STAT 440 - Quiz 2

Handi Gao

2018-01-30

Data and Model

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset consists of 30 features of cell nuclei extracted from 569 digitized images of benign and malignant breast tumors.

```
tumor <- read.csv("wdbc.csv")
dim(tumor)
```

```
## [1] 569  32
```

```
colnames(tumor)
```

```
##  [1] "id"         "diag"       "radM"       "textM"      "perimM"
##  [6] "areaM"      "smoothM"    "compactM"   "concM"      "cptsM"
## [11] "symM"       "fracM"      "radSE"      "textSE"     "perimSE"
## [16] "areaSE"     "smoothSE"   "compactSE"  "concSE"     "cptsSE"
## [21] "symSE"      "fracSE"     "radW"       "textW"      "perimW"
## [26] "areaW"      "smoothW"    "compactW"   "concW"      "cptsW"
## [31] "symW"       "fracW"
```

In addition to the patient ID (variable `id`) and diagnosis (variable `diag`; M = malignant, B = benign), the 30 real-valued cell nuclei features are of the form `feature{M/SE/W}`, where the suffix stands for mean, standard error, and worst along the following ten nuclei characteristics:

- `rad`: radius (mean of distances from center to points on the perimeter).
- `text`: texture (standard deviation of gray-scale values).
- `perim`: perimeter.
- `area`: area.
- `smooth`: smoothness (local variation in radius lengths).
- `compact`: compactness (perimeter^2 / area -1).
- `conc`: concavity (severity of concave portions of the contour).
- `cpts`: concave points (number of concave portions of the contour).
- `sym`: symmetry.
- `frac`: fractal dimension ("coastline approximation" -1).

The purpose of this report is to determine the important predictors of malignant tumors using a penalized logistic regression. The logistic regression model is

$$y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\rho_i), \qquad \rho_i = \text{logit}^{-1}(\boldsymbol{x}_i^{'}\boldsymbol{\beta}) = \frac{1}{1+\exp(-\boldsymbol{x}_i^{'}\boldsymbol{\beta})^{'}}$$

where y is the diagnostic (`diag`) binary response, and $\boldsymbol{x}$ is a vector of 31 predictors (including the intercept term).

The logistic regression model results in a loglikelihood function $l(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X})$. The penalty function on $\boldsymbol{\beta} = (\beta_0, ..., \beta_{30})$ is Elastic Net, such that for fixed α and λ, the penalized likelihood estimator is

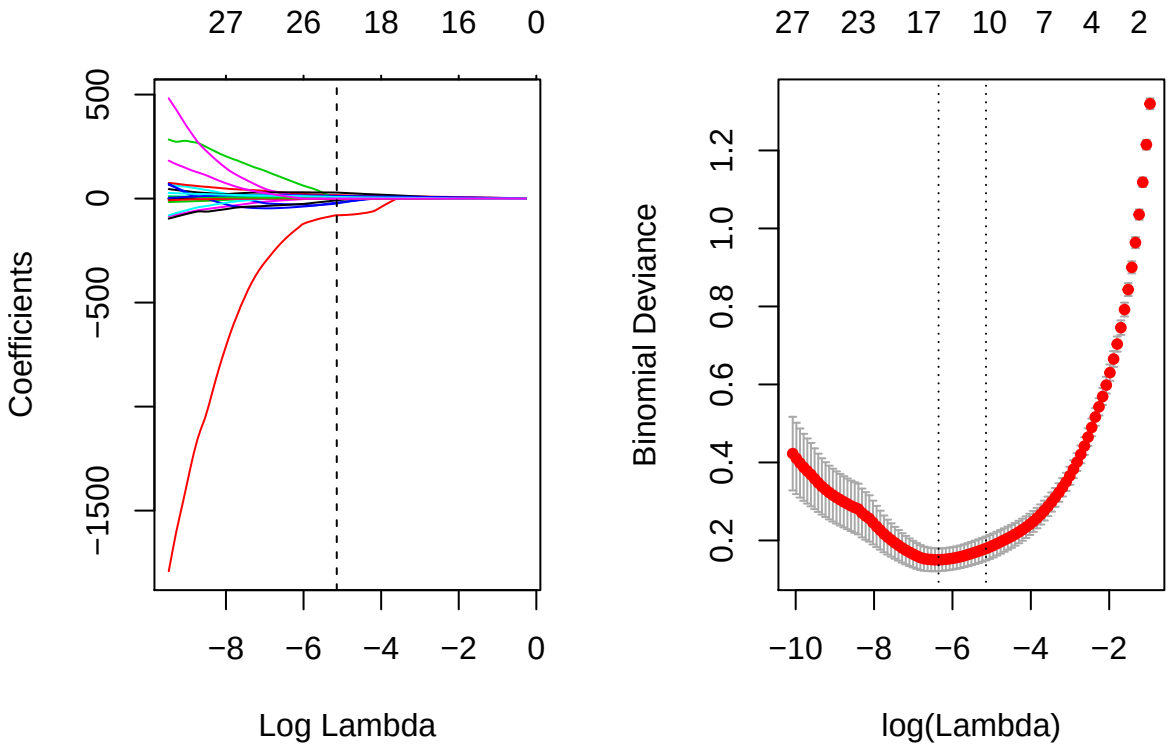
$$\tilde{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left[l(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) - \lambda \sum_{j=1}^{30} (1-\alpha)\beta_j^2 + \alpha|\beta_j| \right]$$

