Course Project - CS534 Machine Learning
Fall 2016

Project Objectives

The course project will provide you with practical experience applying machine learning algorithms and principals to analyze real-world data. You will learn how to deal with issues like data normalization, how to evaluate the performance of prediction models, and how to design experiments to compare models and test hypotheses about algorithms. The project also provides you with the opportunity to investigate approaches that are not covered in the textbook.

Predicting Cancer Outcomes from Genomic and Clinical Features

This project will investigate predicting the survival of cancer patients using high-dimensional cancer genomic features and clinical features (age, gender, etc.) from brain (GBMLGG) and breast (BRCA) cancers. Advances in *genomics* are improving our understanding of cancers by providing insights into cancer genetics and the impact of cancer genetic alterations on the molecular circuits that regulate cell functions. Genomics profiles of cancers are *high-dimensional*, often containing many thousands of features. Analyzing and making predictions from genomic data is challenging due to the abundance of features and relatively small numbers of samples ($p \gg N$). The ability to predict patient survival at the time of diagnosis is critical in guiding cancer treatment, and multiple machine-learning based approaches have been developed to address this problem [1, 2, 3].

Requirements

Your project must address the following requirements:

1. **Comparison to an Alternative Approach.** You must compare the performance of your approach to an alternative method. Your primary approach cannot just use existing software - you much perform some original implementation work here. The alternative approach can utilize "off-the-shelf" software.
2. **Validation Procedures.** Your validation approach has to account for variations due to the random assignment of samples to training, testing, and/or validation sets. Model performance must be assessed using *concordance index*.
3. **Model Selection.** You must use model selection techniques to select the parameters of for your method.

Project Teams

You will work in teams of 4-5 people to carry out the literature survey, design, implementation, experiments and writeup. Work should be distributed fairly and equitably. Team members will evaluate each other's contributions as part of the grading process (see Grading Section).

Grading and Deadlines

Project Plan (25%) - due 10/12 A two-page (max) description of topics you plan to investigate, the methods and tools that you plan to use, and a timeline of project milestones.

Report (50%) - due 12/9. A written report including the following sections: Introduction, Methods, Validation Plan, Results, Discussion and Bibliography. Use the fundamental concepts you learned in class to justify your approach and to provide insights into the results. Why did you choose this approach? Why do some models perform better than others, etc. Software should be submitted to the course TA with instructions for compilation and a list of dependencies.

Presentation (25%) A 20-minute team presentation of your project and findings. These will be scheduled for the last two regular class sessions and the final exam period.

Team Reviews - due last class Each person will rate the contributions of their teammates from 1-5. The lowest review you receive will be dropped, and the remaining reviews will be averaged and used to weight your project score. You will receive your anonymized reviews in an email from the instructor. **The instructor is available to mediate team disputes.**

Possible Sub-Topics

What you do beyond the requirements is up to you. Some ideas for topics to investigate:

1. **Regularization.** Methods for constraining models to avoid overfitting the training data.
2. **Transfer Learning.** How combining unrelated datasets can improve prediction performance (additional data available from instructor).
3. **Model Interpretation.** Explaining what features are important in your model.
4. **Neural Networks.** The use of neural networks for survival analysis.
5. **Cost Functions.** Alternatives to Cox likelihood.
6. **Evaluation Metrics.** Alternatives to concordance index for evaluating model performance.

References

[1] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *Ann. Appl. Stat.*, 2(3):841–860, 09 2008.

[2] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(1):1–13, 2011.

[3] Ludger Evers and Claudia-Martina Messow. Sparse kernel methods for high-dimensional survival data. *Bioinformatics*, 24(14):1632–1638, 2008.