Problem Set #7

MACS 30150, Dr. Evans
Due Monday, Feb. 25 at 11:30am

1. **Multinomial logistic regression and cross validation (6 points).** For this problem, you will estimate the probability that a given wine comes from a given *cultivar*. The data in the file `strongdrink.txt` (taken from the UCI Machine Learning Repository) are the results of a chemical analysis of 176 Italian wines from three known cultivars (a cultivar is a group of grapes selected for desirable characteristics that can be maintained by propagation). The chemical analysis determined the quantities of the following 13 different constituents (the last 13 variables):

Variable	Name	Variable	Name
Alcohol	`alco`	Nonflavanoid phenols	`nonfl_phen`
Malic acid	`malic`	Proanthocyanins	`proanth`
Ash	`ash`	Color intensity	`color_int`
Alkalinity of ash	`alk`	Hue	`hue`
Magnesium	`magn`	OD280/OD315 of diluted wines	`OD280rat`
Total phenols	`tot_phen`	Proline	`proline`
Flavanoids	`flav`		

   (a) Use a multinomial logistic regression model of the following form with the following linear predictor η_j for $j = 1, 2$ (the baseline class is $j = 3$).

$$Pr(cultivar_i = j|X\beta_j) = \frac{e^{\eta_j}}{1 + \sum_{j=1}^{J-1} e^{\eta_j}} \quad \text{for} \quad j = 1, 2$$

$$\text{where} \quad \eta_j = \beta_{j,0} + \beta_{j,1} alco_i + \beta_{j,2} malic_i + \beta_{j,3} tot\_phen_i + \beta_{j,4} color\_int_i$$

   Estimate the model on a 75% sample training set using the following command. Report your two sets of estimated coefficients and intercepts for $j = 1$ and $j = 2$ (not the coefficients for $j = 3$). Report your error rates (1 - precision) on the test set using the code below. Which category(ies) of cultivar is the model best at predicting? Is (are) the most accurately predicted category(ies) the one(s) with the most observations? Report the MSE from the test set.

```
from sklearn.cross_validation import train_test_split
from sklearn.metrics import classification_report

X_train, X_test, y_train, y_test = \
    train_test_split(X, y, test_size = 0.25,
        random_state=20)
print(classification_report(y_test, y_pred))
```

(b) Perform a leave-one-out cross validation (LOOCV) with the model from part (a). Report your error rates (1 - precision) for each category? How do your error rates compare to those from part (a)? Report your LOOCV estimate for the test MSE as the average MSE, where y_i is the left out observation from each test set.

$$CV_{loo} = \frac{1}{N}\sum_{i=1}^{N} MSE_i = \frac{1}{N}\sum_{i=1}^{N}\Big[1 - I\left(y_i = \hat{y}_i\right)\Big]$$

(c) Perform a k-fold cross validation in which the data are divided into $k = 4$ groups. Use the following code. Report your error rates (1 - precision) for each category. How do your error rates compare to those from parts (a) and (b)? Report your k-fold estimate for the test MSE as the average MSE.

```
from sklearn.model_selection import KFold

kf = KFold(n_splits=4, shuffle=True, random_state=10)
kf.get_n_splits(X)
```

$$CV_{kf} = \frac{1}{k}\sum_{i=1}^{k} MSE_i \quad \text{where} \quad MSE_i = \frac{1}{n}\sum_{j=1}^{N}\Big[1 - I\left(y_j = \hat{y}_j\right)\Big]$$

2. **Splines and interpolation (4 points).** A survey was conducted in the year 2019 in which a group of 20 American high school students ranked a representative sample of individuals according to coolness that could range from 0 to 100. The result was an aggregated score for each individual. They called this the Coolness Index. The data are in the comma-delimited data file `CoolIndex.txt`, where each row represents an individual in the sample ($N = 956$), the first column is the age of the individual, and the second column is the corresponding aggregated Coolness Index value.[1]

   (a) Create a scatterplot of the data with *age* on the x-axis and *Coolness Index* on the y-axis. Label your axes, and give the plot a title.

   (b) Use ordinary least squares (OLS) regression to fit a stepwise function to these data. Use 5 bins $[11, 22)$, $[22, 40)$, $[40, 59)$, $[59, 77)$, $[77, 95]$. Remember that your dummy variables must be integer type (0, 1), not boolean type (True, False). Plot this step function on top of the scatterplot of the data from part (a). Label your axes, include a legend, and give the plot a title. Report your estimated step function values for each bin $[\beta_1, \beta_2, \beta_3, \beta_4, \beta_5]$. What is the predicted coolness of a 73-year old from the stepwise function?

[1] As can be plainly seen from the year of the survey, these data are fictitious. They are simulated data that I created. One can take great liberties when one is the teacher.

(c) Fit a linear spline (continuous) to the data over the 5 age bins from part (b). Use the `scipy.interpolate.LSQUnivariateSpline` function with $k = 1$ (linear) and the knots equal to $t =$`[22,40,59,77]`. Plot your continuous linear spline against a scatterplot of the data from part (a) and the estimated step function from part (b). Label your axes, include a legend, and give the plot a title. What is the predicted coolness of a 73-year old from the linear spline?

(d) Fit a cubic spline (continuous) to the data over the 5 age bins from part (b). Use the `scipy.interpolate.LQUnivariateSpline` function with $k = 3$ (cubic) and the knots equal to $t =$`[22,40,59,77]`. Plot your continuous cubic spline against a scatterplot of the data from part (a) and the estimated step function from part (b), and the linear spline from part (c). Label your axes, include a legend, and give the plot a title. What is the predicted coolness of a 73-year old from the cubic spline?