Homework 2

Due: 11.59pm on Friday, February 2

Submission instructions:

- This assignment contains two prediction problems. Create a write-up per group explaining what you have tried for these problems. Submit one write-up per group on gradescope.com. Please do not bring printouts of your solutions to the classroom.
- In addition, you will email your predictions as explained below to boothmlteam@gmail.com.

Files needed for this homework can be downloaded here:
https://github.com/ChicagoBoothML/ML2017/tree/master/hw02

Question 1

In a bike sharing system the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. In this problem, you will try to combine historical usage patterns with weather data to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C.

You are provided hourly rental data collected from the Capital Bikeshare system spanning two years. The file `Bike_train.csv`, as the training set, contains data for the first 19 days of each month, while `Bike_test.csv`, as the test set, contains data from the 20th to the end of the month. The dataset includes the following information:

`daylabel`	day number ranging from 1 to 731
`year`, `month`, `day`, `hour`	hourly date
`season`	1 = winter, 2 = spring, 3 = summer, 4 = fall
`holiday`	whether the day is considered a holiday
`workingday`	whether the day is neither a weekend nor a holiday
`weather`	1 = clear, few clouds, partly cloudy 2 = mist + cloudy, mist + broken clouds, mist + few clouds, mist 3 = light snow, light rain + thunderstorm + scattered clouds, light rain 4 = heavy rain + ice pallets + thunderstorm + mist, snow + fog
`temp`	temperature in Celsius
`atemp`	"feels like" temperature in Celsius
`humidity`	relative humidity
`windspeed`	wind speed
`count`	number of total rentals

Predictions will be evaluated using the root mean squared error (RMSE), calculated as

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \hat{m}_i)^2}$$

where m_i is the true count, $\hat{m}_i$ is the estimate, and n is the number of entries to be evaluated.

Build a model to predict the bikeshare counts for the hours recorded in the test dataset. Save your predicted `count` in a file `hw2-1-<your_uchicago_id>.csv`, where you will need to replace `your_uchicago_id` by your UChicago ID. Your file should contain only one column with a header `count` and 6,493 entries of predicted values. A sample submission `hw2-1-mkolar.csv` can be found on Piazza. This sample submission is created by fitting a linear regression, treating every predictor as numeric, and restricting the predicted values to be positive. It has RMSE of 145.78 on the test set.

You should email your submission file to boothmlteam@gmail.com and another file with the code you used to make predictions.

Some tips:

- It will be helpful to examine the data graphically to spot any seasonal pattern or temporal trend.
- There is one day in the training data with weird `atemp` record and another day with abnormal `humidity`. Find those rows and think about what you want to do with them. Is there anything unusual in the test data?
- It *might* be helpful to transform the `count` to $\log(\texttt{count}+1)$. If you did that, do not forget to transform your predicted values back to count.
- Think about how you would include each predictor into the model, as continuous or as categorical?
- Is there any transformation of the predictors or interactions between them that you think might be helpful?

You will receive points based on your write-up, whether we can compute RMSE based on your submission and your relative ranking in the class.

Question 2

The dataset `MovieReview_train.csv` contains information for 5,000 IMDB movie reviews. The first column `length` contains the length of each review, the next 390 columns contain counts of the 390 most frequent words appearing in the reviews. The last column `sentiment`, which is the target variable, is binary, meaning the IMDB rating < 5 results in a sentiment score of 0, and rating $>= 7$ have a sentiment score of 1. The goal of this analysis is to predict the sentiment of the 5,000 unlabelled reviews in the test dataset `MovieReview_test.csv` based on the bag of words.

Predictions will be evaluated using the overall misclassification rate, calculated as

$$\frac{1}{n}\sum_{i=1}^{n} 1\{y_i \neq \hat{y}_i\}$$

where y_i is the true label, $\hat{y}_i$ is the prediction, and n is the number of entries to be evaluated.

Build a model to predict the sentiment for the reviews in the test dataset. Save your predicted `sentiment` (either 0 or 1) in a file `hw2-2-<your_uchicago_id>.csv`, where you will need to replace `your_uchicago_id` by your UChicago ID. Your file should contain only one column with a header `sentiment` and 5,000 entries of predicted labels. A sample submission `hw2-2-mkolar.csv` can be found on Piazza. This sample submission is created by fitting a logistic regression using all the predictors and has a misclassification rate of 21.34% on the test set.

You should email your submission file to boothmlteam@gmail.com and another file with the code you used to make predictions.

You will receive points based on your write-up, whether we can compute the misclassification rate based on your submission and your relative ranking in the class.