Homework 6

ROhit Kamat rgk359

For the homework I used the `wine` data set. 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.*

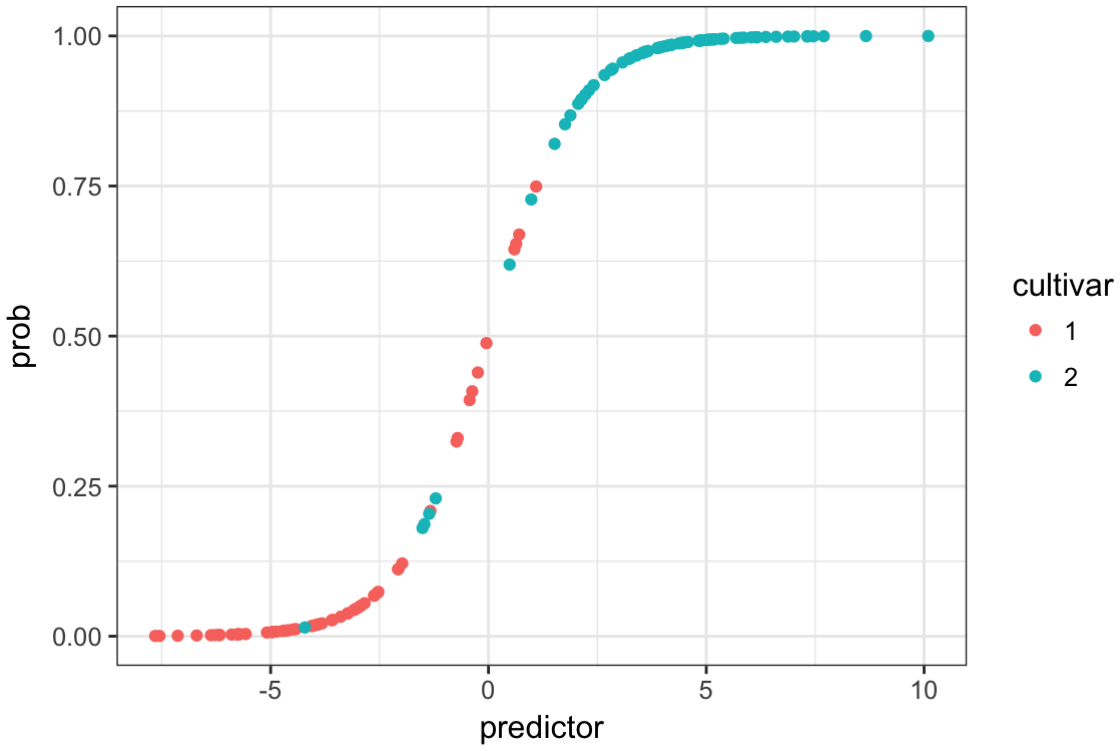
I choose chem1, chem 3, chem 5

```
glm.out<- glm(cultivar ~ chem1 + chem3 + chem5, wine, family=binomial)
summary(glm.out)
```

```
##
## Call:
## glm(formula = cultivar ~ chem1 + chem3 + chem5, family = binomial,
##     data = wine)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.66314  -0.19701   0.03939   0.18832   2.90819
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 68.68560   12.60394   5.450 5.05e-08 ***
## chem1       -4.59484    0.90611  -5.071 3.96e-07 ***
## chem3       -2.73126    1.54446  -1.768    0.077 .
## chem5       -0.02236    0.02792  -0.801    0.423
## ---
## 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:  45.427  on 126  degrees of freedom
## AIC: 53.427
##
## 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, c
ultivar=wine$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 probability cutoff of 0.75.

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

#cutoff at .75
cutoff <- 0.75

#Number of true cultivar 1 samples identified as cultivar 1 (true positives)
pred_data %>% filter(probability <= cutoff & cultivar==1) %>%
  tally() -> cult1_true
#Number of true cultivar 2 samples identified as cultivar 2 (true negatives)
pred_data %>% filter(probability > cutoff & cultivar=="2") %>%
  tally() -> cult2_true

#Total number of true cultivar 1 samples (known positives)
pred_data %>% filter(cultivar==1) %>%
  tally() -> cult1_total
#Total number of true cultivar 2 samples (known negatives)
pred_data %>% filter(cultivar==2) %>%
  tally() -> cult2_total

#calculate the true positive rate and false positive rate

true_positive= cult1_true$n/(cult1_total$n)
true_negative= cult2_true$n/(cult2_total$n)
true_positive
```

