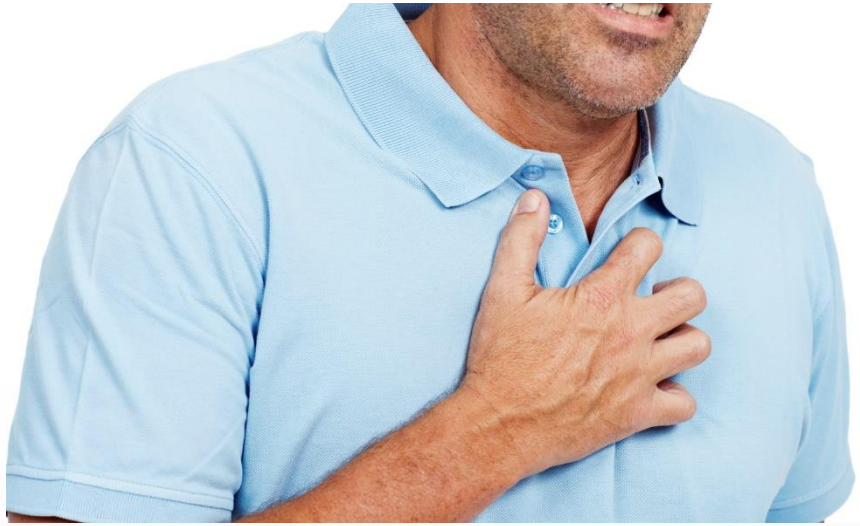
Predicting Death and Other Complications for Myocardial Infarction Patients

Mingxi He
Phillipe Nguyen

Introduction

- Study was done in Quebec to identify the factors causing delays in the treatment of heart attack patients
- Using the gathered data, we will attempt to predict the likelihood of death and other complications for patients
- 50 important variables and around 5000 instances

