CLASSIFICATION MODELS: FURTHER TOPICS

Paul Speaker

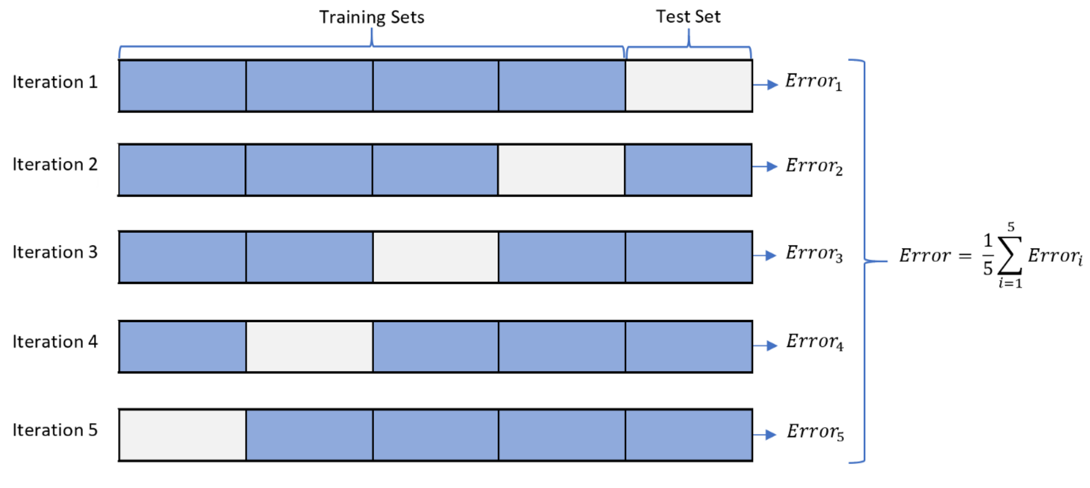
Overview of Classification Model Analysis

- So far we have studied the 3 main groups of classical classification modeling
  - Logistic regression
  - Bayesian methods, including discriminant models
  - knn (nonparametric)

- Today we will look at some additional questions for modeling
  - Cross-validation
  - Abilities of caret package
  - Imbalanced data
  - Multivalued targets

Cross-Validation

- K-Fold Cross-validation is a generalization of the idea of splitting data into train and test datasets
- Basic idea:
  - Split the dataset into K equal pieces
  - Build a model on K – 1 pieces
  - Test the model on last piece
  - (Sometimes) repeat
  - If K = number of data points, this is called Leave One Out CV



Cross-Validation

- For parametric models, the final model's parameters are the average of the individual model parameters
  - Allows for treating all data equally, unlike with train/test split, where parameters are only built on part of data
- For high values of K, computationally very expensive
  - Leave-One-Out only done on small datasets
- Error measure is based only on holdout dataset each time
  - Reduces overfitting

Cross-Validation in R

- You can do the data splitting for cross-validation with a clever use of the sample function in R

- Other packages can do the cross-validation directly

- Even better: packages that do the cross-validation as part of the modeling command

- Enter the caret package

More about the caret package

- We used the caret package for creating Confusion Matrices
- It is actually provides a full suite of modeling capabilities, with a standardized format
- “train” command
  - In most cases models are in other libraries, and caret allow connection to it
  - Can allow for k-fold cross validation
  - Can optimize for other hyperparameters (such as K in knn)
    - “hyperparameter tuning”
    - “grid search”
- No need to split, since caret’s train command will handle do that directly

Other Problems—Imbalanced Datasets

- An imbalanced dataset is a dataset for a classification model where one level of the target variable dominates
  - “dominates” could be 90%, 99% at one level
- Many models will struggle in this setting
  - Bayesian model tend to do better
  - It is generally preferable to process the dataset first
- How to process
  - Differential sampling—build training dataset so that is more balanced
  - Suppose target has 9,000 “yes” points and 1,000 “no” points. We could
    - Undersample the “yes”—train dataset has 80% of no’s (so 800) and only 800 of 9,000 yes’es
    - Oversample the “nos”—train dataset has 80% of yes’es (7,200) and put in the same 800 no’s 9 times each
    - Some will create “synthetic” no’s (Lying with data)

Undersampling and Oversampling

- Undersampling works fine as long as dataset is large enough
  - Are you happy with train dataset which is ~1.5 x the size of your smaller set?
- Oversampling can work ok, but some models struggle more than others with it
  - Creates odd results in knn
    - For k = 9 in the previous example, if smaller class is one of 5 closest, point will be classified as smaller group
      - Lots of false positives
- Bayesian models: if over- or under-sampling is done, priors have to be adjusted to the new proportions
- Another alternative: threshold lowering
  - Rather than having p = 0.5 as decision boundary, adjust so that prediction proportions match original mix
  - To do this properly, you need a 3-way split of data (train, test, validation)

Other Problems—Multivalued Targets

- A multivalued target is a target that has more than 2 levels
- Targets can have 3 or more values
- Most of the methods we worked with can work on multivalued targets
  - Logistic regression cannot directly. 2 alternatives:
    - Multnomial logistic regression
      - Similar structure as logistic, but allowing multiple levels
    - Separate logistic regression for n – 1 of levels
      - Suppose levels are Red, Green, Blue
      - Build logistic regression to predict red/not red
      - Build logistic regression to predict green/not green
      - P(Blue) = 1 – P(Red) – P(Green)
      - For many levels, can run into imbalance problem

Other Problems—Multivalued Targets

- Other models and multivalued targets
  - Bayesian models typically work pretty well
    - More important to have class value-specific statistics (so LDA suffers)
  - K-nearest neighbors is technically possible, but often struggles
    - Especially bad for many-levels targets (if you have 10 neighbors, but there are 5 levels, can be difficult to get support)

- Structure of confusion matrix
  - For multivalued targets, with n possible values, confusion matrix is now n x n matrix