Final Project

Evan Ott

December 5, 2016

Stochastic Gradient Descent (Logistic Regression)

Let's load in the data and Dr. Scott's algorithm:

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

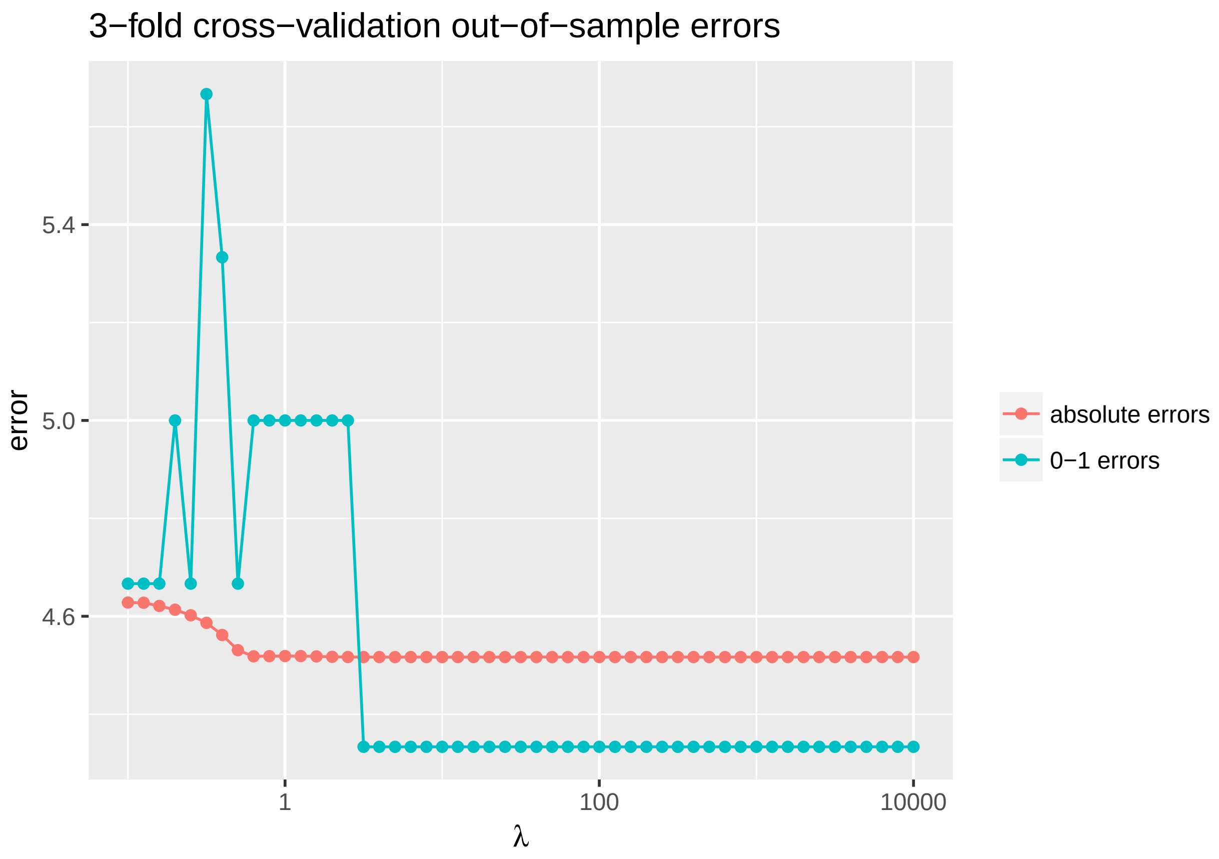
```r
library(latex2exp)
library(Matrix)
library(Rcpp)
library(RcppEigen)
library(reshape2)

Rcpp::sourceCpp("scott_sgdlogit.cpp")

raw_data = read.csv("master.csv", header = TRUE)

# Read specific columns from the data

kegg = as.character(raw_data[ , 1])
gene = as.character(raw_data[ , 2])

data = raw_data[ , 3:ncol(raw_data)]

N = ncol(data)
P = nrow(data)
M = rep(1, N) # Used in logistic regression

# For C++ code, it's easiest to use a sparse matrix
# TODO: might want to consider scaling the counts, either in the way DESeq2 does
# it or something similar.
X = Matrix(as.matrix(unname(data)), sparse = TRUE)

# X[i, j] == data[i, j] is non-negative, so this trims out the genes with no counts
X_nozero = X[which(rowSums(data) > 0), ]

# Convert the column names to 0 = control, 1 = treatment
Ynames = names(data)
Y = stringi::stri_endswith(Ynames, fixed = "T") * 1
```