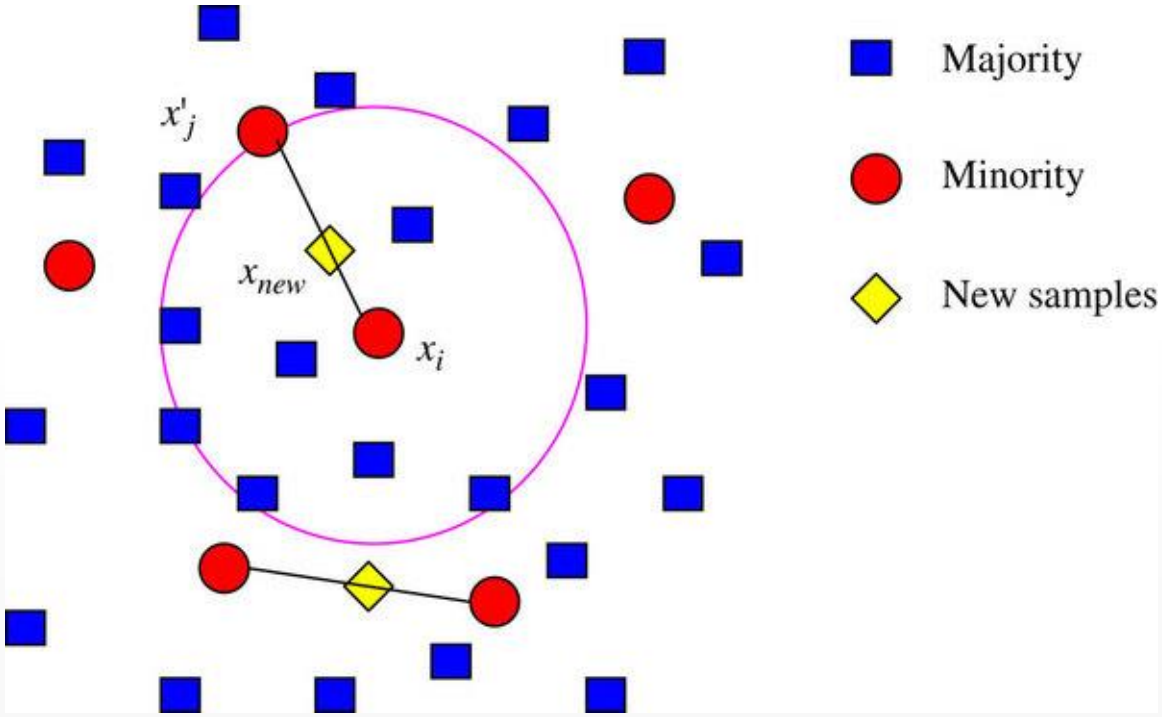


Approach

- Data Cleaning
- k-Fold Cross Validation
- Filling in Missing Data
  - MLE estimates for naïve baye's suggests we use sample mean
  - Training Set vs Testing Set
- Dealing with an under-represented minority class

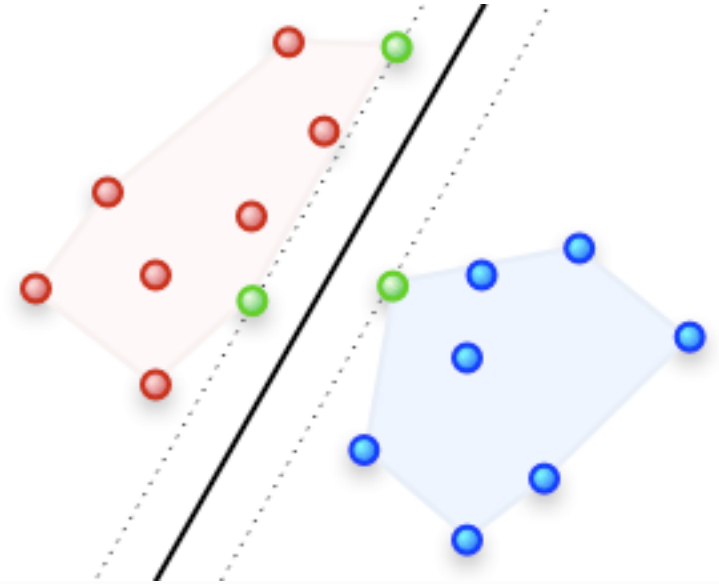
Approach - Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al.)

- For k closest neighbors x_j' of x_i
  - Draw a line between x_j' and x_i
  - Pick a random point x_{new} along that line
  - Add this point as a new instance to the training data set
- k is chosen to generate the desired amount of new data



