Homework 6

Evan Yacek ety78

This homework is due on Mar. 8, 2016 at 11:59pm. Please submit as a PDF file on Canvas.

For this homework you will use the `wine` data set from the previous homework. For this homework, however, we have removed samples from cultivar 3. The `wine` data set contains concentrations of 13 different chemical compounds (`chem1`-`chem13`) in 130 samples of wines grown in Italy. Each row is a different sample of wine, and the data set now contains just two different cultivars (`cultivar`) of wine.

```
wine <- read.csv("http://wilkelab.org/classes/SDS348/data_sets/wine.csv", colClasses = c("cultivar" =
"factor")) %>% filter(cultivar != 3)
head(wine)
```

```
##   cultivar chem1 chem2 chem3 chem4 chem5 chem6 chem7 chem8 chem9 chem10
## 1        1 14.23  1.71  2.43  15.6   127  2.80  3.06  0.28  2.29   5.64
## 2        1 13.20  1.78  2.14  11.2   100  2.65  2.76  0.26  1.28   4.38
## 3        1 13.16  2.36  2.67  18.6   101  2.80  3.24  0.30  2.81   5.68
## 4        1 14.37  1.95  2.50  16.8   113  3.85  3.49  0.24  2.18   7.80
## 5        1 13.24  2.59  2.87  21.0   118  2.80  2.69  0.39  1.82   4.32
## 6        1 14.20  1.76  2.45  15.2   112  3.27  3.39  0.34  1.97   6.75
##   chem11 chem12 chem13
## 1   1.04   3.92   1065
## 2   1.05   3.40   1050
## 3   1.03   3.17   1185
## 4   0.86   3.45   1480
## 5   1.04   2.93    735
## 6   1.05   2.85   1450
```

Problem 1

A. (1 pt) *Make a logistic regression model that predicts the cultivar from the concentrations of* ***three chemical compounds of your choosing*** *(not all of them!) in the `wine` data set. Show the summary (using `summary`) of your model below.*

I choose chem1, chem2, chem3.

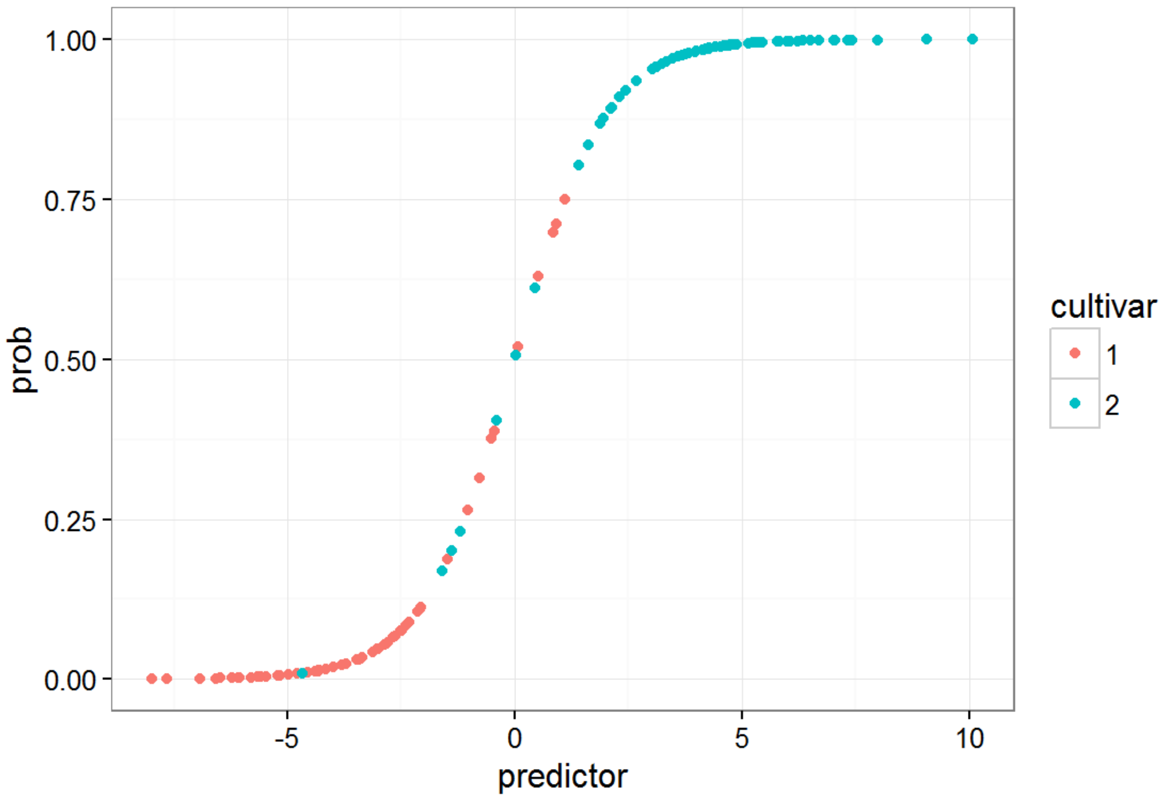
```
glm.out <- glm(cultivar ~ chem1 + chem2 + chem3, data = wine , family = "binomial")
summary(glm.out)
```

```
## 
## Call:
## glm(formula = cultivar ~ chem1 + chem2 + chem3, family = "binomial", 
##     data = wine)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -1.66637  -0.20594   0.03888   0.17769   3.06013  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) 68.88854   12.71587   5.418 6.04e-08 ***
## chem1       -4.68710    0.91379  -5.129 2.91e-07 ***
## chem2       -0.08856    0.33612  -0.263   0.7922    
## chem3       -3.17020    1.45917  -2.173   0.0298 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 179.109  on 129  degrees of freedom
## Residual deviance:  45.927  on 126  degrees of freedom
## AIC: 53.927
## 
## Number of Fisher Scoring iterations: 7
```

