Modern Data Mining - HW 3

Cheshta Dhingra

## Problem 1

We will continue to use the Framingham Data (Framingham.dat) so that you are already familiar with the data and the variables. All the results are obtained through training data.

To keep our answers consistent, use a subset of the data, and exclude anyone with a missing entry. For your convenience, we've loaded it here together with a brief summary about the data.

We note that this dataset contains 311 people diagnosed without heart disease and 1095 with heart disease.

0 1   
 1095 311

After a quick cleaning up here is a summary about the data:

# using the comment=" ", we get rid of the ## in the output.  
summary(hd\_data.f)

HD AGE SEX SBP DBP   
 0:1086 Min. :45.00 FEMALE:730 Min. : 90.0 Min. : 50.00   
 1: 307 1st Qu.:48.00 MALE :663 1st Qu.:130.0 1st Qu.: 80.00   
 Median :52.00 Median :142.0 Median : 90.00   
 Mean :52.43 Mean :148.1 Mean : 90.16   
 3rd Qu.:56.00 3rd Qu.:160.0 3rd Qu.: 98.00   
 Max. :62.00 Max. :300.0 Max. :160.00   
 CHOL FRW CIG   
 Min. : 96.0 Min. : 52.0 Min. : 0.000   
 1st Qu.:200.0 1st Qu.: 94.0 1st Qu.: 0.000   
 Median :230.0 Median :103.0 Median : 0.000   
 Mean :234.6 Mean :105.4 Mean : 8.035   
 3rd Qu.:264.0 3rd Qu.:114.0 3rd Qu.:20.000   
 Max. :430.0 Max. :222.0 Max. :60.000

### Part 1A

Goal: Identify important risk factors for Heart.Disease. through logistic regression. Start a fit with just one factor, SBP, and call it fit1. Let us add one variable to this at a time from among the rest of the variables.

fit1 <- glm(HD~SBP, hd\_data.f, family=binomial)  
summary(fit1)  
fit1.1 <- glm(HD~SBP + AGE, hd\_data.f, family=binomial)  
summary(fit1.1)  
fit1.2 <- glm(HD~SBP + SEX, hd\_data.f, family=binomial)  
summary(fit1.2)  
fit1.3 <- glm(HD~SBP + DBP, hd\_data.f, family=binomial)  
summary(fit1.3)  
fit1.4 <- glm(HD~SBP + CHOL, hd\_data.f, family=binomial)  
summary(fit1.4)  
fit1.5 <- glm(HD~SBP + DBP, hd\_data.f, family=binomial)  
summary(fit1.5)  
fit1.6 <- glm(HD~SBP + FRW, hd\_data.f, family=binomial)  
summary(fit1.6)  
fit1.7 <- glm(HD~SBP + CIG, hd\_data.f, family=binomial)  
summary(fit1.7)

1. Which single variable would be the most important to add? Add it to your model, and call the new fit fit2.

We will pick up the variable either with highest value, or smallest value. From all the two variable models we see that SEX will be the most important addition on top of the SBP. And here is the summary report.

## How to control the summary(fit2) output to cut some junk?  
## We could use packages: xtable or broom.   
library(xtable)  
options(xtable.comment = FALSE)  
fit2 <- glm(HD~SBP + SEX, hd\_data.f, family=binomial)  
xtable(fit2)

1. Is the residual deviance of fit2 always smaller than that of fit1? Why or why not?

Yes. The larger model (fit2) does a better job of predicting our Y variable than fit1 does, so the residual deviance must be lower. A lower deviance always indicates a better fit.

1. Perform both the Wald test and the Likelihood ratio tests (Chi-Squared) to see if the added variable is significant at the .01 level. What are the p-values from each test? Are they the same?