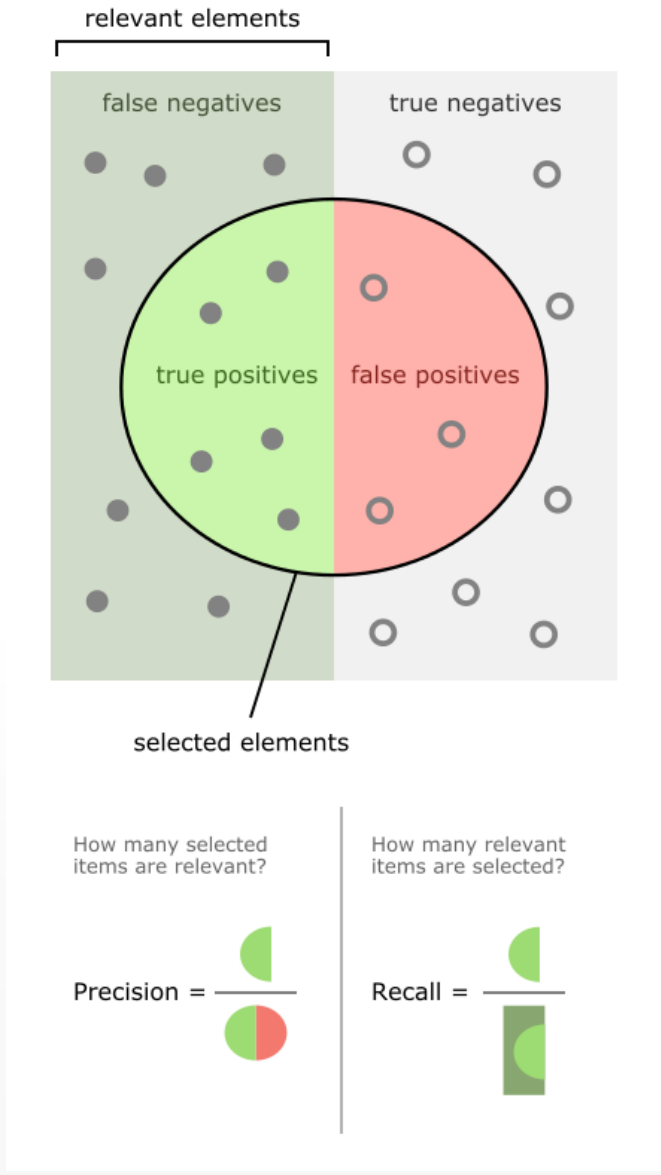
Approach

- Standardize features
- SVM Classifier
  - Attempted different values of C to penalize soft errors and avoid over-fitting
  - Radial basis function as kernel



