Module 2: Supervised learning

Split data into train and test

```
dim(train_x)
```

```
## [1] 106 587
```

```
length(train_y)
```

```
## [1] 106
```

```
dim(test_x)
```

```
## [1]  75 587
```

```
length(test_y)
```

```
## [1] 75
```

LASSO logistic regression

```r
# Choose best lambda using CV.
beta_lasso <- lasso_fit(
  x = train_x,
  y = train_y,
  tuning = "cv",
  family = "binomial"
)
```

```r
# Features Selected.
names(beta_lasso[abs(beta_lasso) > 0])[-1]
```

```
## [1] "NLP93"                  "NLP104"                 "NLP304"
## [4] "main_NLP"               "healthcare_utilization"
```

ALASSO logistic regression

```r
# Fit Adaptive LASSO.
beta_alasso <- adaptive_lasso_fit(
  x = train_x,
  y = train_y,
  tuning = "cv",
  family = "binomial"
)
```

```r
# ALASSO features selected.
beta_alasso[!beta_alasso == 0][-1]
```

```
##                 NLP304               main_NLP healthcare_utilization 
##             -1.0587198              1.2149864             -0.4982002
```

```r
# LASSO features selected.
beta_lasso[!beta_lasso == 0][-1]
```

```
##                  NLP93                 NLP104                 NLP304 
##            -0.01111698            -0.03685247            -0.36065965 
##               main_NLP healthcare_utilization 
##             0.68241957            -0.16359373
```

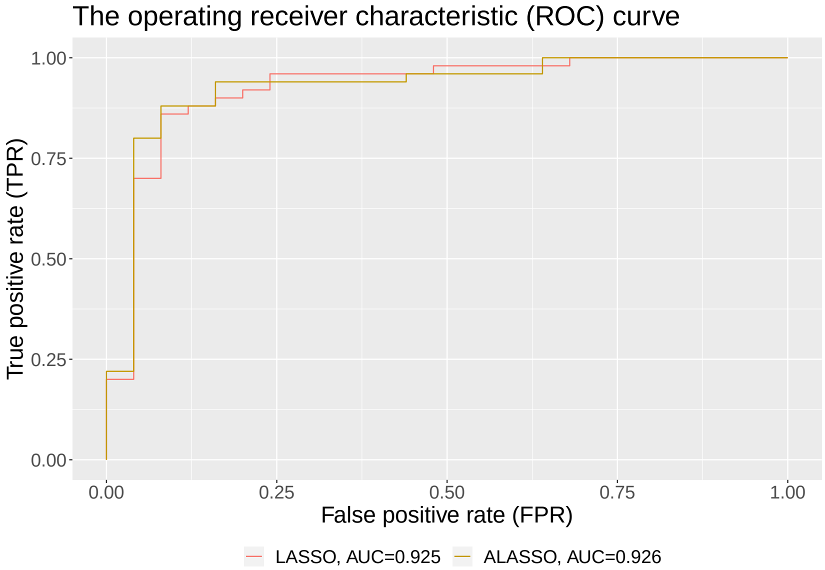
Get model predictions + ROC curve

```r
# Prediction on testing set (LASSO).
y_hat_lasso <- linear_model_predict(
  beta = beta_lasso,
  x = test_x,
  probability = TRUE
)
```

```r
# Prediction on testing set (ALASSO).
y_hat_alasso <- linear_model_predict(
  beta = beta_alasso,
  x = test_x,
  probability = TRUE
)
```

```r
roc_lasso <- roc(test_y, y_hat_lasso)
roc_alasso <- roc(test_y, y_hat_alasso)
```

LASSO vs. ALASSO



LASSO vs. ALASSO at FPR = 0.10

```r
roc_full_lasso <- get_roc(y_true = test_y, y_score = y_hat_lasso) %>% data.frame()
get_roc_parameter(0.1, roc_full_lasso)
```

```
##      cutoff  pos.rate FPR  TPR       PPV       NPV        F1
## 1 0.6637589 0.6066667 0.1 0.86 0.9450549 0.7627119 0.9005236
```

```r
roc_full_alasso <- get_roc(y_true = test_y, y_score = y_hat_alasso) %>% data.frame()
get_roc_parameter(0.1, roc_full_alasso)
```