Result

Penalized Likelihood

In this section, we are going to fit a generalized linear model for the Wisconsin Diagnostic Breast Cancer (WDBC) data, and then estimate the optimal value of λ.

```
attach(tumor)
# build observation matrix
X <- as.matrix(tumor[, -c(1,2)])
# fit model using glmnet with alpha set to 0.5
fit = glmnet(x=X, diag, family = "binomial",
             alpha = 0.5)
# estimate the optimal value of lambda use 15-fold cross-validation
cvfit = cv.glmnet(x= X, diag, family = "binomial", nfolds = 15)
# select lambda hat
lambda.hat <- cvfit$lambda.1se
par(mfrow=c(1,2))
plot(fit, xvar = "lambda", label = TRUE)
abline(v=log(lambda.hat), lty = 2)
plot(cvfit)
```



The above code gives us two plots:

- The plot on the left hand side is the entire solution path for the penalized logistic regression with the WDBC data for $\alpha = 0.5$. As we can see from the plot, as $log(\lambda)$ approaches 0, $\boldsymbol{\beta}$ also approaches $\mathbf{0}$. That is, as the penalty becomes larger, fewer and fewer variables are included in the model. The vertical line in the plot is the estimate of λ selected by cross-validation.
- The plot on the right hand side is the results of a 15-fold cross-validation on the Deviance metric as a function of $log(\lambda)$. From this plot we see that as λ increases the deviance first decreases to its

minimum then increases back up. This tells us that when we either include too many variables or only few variables, the error is large. So we want to balance between the number of variables included and the deviance by choosing the largest value of lambda such that error is within 1 standard error of the minimum, i.e. $\hat{\lambda}_{1se}$.

Variable Selection

The non-zero penalized regression estimates $\tilde{\boldsymbol{\beta}}(\hat{\lambda}_{1se})$ are:

```
# non-zero coefficients except for the intercept
nz <- nonzeroCoef(coef(cvfit,s=lambda.hat))[-1]
coef(cvfit,s=lambda.hat)[nz,]
```

```
##        textM        cptsM        radSE       fracSE         radW
##   0.03391358  14.38544938   4.88784703 -90.07810914   0.68317037
##        textW      smoothW        concW        cptsW         symW
##   0.16384070  22.00861170   1.89132170  16.61084932   5.03938511
```

According to the above result, we can conclude that `fracSE`, `cptsW`, `cptsM`, and `smoothW` are features of tumor cell nuclei that are most predictive of a malignant growth. Please note that the coefficient of `fracSE` is a negtive number with large magnitude, i.e. it negatively affects the response (`diag`) significantly. However, it may not show up in the list under different cross-validations.