Now, let's create a transformed version of X where the values are on the log scale.

```r
# Copy it as to not damage the original
X_nozero_log = X_nozero
# Oooh I figured out a sweet hack. So the dgCMatrix class could be used in the
# following way:
# X@i are the row indices - 1 (that have ncol sections in non-decreasing order)
# X@p are the indices - 1 in X@i where we switch to a new column
#
# OR
# X@x is the (non-zero) data itself, so why not just edit that? =D
# Here, add a small value because log(1) == 0, and we want to maintain a value there.
# #oneliner
X_nozero_log@x = log2(X_nozero_log@x + 0.1)
```

Now, let's try to use cross-validation (on the limited number of samples we have) to determine a good range for λ, the LASSO penalty factor in the stochastic gradient decent logistic regression code.

```r
# Initial guess for the betas.
beta_init = rep(0, P)
# For the weighting in the skip scenario for a feature
# TODO: algorithm is very sensitive to this value
eta = 0.0002
# passes through the data
nIter = 100
# exponentially-weighted moving average factor
# TODO: could explore other values.
discount = 0.01

# X_cv = X_nozero
X_cv = X_nozero_log

folds = 3
col_breaks = cut(1:length(Y), folds, labels = FALSE)
abs_errors = c()
zero_one_errors = c()
lambdas = 10 ^ seq(-1, 4, 0.1)
for (lambda in lambdas) {
  absFoldErrors = c()
  zeroOneFoldErrors = c()
  for (fold in 1:folds) {
    test_ind = which(col_breaks == fold)
    train_ind = which(col_breaks != fold)
    X_nozero_train = X_cv[ , train_ind]
    Y_train = Y[train_ind]
    X_nozero_test = X_cv[ , test_ind]
    Y_test = Y[test_ind]

    train_result = sparsesgd_logit(X_nozero_train, Y_train, M[train_ind], eta, nIter, beta_init, lambda

    intercept = train_result$alpha
    beta = train_result$beta
    prediction_test = 1/(1 + exp(-(intercept + t(X_nozero_test) %*% beta)))

    # metrics to choose from
    zero_one_error_test = sum(Y_test != round(prediction_test))
    abs_error_test = sum(abs(Y_test - prediction_test))

    absFoldErrors = c(absFoldErrors, abs_error_test)
    zeroOneFoldErrors = c(zeroOneFoldErrors, zero_one_error_test)
  }
  abs_errors = c(abs_errors, mean(absFoldErrors))
  zero_one_errors = c(zero_one_errors, mean(zeroOneFoldErrors))
}
```

