Info 573: Data Science I: Theoretical Foundations
Lab: Machine Learning

Instructor: Ben Althouse; Contact: bma85@uw.edu

In this lab you will practice logistic regression and prediction. You will try to predict whether someone will earn more than $50,000/year based on several predictors.

You may work with a partner on this lab, however, you will be asked to submit a copy of your analysis code to Canvas at the end of class - each individual must submit their own version of their code – *please be sure to put your name in the code!* Keep track of all the commands you run using a text editor or R script.

You should comment your code as you run through this exercise. You can do this in R using the # character. Please answer the questions posed in the exercise/lab by adding comments to your R script.

Dataset descriptions: Table below gives the variables in the datasets.

age: continuous.

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

education-num: continuous.

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

sex: Female, Male.

hours-per-week: continuous.

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

1. Load the data.

Make a new column in the dataset that is a 0,1 response for >50K or <=50k. Here is one way to do it:

```
data$income.g50 <- rep(0, nrow(data))

data$income.g50[data$income==" >50K"] <- 1
```

2. Exploring Relationships I: Run a logistic regression looking at the odds ratios of level of education adjusting for age, sex, and race.

mod <- glm(income.g50 ~ education + age + sex + race, data=data[,!colnames(data)%in%"income"], family="binomial")

a. What are the odds ratios for high earnings (remember the output of `summary()` gives log odds ratios) for having a masters degree? Or a 1^{st} – 4^{th} grade education? Are these statistically significant? What about multiple comparisons?

b. What are the effects of age and sex? Again, are they statistically significant? Are they practically significant? Are they fair?

3. Exploring Relationships II: Plot age by the outcome and the observed predicted probabilities. Why are the predicted probabilities so variable?

```
x <- data$age

plot(x, data$income.g50, col="blue")

fits <- fitted(mod)

points(x, fits, pch=19, cex=0.3)
```

4. Explore some cutoffs for the probabilities: Tabulate the outcome with a cutoff of 0.25, 0.5, and 0.75. Which has the lowest percent error?

```
tab <- table(data$income.g50, fits>=0.5)

(tab[1,2]+tab[2,1])/sum(tab)
```

5. Examine this model.

a. Plot the ROC curve and calculate the AUC for this model.

```
library(AUC)

y <- factor(data$income.g50)

rr <- roc(fits, y)

plot(rr)

auc(rr)
```

b. How well does it fit?

6. Let's formulate another model.

a. Fit a model with all covariates (except "income"!). Do you see the same patterns for level of schooling?

```
mod <- glm(income.g50~.,
data=data[,!colnames(data)%in%c("income")], family="binomial")
```

b. Plot the age by the outcome and the observed predicted probabilities. Do the predicted probabilities have the same pattern as the other model? Why or why not?

c. Calculate the percent error as before for cutoffs 0.25, 0.5, 0.75. Which cutoff has the lowest percent error? Does this model perform better than the other model?

d. Plot the ROC and calculate the AUC. Again, does this model out perform the other model?

Extra credit (5 points): Run a k-fold validation on both models and decide which you would prefer to use for predicting high income.