##   
## Call:  
## glm(formula = HD ~ SBP + SEX, family = binomial, data = hd\_data.f)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6408 -0.7373 -0.5726 -0.4169 2.2452   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.570256 0.389727 -11.727 < 2e-16 \*\*\*  
## SBP 0.018717 0.002324 8.053 8.07e-16 \*\*\*  
## SEXMALE 0.903420 0.139762 6.464 1.02e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1469.3 on 1392 degrees of freedom  
## Residual deviance: 1373.8 on 1390 degrees of freedom  
## AIC: 1379.8  
##   
## Number of Fisher Scoring iterations: 4

## Waiting for profiling to be done...

## 0.5 % 99.5 %  
## (Intercept) -5.59732078 -3.58576526  
## SBP 0.01280634 0.02480946  
## SEXMALE 0.54725779 1.26824987

## [1] 1.478913e-22

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: HD  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 1392 1469.3   
## SBP 1 51.864 1391 1417.5 5.949e-13 \*\*\*  
## SEX 1 43.700 1390 1373.8 3.828e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

p-value from Wald test = ?

p-value from Chi-sq test = 1.478913e-22

### Part 1B

Model building: Start with all variables. Our goal is to fit a well-fitting model, that is still small and easy to interpret (parsimonious).

1. Use backward selection method. Only keep variables whose coefficients are significantly different from 0 at .05 level. Kick out the variable with the largest p-value first, and then re-fit the model to see if there are other variables you want to kick out.

After kicking out DBP, FRW and CIG, we get:

##   
## Call:  
## glm(formula = HD ~ AGE + SEX + SBP + CHOL, family = binomial,   
## data = hd\_data.f)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6069 -0.7348 -0.5523 -0.3476 2.4344   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -8.408724 0.908600 -9.255 < 2e-16 \*\*\*  
## AGE 0.056639 0.014500 3.906 9.38e-05 \*\*\*  
## SEXMALE 0.989870 0.145053 6.824 8.84e-12 \*\*\*  
## SBP 0.016956 0.002362 7.179 7.02e-13 \*\*\*  
## CHOL 0.004480 0.001495 2.996 0.00274 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1469.3 on 1392 degrees of freedom  
## Residual deviance: 1349.0 on 1388 degrees of freedom  
## AIC: 1359  
##   
## Number of Fisher Scoring iterations: 4

1. Use AIC as the criterion for model selection. Find a model with small AIC through exhaustive search. Does exhaustive search guarantee that the p-values for all the remaining variables are less than .05? Is our final model here the same as the model from backwards elimination?

##   
## Call:  
## glm(formula = y ~ ., family = family, data = Xi, weights = weights)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.7066 -0.7279 -0.5517 -0.3343 2.4501   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -9.227856 0.996153 -9.263 < 2e-16 \*\*\*  
## AGE 0.061529 0.014775 4.164 3.12e-05 \*\*\*  
## SEXMALE 0.911274 0.157117 5.800 6.63e-09 \*\*\*  
## SBP 0.015966 0.002487 6.420 1.37e-10 \*\*\*  
## CHOL 0.004493 0.001503 2.990 0.00279 \*\*   
## FRW 0.006039 0.004004 1.508 0.13151   
## CIG 0.012279 0.006088 2.017 0.04369 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1469.3 on 1392 degrees of freedom  
## Residual deviance: 1343.3 on 1386 degrees of freedom  
## AIC: 1357.3  
##   
## Number of Fisher Scoring iterations: 4

No, the final model here is different. Using BIC we included Age, Sexmale, SBP and Chol. Using AIC we include Age, Sexmale, SBP, Chol, FRW, CIG. We get a lower AIC from the second model.

1. Use the model chosen from part ii. as the final model. Write a brief summary to describe important factors relating to Heart Diseases (i.e. the relationships between those variables in the model and heart disease). Give a definition of “important factors”.

Final model: HD ~ AGE + SEXMALE + SBP + CHOL + FRW + CIG. Using the estimates for the coefficients on these variables, we see that all of them are associated with an increased probability of heart disease. As an individual's AGE increases by a year, the log odds of HD increases by 0.06. If the individual is MALE, his log odds of HD are higher than a FEMALE's by 0.91. A one unit increase in SBP is associated with a 0.016 increase in logg odds ratio of HD, a one unit increase in CHOL is associated with a 0.004 increase in logg odds ratio of HD, a one unit increase in FRW (Framingham relative weight) is associated with a 0.006 unit increase in log odds ratio for HD and finally, a one unit increase in cigarette smoking is associated with a 0.012 unit increase in log odds ratio for HD.