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

```
#False positive (1-true negative rate)
false_positive = 1 - true_negative
false_positive
```

```
## [1] 0.09859155
```

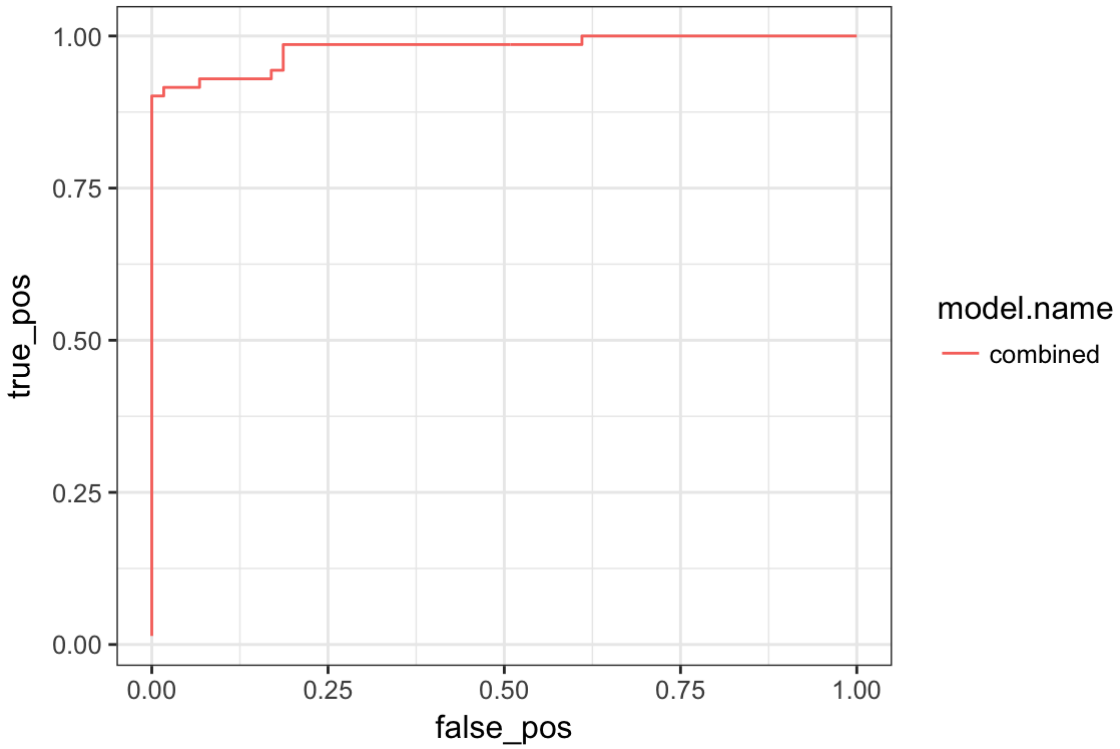
So I compared cultivar 1 and cultivar 2 based on the concentration of wine 1, wine 3, and wine 5. With a probability cutoff of 0.75, I had a true positive of 1 which means 100% of cultivar 1 were correctly identified below the cutoff point of 75%. I receive a false positive of 0.0986 which means 9.86% of the cultivar 2 were incorrectly identified below the cutoff of 75%.

Problem 2

A (1pt). *Using the* `calc_ROC` *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)
  }