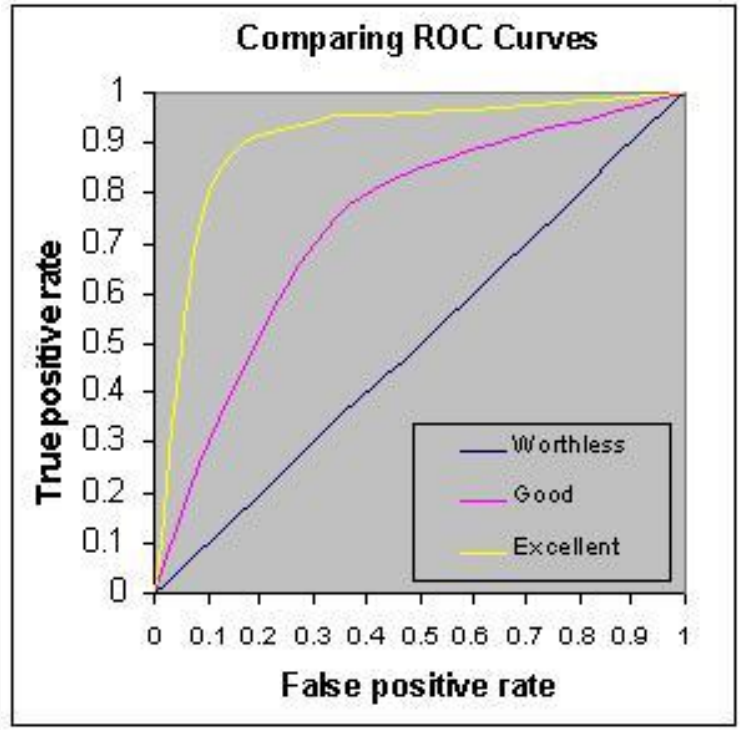
Results – Prediction Accuracy

	Death	Complication
Accuracy	96%	76%
F1 Score	64%	60%
ROC AUC Score	80%	68%



Results – Prediction Accuracy

	Death	Complication
Accuracy	96%	76%
F1 Score	64%	60%
ROC AUC Score	80%	68%

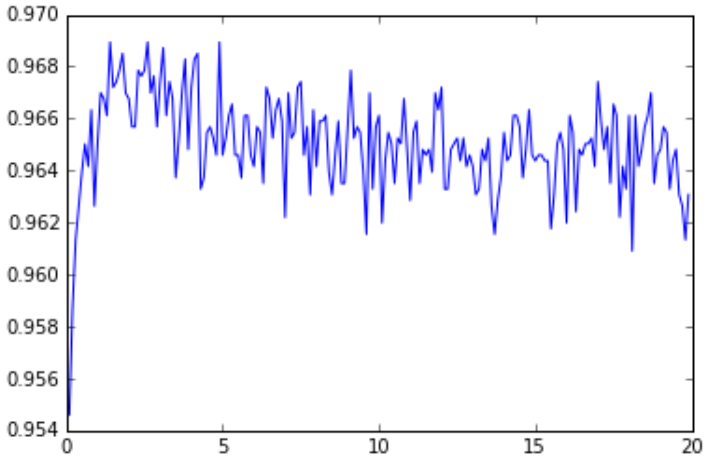


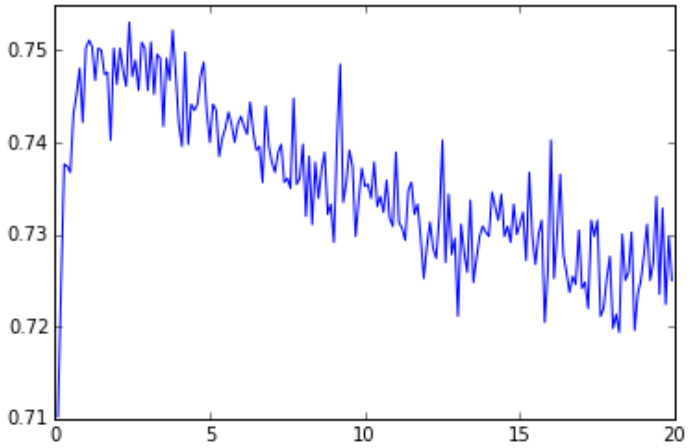
Results – Regularization

Death

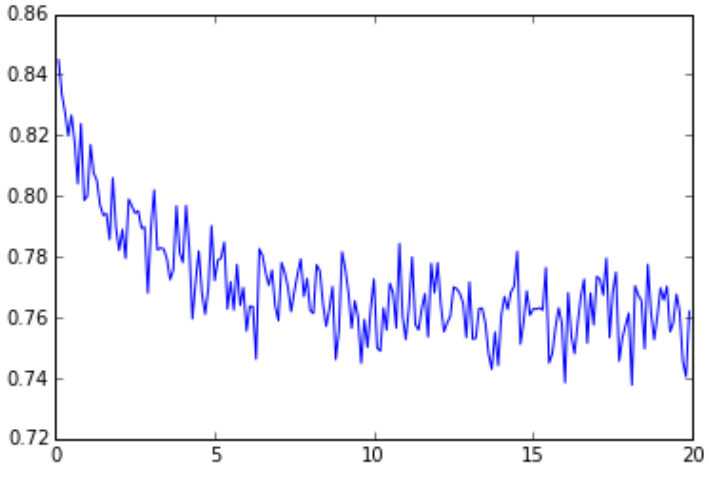
Complications

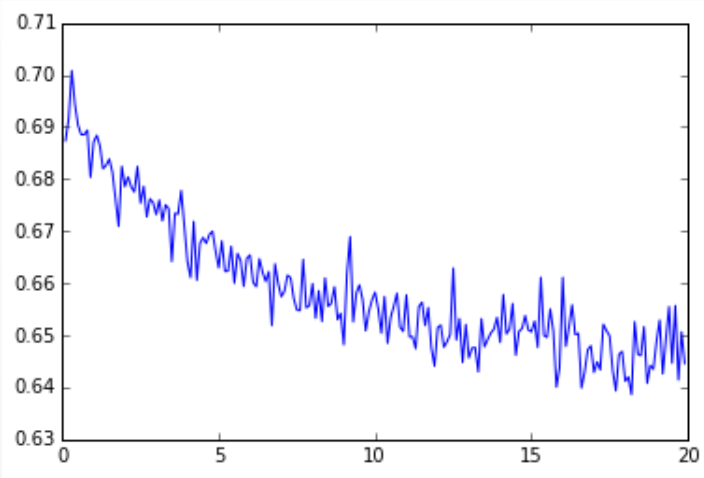
Accuracy





AUC ROC





Conclusions

- All metrics indicate that our classifier works better for “death” than for “complications”
- Accuracy initially increases as C increases, but falls off
- Area Under ROC Curve seems to decrease as C decreases

Moving Forward

- Optimizing over-sampling parameter
- Optimizing over-fitting parameter
- Other Classifiers
- Improve classifier on “complications”
  - PCA

Questions