# Machine Learning

**Benjamin Mayne**

OCEANS AND ATMOSPHERE
www.csiro.au

# Feature Selection

| | Control_1 | Control_2 | Control_3 | Control_4 | Control_5 | Disease_1 | Disease_2 | Disease_3 | Disease_4 | Disease_5 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Gene_1** | 6 | 10 | 0 | 4 | 9 | 16 | 34 | 15 | 7 | 11 |
| **Gene_2** | 15 | 19 | 33 | 23 | 24 | 17 | 25 | 19 | 19 | 33 |
| **Gene_3** | 33 | 11 | 6 | 29 | 13 | 19 | 22 | 13 | 33 | 21 |
| **Gene_4** | 11 | 2 | 33 | 10 | 14 | 9 | 11 | 12 | 10 | 19 |
| **Gene_5** | 24 | 12 | 18 | 28 | 24 | 32 | 28 | 22 | 28 | 30 |
| **Gene_6** | 28 | 11 | 0 | 23 | 8 | 22 | 21 | 22 | 18 | 13 |
| **Gene_7** | 8 | 19 | 12 | 26 | 18 | 23 | 34 | 4 | 23 | 23 |
| **Gene_8** | 26 | 6 | 4 | 16 | 24 | 20 | 7 | 0 | 32 | 35 |
| **Gene_9** | 17 | 1 | 22 | 17 | 12 | 21 | 28 | 11 | 35 | 26 |
| **Gene_N** | 17 | 18 | 20 | 20 | 27 | 13 | 24 | 20 | 16 | 13 |

Feature Selection

Identify highly variable predictors between groups

Recursive Feature Elimination (RFE)

Correlations

Differential Expression

# Data Splitting

| | Control_1 | Control_2 | Control_3 | Control_4 | Control_5 | Disease_1 | Disease_2 | Disease_3 | Disease_4 | Disease_5 |
|---|---|---|---|---|---|---|---|---|---|---|
| Gene_1 | 6 | 10 | 0 | 4 | 9 | 16 | 34 | 15 | 7 | 11 |
| Gene_2 | 15 | 19 | 33 | 23 | 24 | 17 | 25 | 19 | 19 | 33 |
| Gene_3 | 33 | 11 | 6 | 29 | 13 | 19 | 22 | 13 | 33 | 21 |
| Gene_4 | 11 | 2 | 33 | 10 | 14 | 9 | 11 | 12 | 10 | 19 |
| Gene_5 | 24 | 12 | 18 | 28 | 24 | 32 | 28 | 22 | 28 | 30 |
| Gene_6 | 28 | 11 | 0 | 23 | 8 | 22 | 21 | 22 | 18 | 13 |
| Gene_7 | 8 | 19 | 12 | 26 | 18 | 23 | 34 | 4 | 23 | 23 |
| Gene_8 | 26 | 6 | 4 | 16 | 24 | 20 | 7 | 0 | 32 | 35 |
| Gene_9 | 17 | 1 | 22 | 17 | 12 | 21 | 28 | 11 | 35 | 26 |
| Gene_N | 17 | 18 | 20 | 20 | 27 | 13 | 24 | 20 | 16 | 13 |

Split data

| | Control_1 | Control_2 | Disease_1 | Disease_2 | Disease_5 |
|---|---|---|---|---|---|
| Gene_1 | 6 | 10 | 16 | 34 | 11 |
| Gene_2 | 15 | 19 | 17 | 25 | 33 |
| Gene_3 | 33 | 11 | 19 | 22 | 21 |
| Gene_4 | 11 | 2 | 9 | 11 | 19 |
| Gene_5 | 24 | 12 | 32 | 28 | 30 |
| Gene_6 | 28 | 11 | 22 | 21 | 13 |
| Gene_7 | 8 | 19 | 23 | 34 | 23 |
| Gene_8 | 26 | 6 | 20 | 7 | 35 |
| Gene_9 | 17 | 1 | 21 | 28 | 26 |
| Gene_N | 17 | 18 | 13 | 24 | 13 |