These 6 factors are considered important, because they ensure that the predicted probability for Heart Disease for each individual corresponds as closely as possible to the indicidual's observed HD status. In other words, the coefficients on the intercept and these variables are such that plugging these estimates into the model for p(HD) = (e^(β0+β1X)) / (1 + e^(β0+β1X)), yields a number close to 1 for all individuals with heart disease and a number close to 0 for all individuals who did not.

### Part 1C

Liz is a patient with the following readings: AGE=50, GENDER=FEMALE, SBP=110, DBP=80, CHOL=180, FRW=105, CIG=0. What is the probability that she will have heart disease, according to our final model?

fit3.predict <- predict(fit3, hd\_data.new, type="response")  
fit3.predict

## 1407   
## 0.05035437

Liz will have Heart disease with a probability of 0.05 according to our model.

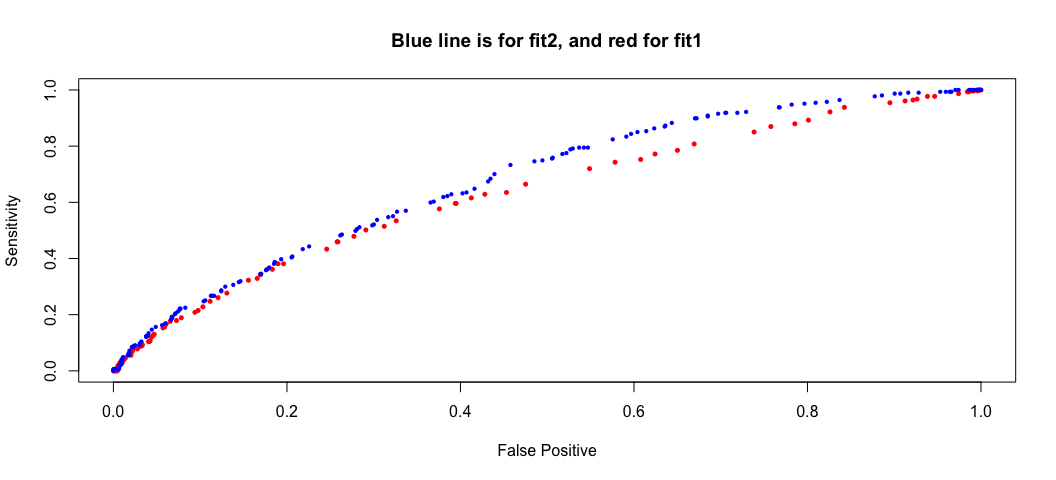
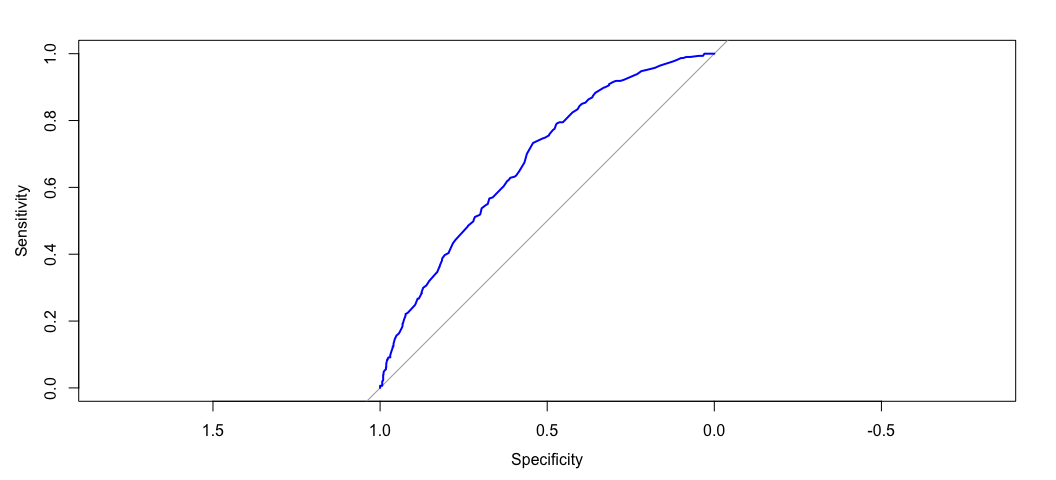
### Part 2

Classification analysis

1. Display the ROC curve using fit1. Explain what ROC reports and how to use the graph. Specify the classifier such that the False Positive rate is less than .1 and the True Positive rate is as high as possible.

