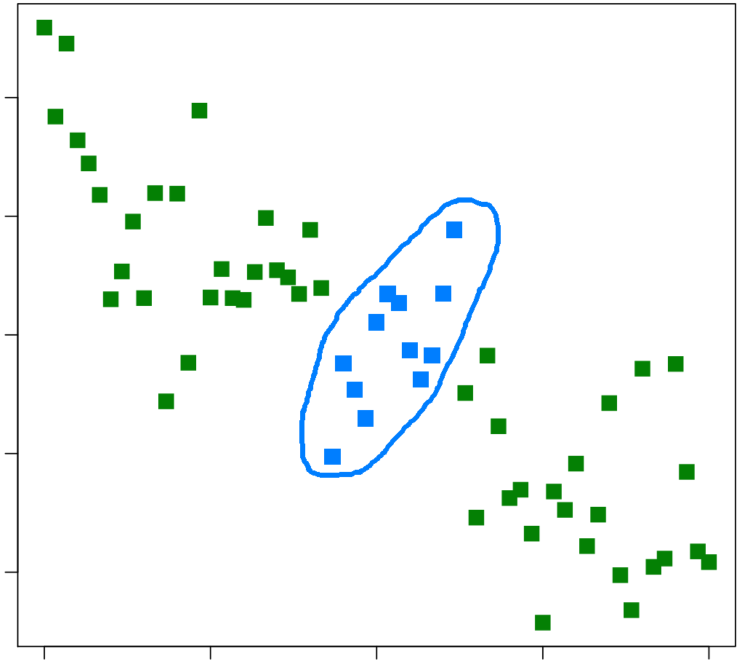
Some time last night, you probably made a couple of decisions. Before talking about what those decisions ended up being, let's take a look at some practical features that probably influenced them:

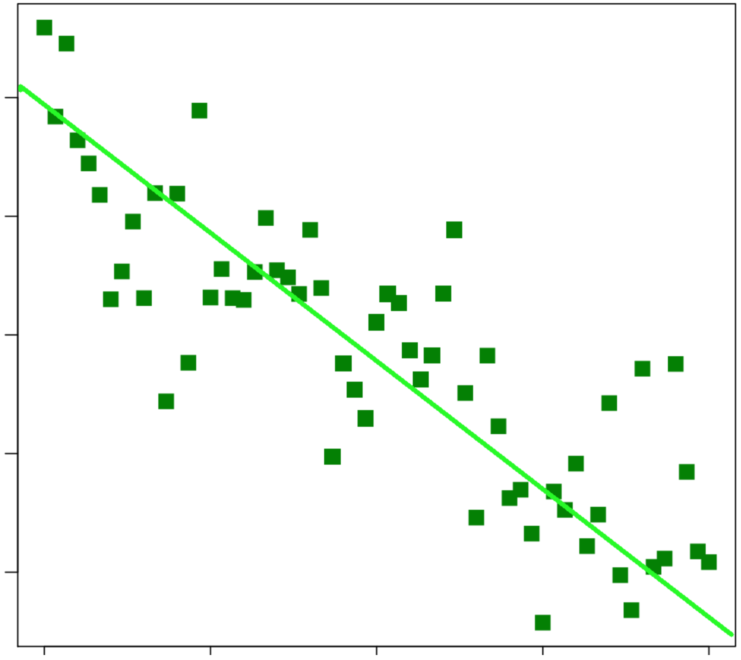
1. Are there better cooks than you in the house? Do you even know how to cook?
2. Do you live near good ranked, and affordable restaurants?
3. How much spending money do have on hand?
4. Are there any decent leftovers in the fridge?
5. How badly do you hate doing the dishes?
6. What ingredients do you have at home?
7. How is the current weather outside?
8. How hungry are you right now?

Armed with the answers to the above features, you are ready to make a few decisions, such as:

- *Should I eat out tonight, or cook at home?*
- *Should I cook something new, or heat leftovers up?*
- *Should I cook tonight, or ask someone else in the house to?*

These questions are all examples of categorical decisions you can calculate with a supervised classification algorithm. Such algorithms derive weights for the contribution each feature has to determining the overall outcome. You can either out in a restaurant or eat at home, but you can't eat out *and* eat at home simultaneously; only a single decision at a time.





The main difference between classification and regression algorithms is that regression aims to compute a continuous output, but the goal of classification is to predict a discrete, categorical output. Using classification, samples get labeled depending on a decision boundary test that separates your data into a range of space. With regression, a continuous value output is calculated from a best fit curve function that runs through your data. In the special case of linear regression, the curve is restricted such that it is linear. Given the features listed above, a regression algorithm would enable you to calculate continuous values like:

- *How far are you willing to drive to eat out?*
- *How much money can you save by cooking at home?*
- *How much time are you willing to invest cooking at home?*

Effectively predicting the future, known as *extrapolating*, or identifying a trend in your existing data, known as *interpolating*, requires there be a statistically significant, linear correlation between your features. Without a decent correlation, linear regression isn't able to benefit to you. Check the Visualization module's Higher Dimensionality 'imshow' section to see methods of discerning correlation.

You may also have heard the phrase, correlation doesn't imply causation. This is an easy trap to fall into when using regression. You can build a linear regression model that fits a relationship between university student's GPA and their first job's annual salary. But simply having a high GPA in doesn't *cause* someone to have a high paying job, although there is probably some significance between the two.