Feedback kNN In-Class

Jack Jewson • Jan 26

Dear Students

I just posted some brief feedback to your in-class assignments. Please feel free to respond if you don't understand some of my comments.

In general, I was very happy
+ All of your dataset management skills are excellent
+ Most students showed a good understanding of the kNN model and the steps important to preprocess the data.

Some general feedback and advice
+ Generally we can do better than just removing observations with missing data. As I said in class you can't remove the testing set observations with missing data (although you can remove columns), and it feels strange to remove training observation and impute testing observations. It would be good for you to investigate some methods for imputing the missing data as part of your project
+ When imputing testing data using summaries (like means or medians), you should use the summaries from the training set rather than the testing set. This is exactly analogous to standardisation (which you all did correctly)
+ Unless you have a good, well explained reason, you don't need to create your own training and testing sets. this has been done for you. If you create your own, this often means you don't train the final model on all the data so you are wasting data.
+ Take a look at the pipeline as a way to streamline your data preprocessing.

Good luck for the rest of the project,

Jack

Class comments

Add class comment...