Training Data
70%

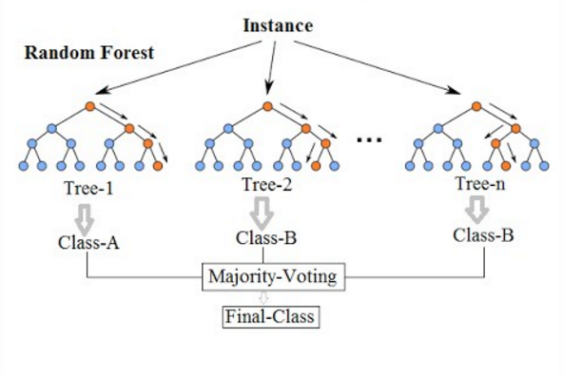
| | Control_4 | Control_5 | Disease_3 | Disease_4 |
|---|---|---|---|---|
| Gene_1 | 4 | 9 | 15 | 7 |
| Gene_2 | 23 | 24 | 19 | 19 |
| Gene_3 | 29 | 13 | 13 | 33 |
| Gene_4 | 10 | 14 | 12 | 10 |
| Gene_5 | 28 | 24 | 22 | 28 |
| Gene_6 | 23 | 8 | 22 | 18 |
| Gene_7 | 26 | 18 | 4 | 23 |
| Gene_8 | 16 | 24 | 0 | 32 |
| Gene_9 | 17 | 12 | 11 | 35 |
| Gene_N | 20 | 27 | 20 | 16 |

Testing Data
30%

Important to keep samples evenly distributed based on other outcomes

# Model Training

- Random Forest




- Regression Models

- Elastic net regression model

$$y = Predictor1\chi + Predictor2\chi + Predictor3\chi + c$$

- Cross Validation
- 10 Fold, often in-built into functions

# Model Testing

- Confusion Matrix and ROC Curve

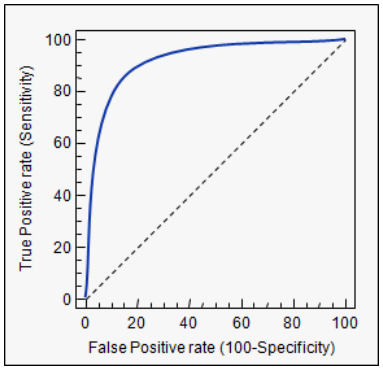
**Confusion Matrix and ROC Curve**

| | | Predicted Class | |
|---|---|---|---|
| | | No | Yes |
| Observed Class | No | TN | FP |
| | Yes | FN | TP |

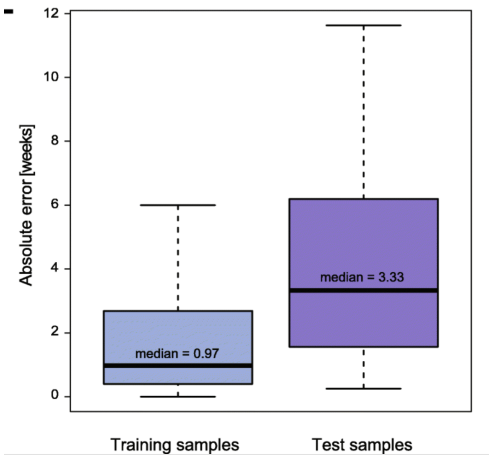
| | |
|---|---|
| TN | True Negative |
| FP | False Positive |
| FN | False Negative |
| TP | True Positive |

**Model Performance**

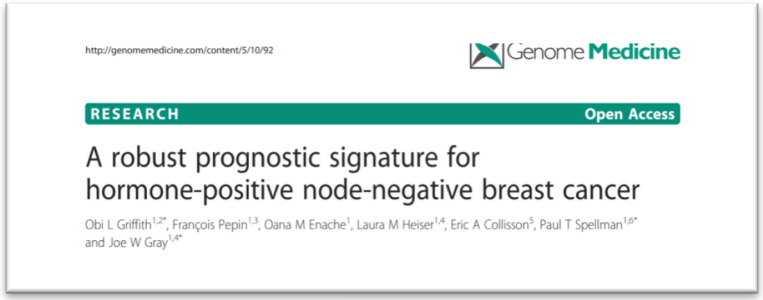
| | |
|---|---|
| Accuracy | = (TN+TP)/(TN+FP+FN+TP) |
| Precision | = TP/(FP+TP) |
| Sensitivity | = TP/(TP+FN) |
| Specificity | = TN/(TN+FP) |




- Median Absolute Error



# Tutorials




- Machine Learning For Cancer Classification

(https://www.biostars.org/p/85124/)


- Additional File 2, page 14

(https://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-10-r115)

# R Packages

- caret (Classification And REgression Training)

http://topepo.github.io/caret/index.html

- Elastic Net Regression model

https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html

- pROC

# Thank you