The ROC curve here plots Sensitivity (True positive rate) vs. Specificity (True Negative rate). The x axis can be thought of as the False Positive rate with values going from 0 to 1. (FP = 1-specificity). The classifier for which the FP rate is restricted to less than 0.1 is 0.3. This means that if our probability for HD given SBP exceeds 0.3, we will label that individual as 1 for HD.

hat Y=1 if p(y=1|x) > 0.3

1. Overlay two ROC curves: one from fit1, the other from fit2. Does one curve always contain the other curve? Is the AUC of one curve always larger than the AUC of the other one? Why or why not? 

## Area under the curve: 0.6357

## Area under the curve: 0.6801

It is not necessary that the new ROC curve will lie above the old one, or that AUC will always increase. This is because we are applying our model to a test set, rather than the training set. Adding more variables does not necessarily yield a better model.

AUC(fit1.roc) = .636 and AUC(fit2.roc) = 0.680

1. Estimate the Positive Prediction Values and Negative Prediction Values for fit1 and fit2 using .5 as a threshold. Which model is more desirable if we prioritize the Positive Prediction values?

Fit 1: positive prediction = 0.45, negative prediction = 0.783. Fit 2: positive prediction = 0.47, negative prediction = 0.786. Fit2 is slightly more desireable since it has a higher positive prediction value

### Part 3

Bayes rules with risk ratio a10/a01=10 or a10/a01=1. Use your final model obtained from 1 B) to build a class of linear classifiers.

1. Write down the linear boundary for the Bayes classifier if the risk ratio of .

The Bayes linear boundary is -9.227 + 0.0615AGE + 0.9113SEXMALE + 0.0159SBP+ 0.0045CHOL + 0.006FRW + 0.0123CIG = -1.0047

1. What is your estimated weighted misclassification error for this given risk ratio?

fit6.pred.bayes=rep("0", 1393)  
fit6.pred.bayes[fit6$fitted > .09]="1"   
MCE.bayes=(sum(5\*(fit6.pred.bayes[hd\_data.f$HD == "1"] != "1"))   
 + sum(fit6.pred.bayes[hd\_data.f$HD == "0"] != "0"))/length(hd\_data.f$HD)  
MCE.bayes

## [1] 0.684135

MCE.bayes = 0.684

1. Recall Liz, our patient from part 1. How would you classify her under this classifier?

Liz would be classified as not having HD, since her p(HD) = 0.05 < 0.09.

Now, draw two estimated curves where x = posterior threshold, and y = misclassification errors, corresponding to the thresholding rule given in x-axis.

1. Use weighted misclassification error, and set . How well does the Bayes rule classifier perform?
2. Use weighted misclassification error, and set . How well does the Bayes rule classifier perform?

## Problem 2

## 1.Goal

The goal of this study is to determine the factors that are important in predicting whether or not a bill gets passed through the Pennsylvania legislature. We examine about 8000 bills proposed since 2009, with the goal of building a classifier which has decent power to forecast which bills are likely to be passed. We have information regarding the session in which the bill was introduced, the spondoring party, number of cosponsors (split among republicans and democrats), the word count of the bill's title, the originating committee, day of the week the bill was introduced, number of amendments it went through, whether the sponsor is in leadership, how many of the sponsors sit on the committee to which the bill is referred (split among democrats and republicans). In this report I will choose a best set of classifiers such that:

* The testing ROC curve pushes to the upper left corner the most, and has a competitive AUC value.
* Propose a reasonable loss function, and report the Bayes rule together with its weighted MIC.
* You may also create some sensible variables based on the predictors or make other transformations to improve the performance of your classifier.

## 2. Summary

Preliminary summary of the data: Sponsoring Party: About an equal split, with 46% of the bills coming from Democrats and 54% from Republicans. Num cosponsors: Between 0 and 165 with a mean of 23. Democrat cosponsors: Between 0 and 90 with mean of 10.3. Republican cosponsors: Between 0 and 99 with mean of 13. Word count of the title: Between 6 and 751, mean of 34. Number of amendments: Between 0 and 8, mean of 0.18. % of sponsors in leadership: 62%

## 3. Building the classifier

### Model building

Start with all variables. Our goal is to fit a well-fitting model, that is still small and easy to interpret (parsimonious).