B. (1 pt) *Make a plot of the fitted probability as a function of the linear predictor, colored by cultivar.*

```
lr_data <- data.frame(predictor=glm.out$linear.predictors, prob=glm.out$fitted.values, cultivar = win
e$cultivar)

ggplot(lr_data, aes(x=predictor, y=prob, color=cultivar)) + geom_point()
```



C. (3 pts) *Choose a probability cut-off for classifying a given sample of wine as cultivar 1 or cultivar 2. State the cut-off that you chose. Calculate the* ***true positive rate*** *and* ***false positive rate*** *and interpret these rates in the context of the* `wine` *data set. Your answer should mention something about cultivars and the three chemical compounds you chose in part A.*

I choose a cut-off of 0.6.

```
cutoff <- 0.6

pred_data <- data.frame(probability=glm.out$fitted.values, cultivar=wine$cultivar)


pred_data %>% filter(probability < cutoff & cultivar==1) %>%
  tally() -> true_pos


pred_data %>% filter(probability >= cutoff & cultivar==2) %>%
  tally() -> true_neg


pred_data %>% filter(cultivar == 1) %>%
  tally() -> pos_total


pred_data %>% filter(cultivar == 2) %>%
  tally() -> neg_total

true_pos_rate <- true_pos$n/pos_total$n
true_neg_rate <- true_neg$n/neg_total$n

true_pos_rate
```

```
## [1] 0.9322034
```

```
true_neg_rate
```

```
## [1] 0.915493
```

