Data Fusion for Predicting Breast Cancer Survival
Linbailu Jiang, Yufei Zhang, Siyi Peng

Abstract
- In this project, we want to understand a patient's survival rate given his/her genome pattern and time since diagnosis. Our ultimate goal is to predict the survival rate of a patient with breast cancer changing over time based on some related genomic data.
- By training logistic regression models and cox models on each feature separately and ranking them by lowest CV error, we largely reduced the number of features in our models.
- We combined the timeline and status of each event, trained cox models, plotted survival curves for test samples, and then calculated test errors to evaluate our models.

Background
- Cancer is usually considered as one of the most terrifying diseases in our current society:
  As severe as it is lethal in general, there are many different pathways that may affect a patient's survival.
  Age, treatment pathway, genome patterns...
  Is it possible for us to create a model to predict the survival rate of a patient just by analyzing his/her genetic data?

Data & Preprocessing
- Data: The data of our project is from NIH (National Institutes of Health) Project, and we obtained them from Professor Olivier Gevaert.
  The data are all pre-possessed, log-transformed, and well-separated based on cancer type into 11 datasets.

Result & Discussion
- Main Work:
  A. Feature Selection:
  - The sample size of the gene expression dataset for breast cancer is only 985, while the corresponding feature size is 1620, which is much larger than the sample size here.
  - A potential problem that can be caused by this is over-fitting.
  - Method 1: find variance of each feature and get rid of the features with low variance.
  - Method 2: train logistic regression models & cox models on each feature separately and ranking them by lowest CV error.
  - B. Surv Object & Cox Model:
  - Use Surv object to combine the information of both timeline & status of events
  - Fit cox models and analyze the survival curves for test data

Method Overview
- Feature selection (by removing low variance features)
- SVM & Naïve Bayes
- Improved feature selection (ranked by Logistic Regression model)
- A more comprehensive response type (Sun)
- Cox model & survival Curves

Results & Discussion
- Analyze survival curves

Conclusion & Future work
- Conclusion
  Best combination of algorithms:
  Select top logistic/cox-ranked features + treat contact time as part of response (in Surv objects) + Cox model

- Future Work
  Combine gene expression & methylation datasets to fit models

Acknowledgement
Irene (Mentor), Professor Andrew Ng, CS 229 Teaching Staff, Professor Olivier Gevaert, Magali Claire Champion, NIH (National Institutes of Health).