Feedback SVM In-Class

Jack Jewson • 10:45 AM

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 am really happy with your basic dataset preprocessing skills - everybody appears to have a good understanding of how to minimally standardise, create dummies and deal with missing data - part of your extended assignment could look at more sophisticated ways to do these
+ Most students showed a good understanding of the SVM and the options you have to tune its performance

Some general feedback and advice
+ Many students used the class_weight = 'balanced' option to deal with missing data, but very few students then reweighed their probabilities. Exactly what you should do here is not quite so clear cut as with Logistic regression, as SVM doesn't explicitly build a model for probabilities. What I want to see in the extended version is evidence that you considered reweighing your probabilities, but you are free to then ignore this if it doesn't improve your predictions - you could also play with different methods to balance the data if you like...
+ I gave full marks for quite small Grid searches for the In-Class assignment. For the extended assignment I want to see you have considered a wide rage of possible options, feel free to comment out old GridSearches so I can see what you have taken into account.
+ Unsurprisingly, the best In-Class predictions came when students used the Diagnosis in the model. But you have to be careful, if you blindly turn the diagnosis into dummies you end up with > 7000 columns which will take too long to fit. One option you could consider is a different type of encoding. I believe Laura showed you Binary encoding (which I will explain later in the course) there are other types of encoding - one that works particularly well... (https://contrib.scikit-learn.org/category_encoders/)
+ Another consideration that may improve predictions is finding some way to incorporate the comorbidities extra data set into the main data.frame. Many of you did so for the unsupervised project at the end of last term.

Good luck for the rest of the project, please let me know if you have nay questions.

Jack