glm.out <- glm(cultivar ~ chem1 + chem3 + chem5,
               data =wine,
               family = binomial)
ROC1 <- calc_ROC(probabilities=glm.out$fitted.values, known_truth=wine$cultivar,
                 model.name="combined")
ggplot(data=NULL, aes(x=false_pos, y=true_pos)) + geom_line(data=ROC1, aes(color=model.n
ame))
```



This model is better than a model that randomly identifies sample as cultivar 1 and cultivar 2. A model that randomly identifies samples as cultivar 1 and cultivar 2 is a straight line. The model I created is more curved which represents a better model for identifying cultivar 1 than cultivar 2.

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****.*

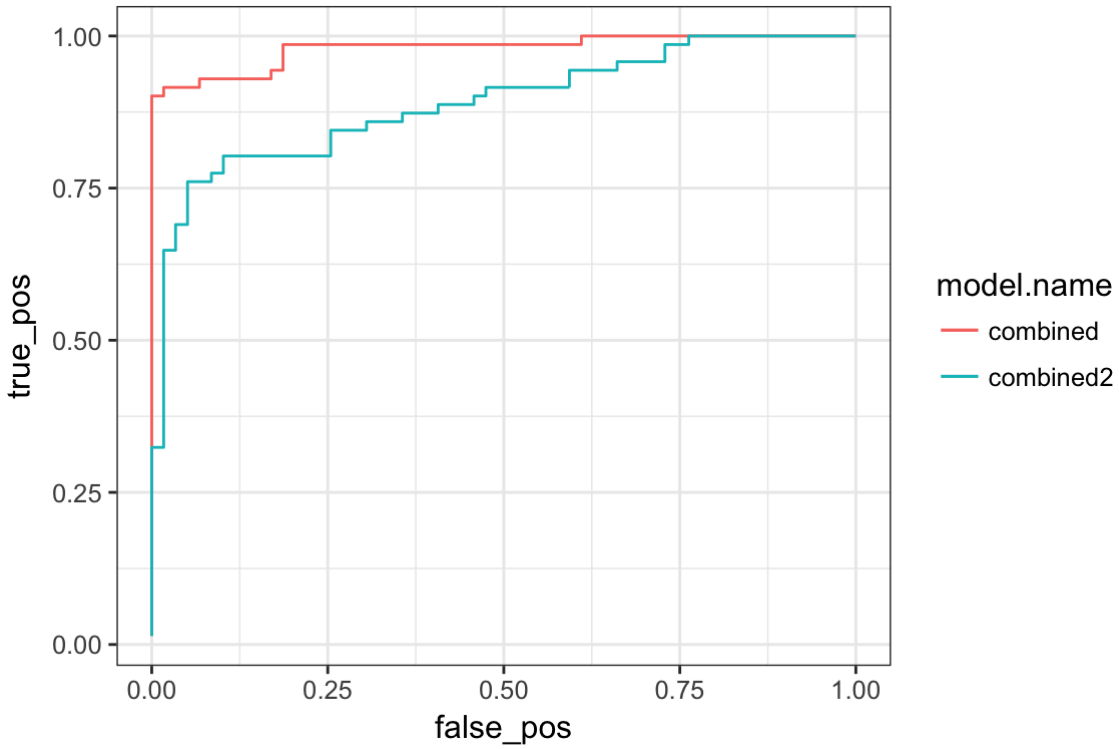
I choose Chem 2, Chem 4, Chem 6

```
# create a logistic regression model
glm.out<- glm(cultivar ~ chem2 + chem4 + chem6, wine, family=binomial)
summary(glm.out)
```

```
##
## Call:
## glm(formula = cultivar ~ chem2 + chem4 + chem6, family = binomial,
##     data = wine)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3382  -0.7186   0.1340   0.6537   2.2394
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.7426     2.1755   0.341    0.733
## chem2        -0.3276     0.2650  -1.236    0.216
## chem4         0.4203     0.1016   4.138 3.50e-05 ***
## chem6        -3.0087     0.6499  -4.629 3.67e-06 ***
## ---
## 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: 107.81  on 126  degrees of freedom
## AIC: 115.81
##
## Number of Fisher Scoring iterations: 5
```

```
#Plot ROC curve
ROC2 <- calc_ROC(probabilities=glm.out$fitted.values, known_truth=wine$cultivar,
                 model.name="combined2")

ggplot(data=NULL, aes(x=false_pos, y=true_pos)) +
  geom_line(data=ROC1, aes(color=model.name)) +
  geom_line(data=ROC2, aes(color=model.name))
```




```
#calculate area of roc curve

ROC1 %>% mutate(delta=false_pos-lag(false_pos)) %>%
  summarize(AUC=sum(delta*true_pos, na.rm=T))
```

```
##         AUC
## 1 0.9799475
```

```
ROC2%>% mutate(delta=false_pos-lag(false_pos)) %>%
  summarize(AUC=sum(delta*true_pos, na.rm=T))
```

```
##         AUC
## 1 0.8892337
```

From the plot I can conclude that the combined model, the ROC Curve from problem 1A in the model, was the model that performed the best. This is because the area under the combined model is greater than the area under the combined2 model, and the area under the curve tells us how good the model's prediction are. The combined model had an area under the curve of .9799, while combined2 model had an area under the curve of 0.8892. Therefore I conclude that wine 1,3, and 5 are able to predict the cultivar better than wine 2,4, and 6.