Linh Toan

DSCI 6423: Data Preprocessing

5/10/2024

1. Define the purpose and goals of feature engineering in your own words. (approx. 200 words)

The goals of feature engineering is to discover which features are more important for model performance than others and/or create features to gain more insights. This can be done by simply removing features that seemingly have little to no relevance, or if you want to get more rigorous then you can run statistical tests to see how strongly each feature correlates with each other.

You can also try feature elimination by removing a feature, training the model without it, then do that for every feature and compare model performance. You can create new features by aggregating current features, you can remove features by doing principal component analysis or linear discriminant analysis.

An important step is to encode categorical variables into numbers, using the get\_dummies method, label encoding, DictVectorizer, or you can even manually create a conversion dictionary and apply that to categorical variables.

Another important step is filling in null values. Ideally your client or vendor should give you a full dataset, but in the real world you will have null values and it’s just unavoidable. Easy ways to resolve null values is to replace them with a mean or median value for that feature, but you can also try out more advanced techniques such as Vandermonde interpolation, cubic spline, Newtonian interpolation, Lagrange interpolation, etc.

Very rarely will you be creating the machine learning model from scratch. It takes a lot of computational resources to actually do that, so the vast majority of people using these models will be using ones that are pre-made by a much smarter computer scientist or mathematician. The only thing we can really do is change the inputs, which is why feature engineering is so important to data science.

1. Create an annotated list of the feature engineering strategies that the course author discusses in Sections 9 and 10. Define each strategy in 2-3 sentences and explain when it is an appropriate strategy to choose. (Be comprehensive in listing these.)
2. Reflect on one or more of the strategies that particularly stood out to you. (approx. 300 words)
3. Answer the question: Would it have been helpful to you to have reversed the modules in this course, by finishing this video series before tackling the Superstore Feature Engineering project? (This can be succinct.)

I think in general it’s useful to get an idea of the concepts first, then apply them. So watching a video series before tackling the Superstore Feature Engineering project might be a good order. I find it useful to not only follow along the video, but then have us do the same thing but with a new dataset. That is, the video series does various transformations with one dataset, but then our assignment is to do those transformations on a different dataset. So that way we’re not just copying exactly from the videos.

1. Reflect on the interrelationship between this course (Data Preprocessing) and the course, Applied Machine Learning 1. Do these align well together, or would it be preferable to take one first, and then the other? Which one first? (100-200 words)

I think taking data preprocessing before applied machine learning 1 would work better for the timeline. At the beginning of the courses, it felt very repetitive to do feature engineering and EDA for two courses and I often got confused on which assignment is for which course.

Taking data preprocessing first will do a better job at keeping the two courses distinct, and if you make data preprocessing a pre-requisite for applied machine learning 1, then you can spend less time on feature engineering in applied machine learning 1 too. This would also allow students to do more advanced feature engineering from the start in applied machine learning 1, so students can spend time learning about more models and spend more time on model evaluation.