1. Use backward selection method. Only keep variables whose coefficients are significantly different from 0 at .05 level.

After kicking out num\_r\_cosponsors, num\_originating\_committee\_cosponsors\_r, num\_cosponsors, day.of.week.introduced, is\_sponsor\_in\_leadership, num\_originating\_committee\_cosponsor:

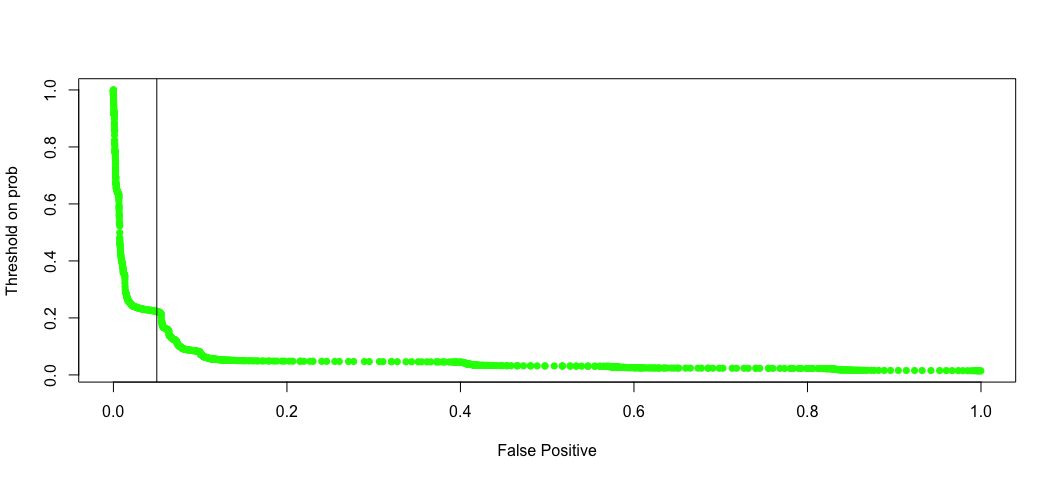
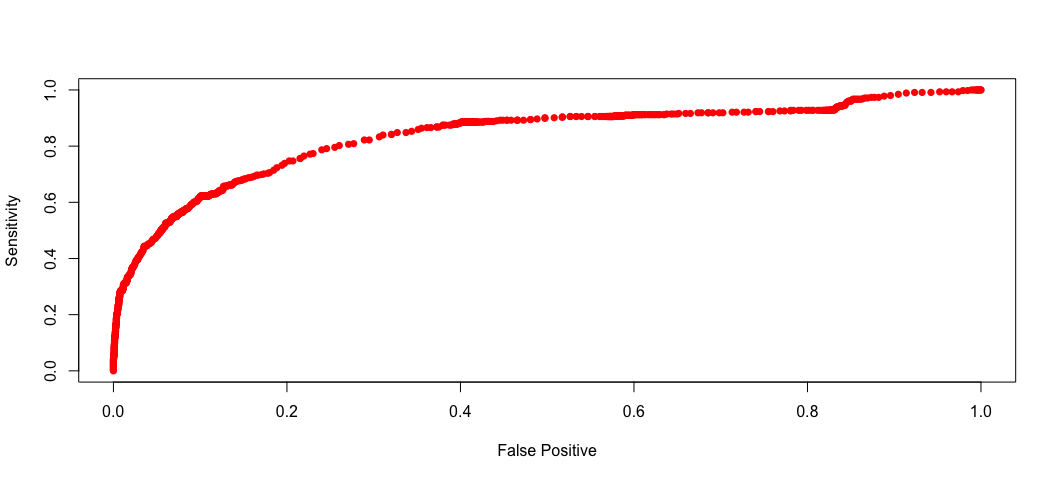
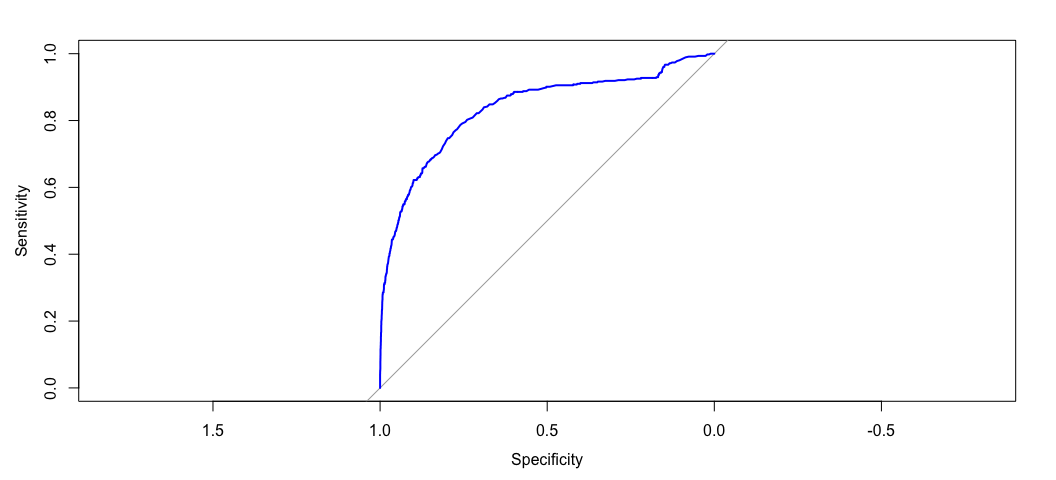
Final model: status ~ sponsor\_party + session + title\_word\_count + num\_amendments

