COMP 755 Final Project

Kevin O'Connor

11/8/2018

In this document, we apply three different classification methods on The Cancer Genome Atlas (TCGA) Breast Cancer data. Specifically, we will use gene expression levels to predict breast cancer subtype which falls into one of five different groups: Luminal A, Luminal B, Basal, Normal, and HER2.

After reading in the data and removing genes with extremely small or large variance, we are left with $p=$4101 variables and sample sizes for each group given by

```
kable(data.frame(group=c("Luminal A", "Luminal B", "Basal", "Normal", "HER2"), sample.size=c(ncol(dat.lu
```

group	sample.size
Luminal A	572
Luminal B	219
Basal	191
Normal	137
HER2	82

Yielding a total sample size of $n=$1201. Note that we have normalized the row of each dataset.

Multinomial Logistic Regression

First we will perform logistic regression to try and predict the subgroup which a given set of gene expression values belongs to. We can write our model as

$$\mathbb{P}(y_i = k | X_i, \beta) = \frac{\exp\{-X_i\beta_k\}}{\sum_{k'} \exp\{-X_i\beta_{k'}\}}$$

Which gives a negative log-likelihood of

$$\mathcal{NLL}(\beta) = \sum_{i=1}^{n} -X_i\beta_k - \sum_{i=1}^{n} \log\left(\sum_{k'} \exp\{-X_i\beta_{k'}\}\right)$$

```
neg_log_likelihood <- function(beta, X, y){
  # Returns negative log-likelihood of the logistic regression model.
  #
  # Args:
  #  -beta: p by K matrix where the k'th column gives beta_k as in the equation
  #    above.
  #  -X: p by n data matrix.
  #  -y: n by 1 vector giving the labels of the data.
  # Returns:
  #  A number corresponding to the negative log-likelihood.
}

get_log_likelihood_w_pen <- function(beta, X, y, pen_type="l1", lambda){
  # Returns penalized log-likelihood of the logistic regression model.
```

```r
  #
  # Args:
  #  -beta: p by K matrix where the k'th column gives beta_k as in the equation
  #    above.
  #  -X: p by n data matrix.
  #  -y: n by 1 vector giving the labels of the data.
  #  -pen_type: Either "l1" or "l2" corresponding to the type of penalty term.
  #  -lambda: Penalty parameter.
  # Returns:
  #  A number corresponding to the penalized negative log-likelihood.

  nll <- neg_log_likelihood(beta, X, y)
  if(pen_type == "l1"){
    pen <- lambda*sum(abs(beta))
  } else if(pen_type == "l2"){
    pen <- lambda*sum(beta^2)
  } else {
    print("ERROR: Unknown penalty type!")
    return(NaN)
  }

  return(nll + pen)
}

# TODO: A test (?)
```