**Video Questions**

Questions start at around 1:07:00

Strategy for solving 1st question – KNN not solve well

Method:

Write down the strengths and weaknesses of each type of learner, especially consider the weaknesses. An obvious strength for linear regression is if the data is linear then it solves it pretty well.

Weaknesses: if data’s not linear, it’s a problem

Similarly with strength and weaknesses of KNN

K doesn’t have to be set to 3. Think about other weaknesses of KNN

For 2nd one, where we want the data to work better with knn, should it work better both in and out sample?

Don’t do the following:

(insert picture)

They have 60%

It’s possible to create a dataset … without loading the test set

Essentially grading your data differently for the 40% than the 60%

You should draw from the same distribution for both 40 and 60%

Linear regression is better in certain places than KNN

If you choose your features correctly, you’re in some extent linearizing the problem

So if your features are particularly non-linear, you can turn your non-linear problem into linear problem by ways of features. (Kernalizing is 1 way)

Project is underspecified to leave room for creativity

**What I’m looking for:**

I’m coming up with some great data that’s going to hurt KNN

It would be a nice to show a picture of what the data is like

And color in and out sample with different dots so that I know they’re drawn from different distribution

Explain in words why this is better for linreg and not good for KNN

Apply the learners. And calculate error and correlation

Maybe baging helps? Maybe not? Try different bags?

Don’t need to try all of them, but want to see creativity on your part.

Can use any visualization tools you like.

You could also grab data off of the web.

Q: Expand on when you can use error vs. correlation?

When is error or correlation a good metric

It’s easy to create a dataset where error is huge. Eg. All data is around 1 million, - small error

But if it’s jumping +/- 50,000 you can have an error of 50,000 but that’s a small percentage of error

Within the same dataset, comparing error is valid

Can’t compare across datasets

In general correlation is a good metric …