Problem Set #5

MACS 30100, Dr. Evans
Due Monday, Feb. 19 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 for $j = 1$ and $j = 2$. Report your error rates (1 - precision) on the test set using the code below. Which category of cultivar is the model best at predicting? Is the most accurately predicted category the one with the most observations?

```
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=3, 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. **Spline and kernel density interpolation (4 points).** [TODO: will finish soon.]