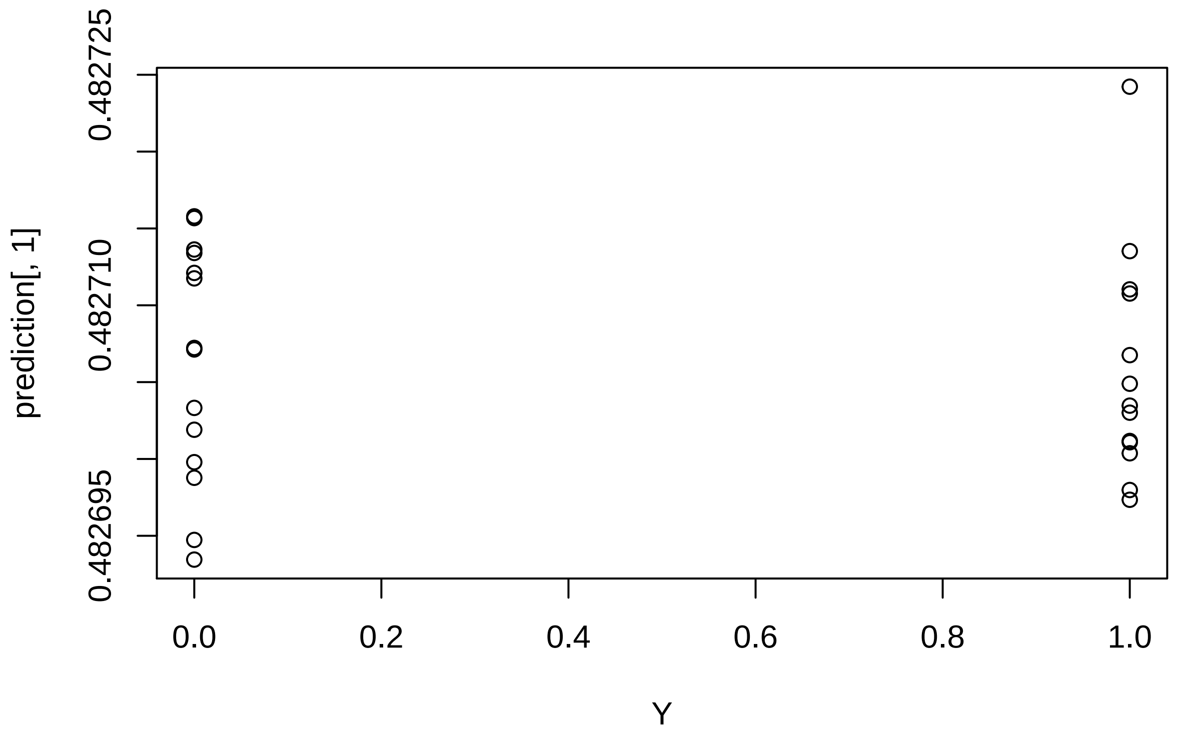


So based on 3-fold cross-validation, the absolute error $\sum_i |y_i - \hat{y}_i|$ seems to reach its best (minimal) value when $\lambda = 10^{-0.2}$ while the 0-1 error $\sum_i \delta(y_i, \text{ round}(\hat{y}_i))$ (number of incorrect predictions) exhibits the same behavior around $\lambda = 10^{0.5}$.

Let's take the latter, since the absolute error shouldn't be different.

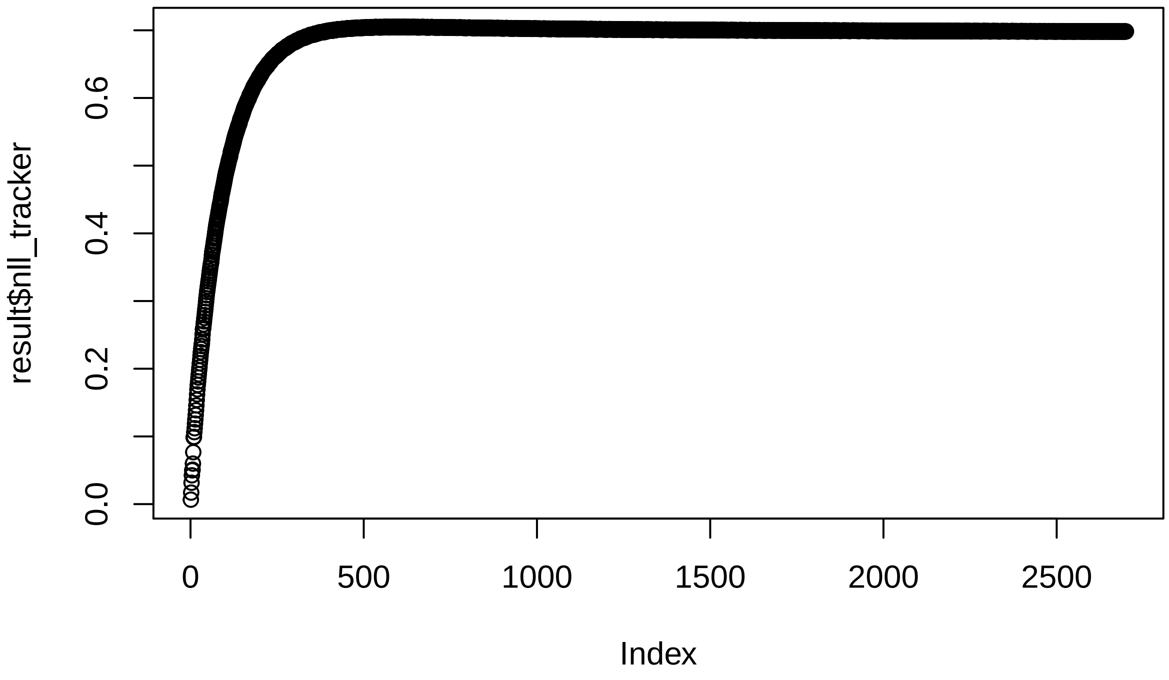
```
# Initial guess for the betas.
beta_init = rep(0, P)
# For the weighting in the skip scenario for a feature
# TODO: algorithm is very sensitive to this value
eta = 0.0002
# passes through the data
nIter = 100
# L1 regularization
lambda = 10^0.5
# exponentially-weighted moving average factor
# TODO: could explore other values.
discount = 0.01

# Run the algorithm
result = sparsesgd_logit(X_nozero_log, Y, M, eta, nIter, beta_init, lambda, discount)
intercept = result$alpha
beta = result$beta
prediction = 1/(1 + exp(-(intercept + t(X_nozero_log) %*% beta)))
```

```
plot(prediction[ , 1] ~ Y)
```



