Problem Set #7

MACS 30100, Dr. Evans
Due Monday, Mar. 5 at 11:30am

1. **Classifier "horse" race (10 points).** For this problem, you will use the 397 observations from the Auto.csv dataset.[1] This dataset includes 397 observations on miles per gallon (`mpg`), number of cylinders (`cylinders`), engine displacement (`displacement`), horsepower (`horsepower`), vehicle weight (`weight`), acceleration (`acceleration`), vehicle year (`year`), vehicle origin (`origin`), and vehicle name (`name`). We will study the factors that make miles per gallon high or low. Create a binary variable `mpg_high` that equals 1 if `mpg_high`$\geq$ median(`mpg_high`) and equals either 0 if `mpg_high`$<$ median(`mpg_high`).

   (a) Use `sklearn.linear_model.LogisticRegression` to fit a logistic model of `mpg_high` on features number of cylinders (`cyl`), engine displacement (`dspl`), horsepower (`hpwr`), vehicle weight (`wgt`), acceleration (`accl`), vehicle year (`yr`), vehicle origin (`orgn`). Make sure to include a constant term. Fit the model using k-fold cross validation with $k = 4$ folds.[2]

   ```
   kf_log = KFold(n_splits=4, shuffle=True, random_state=15)
   ```

   Report the MSE of the model as the average MSE across the $k = 4$ test sets, and report the error rates for each category of `mpg_high` as the average error rate for that category across the $k = 4$ test sets.

   $$Pr(mpg\_high = 1|\boldsymbol{X\beta}) = \frac{e^{\boldsymbol{X\beta}}}{1 + e^{\boldsymbol{X\beta}}}$$
   $$\text{where} \quad \boldsymbol{X\beta} = \beta_0 + \beta_1 cyl_i + \beta_2 dspl_i + \beta_3 hpwr_i + \beta_4 wgt_i + \beta_5 accl_i + \beta_6 yr_i + \beta_7 orgn_i$$

   (b) Use `sklearn.ensemble.RandomForestClassifier` to fit a random forest model of `mpg_high` on `max_features=2` out of the seven possible features used in part (a). Set `n_estimators=20`, set `bootstrap=True`, set `oob_score=True`, and set `random_state=25`. Report the MSE of the random forest model as the MSE from the `.oob_prediction_` object, and report the error rates for each category of `mpg_high` from the `.oob_prediction_` object.

   (c) Use `sklearn.svm.SVC` to fit a support vector machines model of `mpg_high` with a Gaussian radial basis function kernel `kernel='rbf'` on the seven features used in part (a). Set the penalty parameter to `C=1` and set `gamma=0.2`. Fit the model using k-fold cross validation with $k = 4$ folds exactly as in part (a).

[1] The Auto.csv dataset comes from James et al. (2017, Ch. 3) and is available at http://www-bcf.usc.edu/ gareth/ISL/data.html.

[2] `sklearn.model_selection.KFold`.

```
kf_svm = KFold(n_splits=4, shuffle=True, random_state=15)
```

Report the MSE of the model as the average MSE across the $k = 4$ test sets, and report the error rates for each category of `mpg_high` as the average error rate for that category across the $k = 4$ test sets.

(d) Which of the above three models do you think is the best predictor of `mpg_high`? Why?

References

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R* Springer Texts in Statistics, Springer, 2017.