Neural Networks and Machine Learning
Applied to Classification of Cancer

Sachin Govind, Advisor: Namrata Pandya, IMSA
Cancer Screening

• Current methods
  – Invasive techniques (biopsy, colonoscopy, etc.)
  – Helical CT
  – Mammography
  – Blood/skin tests

• Dependent on an oncologist’s ability

• New methods – data mining and machine learning
Data Mining in Medicine

- Extraction of useful data from large datasets
- Application of computing technology to medicine
- Often uses gene expression signatures
- Why gene expression?
- R is particularly powerful
Machine Learning = Learning from data

- Raw Data → Scaling and Cleaning
- New data → Develop Model
  - Training Set
  - Testing Set
- Develop Model → Validate
- Validate → Apply Model
Focusing Question
How can machine learning algorithms and methods be effectively adapted to predict genetic predisposition to certain cancers?
Outline of Investigation

1. Obtain raw data and clean
2. Split into disjoint training and testing sets
3. Develop multiple models
4. Test models and assess
5. Gene significance and location analysis
Datasets

• Broad Institute
  – Cancer gene expressions signatures
  – 15 types of cancer
  – 16,000 genes and corresponding expression

• COSMIC
  – Catalogue Of Somatic Mutations in Cancer
  – Gene mutation data
Programming Work

• All work was done in R and in Java
• Data organized in random order to prevent classification bias
• Train : Test ratio is held constant (unless specified) at 7 : 1
• 8 fold cross validation performed on all models
• Average across 8 validation tests is presented
Multinomial Logistic Regression

- Logistic = two classes (this class, or not this class)
- Multinomial = multiple variables affect output
- Assesses each individual and optimizes weight coefficient matrix
- Infeasible for large gene pool

\[ w = (w_0, w_0, w_0, \ldots, w_K) \]

Model Output

\[
 P(y | x) = \frac{\exp(xw_k)}{1 + \sum \exp(xw_i)}
\]
Results of Multinomial Logistic Regression

Accuracy of Multinomial Logistic Regression Model with Varying Size of Genes

Average accuracy over MLR model 15 times with genes randomly selected
Decision Tree

- Goes down list in order
- Binary decision at each split
- Example: dichotomous key
- Determine which characteristics are most important
- Evaluate possibilities at each split and reassess every time
- Allows for multiple outputs
- Computationally inexpensive
Results of Decision Tree
Hidden Neural Network

• Black-box approach
• Mimics neuron interaction
• Voting
• Multiple layers
  – Input layer (senses)
  – Hidden layers (neurons and signaling)
  – Output layer (decision)
• Can vary size of hidden layer or number of layers
Hidden Neural Network Example

Input layer

Hidden Layer, n = 20

Output Layer
Hidden Neural Network Results

Number of Genes versus Accuracy for Varying Size of Hidden Layer

Average Accuracy

Number of Genes (Randomly Selected)

Size = 10  Size = 20  Size = 30
Rooted Random Forests

- Lots of rooted decision trees
- Bootstrapping of features with random selection
- Allows for calculation of importance of genes
- Ran with 1000 random rooted trees
Rooted Random Forest Results
Self-Organizing Map

- Dimensionality reduction
- Visualization of interactions between neurons
- Unsupervised (non-classification)
- Produces a low-dimension representation of the data
Self-Organizing Map Results
Extra: Mutation Location Analysis

- COSMIC Database
- Analyzed the prevalence of tumor-related mutations
- Sample size > 10^6
- Computed probability of tumorigenic mutation at given locations
- Drawback: takes significant amount of time to compute
Example: Chromosome 17

- Centromere
- Telomere Shortening
- BRCA1 Gene
Key Conclusions

• Methods:
  – MLR: Infeasible but accurate
  – Decision tree: mediocre predictor but quick to construct
  – Random Forest – best predictor
  – SOM – gives good insight into gene interactions
  – Hidden Neural Network – best predictor

• Analysis of gene location can prove useful
References

- http://www.inside-r.org/packages/cran/randomforest/docs/randomforest
Acknowledgements

• SIR Staff for providing this opportunity to me for the fourth and final time
• Ms. Pandya for offering guidance and support in my investigation
• The internet for helping me with my troublesome code at any hour of the night