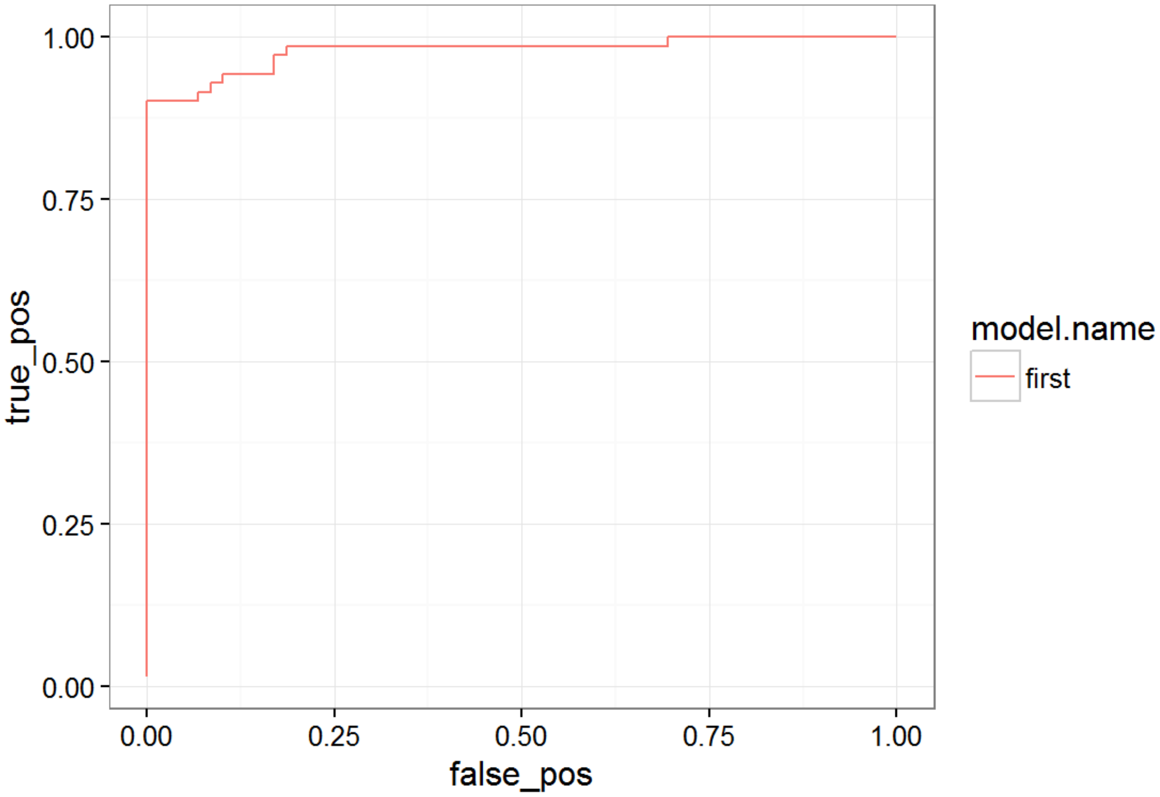
The true positive rate is 93.2% and the true negative rate is is 91.5%. Thus, the model from correctly identifies 93.2% of cultivars from 1, and 91.5% of cultivars from 2 both using chem1, chem2, and chem3 as predictors. ###Problem 2

A (1pt). *Using the `roc_curve` function below (which we also used in class), plot an ROC curve for the model that you created in Problem 1A. Does the model perform better than a model in which you randomly classify a wine sample as cultivar 1 or cultivar 2? Explain your answer in 1-2 sentences.*

```
calc_ROC <- function(probabilities, known_truth, model.name=NULL)
  {
  outcome <- as.numeric(factor(known_truth))-1
  pos <- sum(outcome) # total known positives
  neg <- sum(1-outcome) # total known negatives
  pos_probs <- outcome*probabilities # probabilities for known positives
  neg_probs <- (1-outcome)*probabilities # probabilities for known negatives
  true_pos <- sapply(probabilities,
                     function(x) sum(pos_probs>=x)/pos) # true pos. rate
  false_pos <- sapply(probabilities,
                     function(x) sum(neg_probs>=x)/neg)
  if (is.null(model.name))
    result <- data.frame(true_pos, false_pos)
  else
    result <- data.frame(true_pos, false_pos, model.name)
  result %>% arrange(false_pos, true_pos)
  }


ROC.final <- calc_ROC(probabilities=glm.out$fitted.values,
                 known_truth=wine$cultivar,
                 model.name="first")


ggplot(data=ROC.final, aes(x=false_pos, y=true_pos, color=model.name)) +
  geom_line()
```



Yes. Randomly picking which cultivar the wine was in would yield about a 50% chance. However, our model correctly identifies the wine cultivar(true_positive) at a much higher rate than 50%, and has a lower chance of a false positive.At around 80-90 % our model is predicting a true postive with almost 0 percent false positives.

B. (4 pts) *Choose a new set of predictor variables (different from the variables that you chose in Problem 1A), and create a logistic regression model. Plot an ROC curve for your newly-created model and, on the same plot, add an ROC curve from your model in Problem 1A. What can you conclude from your plot? Which model performs better and why? Support your conclusions* ***with AUC values for each model****.*