summary(fit0.6)

##   
## Call:  
## glm(formula = status ~ sponsor\_party + session + title\_word\_count +   
## num\_amendments, family = binomial, data = databills)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.7280 -0.3109 -0.2502 -0.2135 2.9013   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.269217 0.137475 -31.054 < 2e-16 \*\*\*  
## sponsor\_partyRepublican 0.726797 0.129537 5.611 2.01e-08 \*\*\*  
## session2011-2012 0.419209 0.140381 2.986 0.00282 \*\*   
## session2013-2014 0.415172 0.158068 2.627 0.00863 \*\*   
## title\_word\_count 0.004709 0.001097 4.292 1.77e-05 \*\*\*  
## num\_amendments 1.785032 0.077224 23.115 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3318.4 on 6646 degrees of freedom  
## Residual deviance: 2456.3 on 6641 degrees of freedom  
## AIC: 2468.3  
##   
## Number of Fisher Scoring iterations: 6

Can also use AIC as the criterion for model selection. Find a model with small AIC through exhaustive search (chunk not evaluated here)

### Classification analysis

ROC curve using fit0.6: 

## Area under the curve: 0.8363

I chose to specify the classifier such that the False Positive rate is less than 0.05 and the True Positive rate is as high as possible. Area under the curve: 0.8363.

### Positive and Negative prediction values

I will estimate the Positive Prediction Values and Negative Prediction Values for fit0.6 and using .33 as a threshold, since this is approximately the threshold that corresponds to a FP rate of no more than 0.05.

Fit 0.6: positive prediction = 0.63, negative prediction = 0.95.

Bayes rules with risk ratio a10/a01=10 or a10/a01=1. Use your final model obtained from 1 B) to build a class of linear classifiers.

### Loss function

Let the risk ratio of .

The Bayes linear boundary is -4.269 + 0.7267sponsor\_party + 0.42session2011-2012 + 0.415session2013-2014 + 0.0047title\_word\_count + 1.785num\_amendments = -1.0047

### MCE

What is your estimated weighted misclassification error for this given risk ratio?

fit0.6.pred.bayes=rep("0", 6647)  
fit0.6.pred.bayes[fit0.6$fitted > .09]="1"   
MCE.bayes=(sum(5\*(fit0.6.pred.bayes[databills$status == "1"] != "1"))   
 + sum(fit0.6.pred.bayes[databills$status == "0"] != "0"))/length(databills$status)  
MCE.bayes

## [1] 0.2225064

Weighted MCE = 0.222.

*Final notes*: The data is graciously lent from a friend. It is only meant for you to use in this class. All other uses are prohibited without permission.