So, these predictions are, to use a technical term, "totally garbage." Every single prediction is just slightly below $\frac{1}{2}$ so the algorithm is effectively saying, "I dunno, but maybe this data is control not treatment" so while the results below are potentially a starting point, it'd be rather foolish to grant them much authority.

```
## How many genes are in the final model?  25
## Which genes?
## adef_c_5_1802
## aot180_c_13_1642
## bdim_c_51_1620
## belk_c_262_3655
## belk_c_623_6138
## blon2249_c_1_1767
## ecas2549_c_2_485
## ecas2549_c_2_487
## ecas2549_c_2_492
## ecas2549_c_3_766
## efae1080_c_1_429
## gmor_c_6_936
## lfus_c_18_3454
## mneo_c_1_1955
## mneo_c_1_4619
## NCBIABWK_c_4_1154
## NCBIACFE_c_1_366
## NCBIACFU_c_47_2923
## not020_c_81_1706
## nsic_c_7_1124
## pmir_c_10_1585
## pot302_c_4_620
## psal_c_109_2279
## ptan_c_1_266
## rpic_c_1_2351
```



DESeq2

Now, let's use a pre-fabricated solution instead.

```r
library(DESeq2)
#library(BiocParallel)
# Allow for parallelization of DESeq2 code.
#register(MulticoreParam(4))

# Get the counts
count_data_raw = raw_data[ , 3:ncol(raw_data)]

########################################
#  Format the data in the way DESeq2 expects
########################################
new_colnames = rep(NULL, N)
condition = rep(NULL, N)
control = 0
treatment = 0
for (i in 1:N) {
  base = c("control", "treatment")[Y[i] + 1]
  val = 0
  if (base == "control") {
    control = control + 1
    val = control
  } else {
    treatment = treatment + 1
    val = treatment
  }
  new_colnames[i] = paste0(base, val)
  condition[i] = base
}
count_data = count_data_raw
colnames(count_data) = new_colnames
rownames(count_data) = gene

col_data = data.frame(condition)
rownames(col_data) = new_colnames

# remove the unobserved genes from the get-go
count_data = count_data[which(rowSums(data) > 0), ]

# Run DESeq2
dds = DESeqDataSetFromMatrix(countData = count_data,
                             colData = col_data,
                             design = ~ condition)
dds = DESeq(dds)#, parallel = TRUE)
res = results(dds)#, parallel = TRUE)

# plot(sort(res$pvalue))
# points(sort(res$padj), col = "red")
# P - sum(is.na(res$pvalue))
# P - sum(is.na(res$padj))
```

```
## How many genes are in the final model?  61
```

```
## Which genes in default DESeq2?
## abau_c_1_46
## ager2575_c_12_2315
## aisr_c_34_2576
## anae_c_1_1654
## anae_c_1_728
## aomn_c_8_885
## aori_c_393_2362
## aot178_c_12_438
## aot448_c_14_1554
## blon_c_1_69
## bper_c_1_3775
## bpyo_c_2_358
## chom_c_30_1496
## cot332_c_4_252
## cot412_c_6_689
## cure_c_1_1750
## ecol1511_c_1_2410
## ehor_c_1_170
## ehor_c_52_4399
## elim_c_1_3325
## haeg_c_4_726
## hpar1239_c_1_1267
## hpyl2409_c_1_1167
## kpne1249_c_1_4568
## ksed_c_1_1420
## line_c_2_573
## llac1719_c_1_2007
## llac1719_c_1_2586
## lmon_c_1_2434
## lot107_c_1_104
## lot225_c_32_1105
## lpar_c_57_2568
## mot186_c_3_1752
## NCBIABZW_c_1_2083
## NCBIACDZ_c_2_749
## nsic_c_2_418
## oant1457_c_1_1135
## pasa2160_c_1_1685
## pbue_c_15_675
## pflu_c_1_2697
## pflu_c_1_3543
## pflu_c_1_3638
## pflu_c_1_3948
## pflu_c_1_4139
## pflu_c_1_4394
## pflu1300_c_1_2147
## pgin_c_1_669
## pgin33277_c_1_480
## pori2683_c_4_744
## pori2683_c_7_1151
## pot278_c_8_458
## pot317_c_73_2966
## pot472_c_11_799
## pot472_c_2_167
## pot472_c_26_1409
## pot472_c_28_1476
## pot472_c_39_1762
## ppal_c_18_1062
## ppis_c_23_1295
## psac2685_c_5_1034
## pver2511_c_3_830
```

```
# Not exactly sure what this plot does either.
# plotMA(res, main="DESeq2", ylim=c(-3,3))

# Benjamini-Hochberg by hand (DESeq2 does something similar...)
alpha = 0.1
sorted_pval = sort(res$pvalue, na.last = TRUE)
numNA = sum(is.na(res$pvalue))
bh = sapply(1:P, function(i){ sorted_pval[i] <= alpha * i / (P - numNA)})
threshold = sorted_pval[max(which(bh))]
```

```
## How many genes in customized Benjamini-Hochberg procedure
## 1
```

```
## Which genes in customized Benjamini-Hochberg procedure
## pflu_c_1_4139
```

```
# These two are equal -- throws out genes that were never observed section 1.5.3 of DESeq2 paper
print(paste(sum(rowSums(count_data) == 0),
            sum(is.na(res$pvalue))))
```

```
## [1] "0 0"
```

```r
# 118687 genes are thrown out by the "independent filtering" for having a low
# mean normalized count
sum(is.na(res$padj)) - sum(is.na(res$pvalue))
```

```
## [1] 118687
```

```r
# Not entirely sure what this plot means.
# plot(metadata(res)$filterNumRej,
#      type = "b", ylab = "number of rejections",
#      xlab = "quantiles of filter")
# lines(metadata(res)$lo.fit, col = "red")
# abline(v = metadata(res)$filterTheta)

#nofilter
# If you turn off the filtering, only two of the adjusted p-values are less than 0.1
# where the adjustment basically just accounts for Benjamini-Hochberg, I think
resNoFilt <- results(dds, independentFiltering = FALSE)
addmargins(table(filtering = (res$padj < .1),
                 noFiltering = (resNoFilt$padj < .1)))
```

```
##          noFiltering
## filtering FALSE  TRUE   Sum
##     FALSE 33700     0 33700
##     TRUE     59     2    61
##     Sum   33759     2 33761
```

```
## How many genes without independent filtering
## 2
```

```
## How many genes without independent filtering
## cot412_c_6_689
## pflu_c_1_4139
```