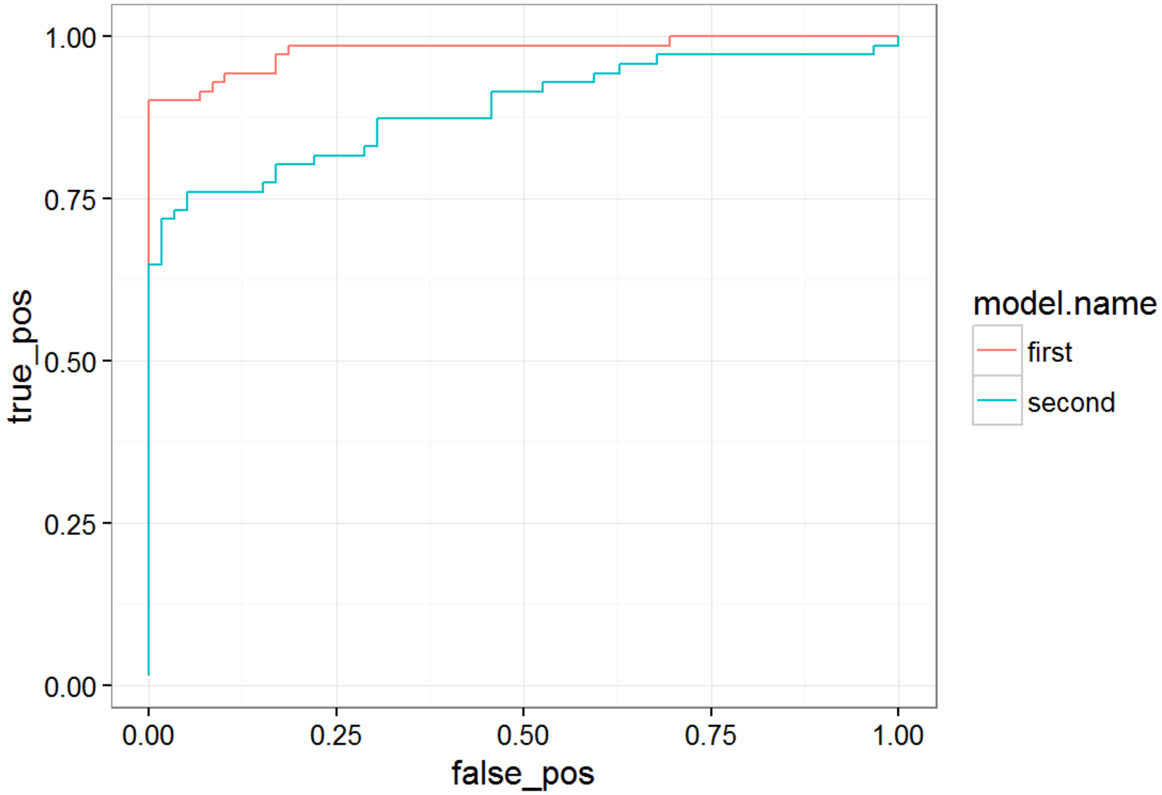
I choose chem6, chem7, chem8

```
glm.out2 <- glm(cultivar ~ chem6 + chem7 + chem8, data = wine , family = "binomial")
summary(glm.out2)
```

```
##
## Call:
## glm(formula = cultivar ~ chem6 + chem7 + chem8, family = "binomial",
##     data = wine)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6035  -0.6907   0.1865   0.5975   3.5442
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.7689     2.0091   2.374 0.017613 *
## chem6         0.6341     0.9491   0.668 0.504069
## chem7        -2.9881     0.8100  -3.689 0.000225 ***
## chem8         4.5002     2.9008   1.551 0.120817
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 179.11  on 129  degrees of freedom
## Residual deviance: 114.87  on 126  degrees of freedom
## AIC: 122.87
##
## Number of Fisher Scoring iterations: 5
```

```
ROC.final2 <- calc_ROC(probabilities=glm.out2$fitted.values,
                 known_truth=wine$cultivar,
                 model.name="second")


ROCs <- rbind(ROC.final, ROC.final2)
ggplot(data=ROCs, aes(x=false_pos, y=true_pos, color=model.name)) +
  geom_line()
```



```
ROCs%>% group_by(model.name) %>%
  mutate(delta=false_pos-lag(false_pos)) %>%
  summarize(AUC=sum(delta*true_pos, na.rm=T)) %>%
  arrange(desc(AUC))
```

```
## Source: local data frame [2 x 2]
##
##   model.name       AUC
##       (fctr)     (dbl)
## 1      first 0.9792313
## 2     second 0.8887563
```

After creating the second model, and plotting the roc curve we can see that the first model was better. Not only is this more apparent by the plot, but also the AUC value 0.979 is greater than the second model 0.88. This is probably due to the fact the first model had much lower p values in the glm.