Comparion of models

Brooke Anderson

January 29, 2016

Comparison of models for categorical outcomes

Model	Example applications
Logistic regression (LR)	fraud detection in credit card applications
Linear discriminant analysis (LDA)	...
Naive Bayes (NB)	spam filters, other document classification tasks
k-Nearest Neighbors (k-NN)	face recognition
Classification tree	...

Advantages / disadvantages of different methods

Key: + : true for model; 0 : false for model; - : unclear

—	LR	LDA	NB	k-NN	tree
Positives					
Natural fit with categorical predictors	+	0	+	0	+
Natural fit with continuous predictors	+	+	+	+	+
Quick to train	+	+	+	+	+
Quick to predict	+	+	+	0	+
Interpretable coefficients	+	+	0	0	+
Easy to explain results	+	+	+	0	+
Top choice with linear decision boundary	+	+	-	0	0
Top choice with non-linear decision boundary	0	0	-	+	+
Automatically learns interactions	0	0	0	0	+
Easy to do >2 outcome classes	0	0	+	+	0
Works okay with missing values	0	0	0	0	+
Good for determining class probabilities	+	+	+	0	+
Works okay with smaller training datasets	+	+	+	0	0
Does okay handling irrelevant predictors	0	0	+	0	-
Does okay handling lots of predictors	0	0	+	0	0
Easy to update with new training data	0	0	+	+	0
Does okay if training data is not random sample	0	0	0	+	0
Negatives					
Assumes independence in predictors	NA	0	+	+	0
Assumes distribution for continuous predictors	+	+	+	+	0
Requires lots of memory	0	0	0	+	0

"A good first rule for choosing a method is choose one that you understand very well. ... Put an SVM in Vladimir Vapnik's hands, a classification-regression tree in Jerome Friedman's hands or deep learning in Geoffrey Hinton's hands and they're all going to positively destroy the rest of you on Kaggle and have a hard time beating each other very consistently..." - Rick Barber, Quora post

Characteristics of Titanic data

- Classifying between two outcome classes
- Training data is random sample of the population to predict
- Mixture of categorical, ordinal, and continuous predictors
- Missing data in training dataset
- Missing data in testing dataset
- Some categorical variables with many levels (e.g., `Ticket`)
- More observations than predictors
- Continuous variables with non-normal distributions
- Pretty balance between two outcome classes
- One of the categorical variables (`Sex`) seems very important
- We need to predict outcome classes, not class probabilities