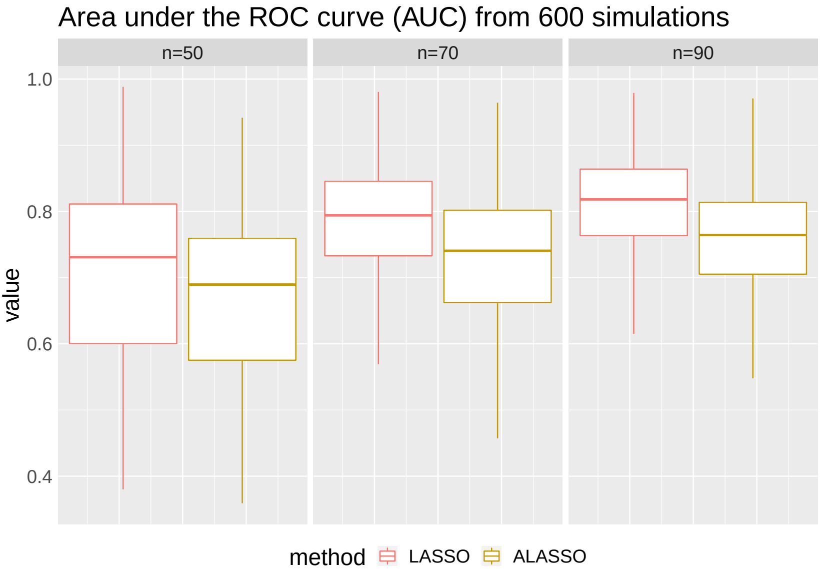
```
##      cutoff pos.rate FPR  TPR       PPV       NPV        F1
## 1 0.7120506     0.62 0.1 0.88 0.9462366 0.7894737 0.9119171
```

LASSO vs. ALASSO with different training set size

- Randomly sample training size = 50, 70, 90
- Use the remaining data as the test set
- Repeat 600 times

```r
auc_supervised <- validate_supervised(
  dat = labeled_data,
  nsim = 600,
  ntrain = c(50, 70, 90)
)
```

LASSO vs. ALASSO with different training set size

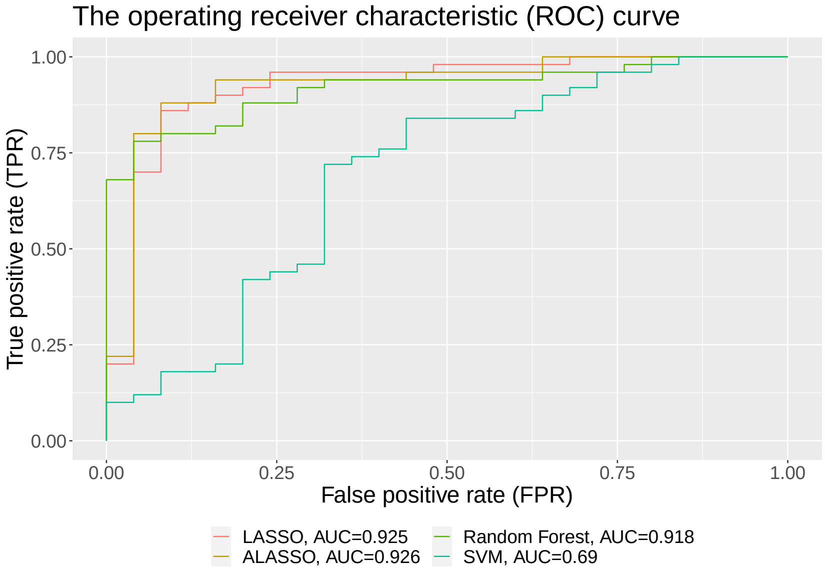


Random Forest and SVM

```r
# Random forest.
model_rf <- rfsrc(y ~ ., data = data.frame(y = train_y, x = train_x))
y_hat_rf <- predict(model_rf, newdata = data.frame(x = test_x))$predicted
roc_rf <- roc(test_y, y_hat_rf)
```

```r
# SVM.
model_svm <- SVMMaj::svmmaj(X = train_x, y = train_y)
y_hat_svm <- predict(model_svm, test_x)
roc_svm <- roc(test_y, y_hat_svm)
```

ROC curves


The operating receiver characteristic (ROC) curve
True positive rate (TPR)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
False positive rate (FPR)
LASSO, AUC=0.925
ALASSO, AUC=0.926
Random Forest, AUC=0.918
SVM, AUC=0.69