CS 189
Fall 2017
Extra Credit

Introduction to Machine Learning
Ashwinee Panda, Amog Kamsetty, Rohan Taori and Gokul Swamy

Instructions: You are welcome to form small groups (up to 4 people total) to work through the homework, but you **must** write up all solutions by yourself. List your study partners for homework on the first page, or "none" if you had no partners.

If using LaTeX (which we recommend), you may use the homework template linked on this Piazza post to get started.

Begin each problem on a new page. Clearly label where each problem and subproblem begin. The problems must be submitted in order (all of P1 must be before P2, etc).

No late homeworks will be accepted. No exceptions. This is not out of a desire to be harsh, but rather out of fairness to all students in this large course. Out of a total of approximately 12 homework assignments, the lowest two scores will be dropped.

Due Friday, September 16 at 11:59am

1. **Where there's ( ), there's fire** For this problem we will be using a data set of forest fires in the northeastern part of Portugal. In the real world you can collect many features for any given data point; in this case, we have 12. Our goal is to find out which indicators best predict the area of a forest fore -or in other words, redefine the old adage: **Where there's smoke, there's fire.**
   This dataset contains twelve features, and we want to predict the total burnt area of the forest. This problem combines multiple concepts we have gone over, including but not limited to: multivariate regression, k-fold cross validation, and ridge regression.
   To check out more information about the twelve features, check out this archive.
   Since we have not worked with categorical data yet, we will ignore the month and the day. Thus, we are left with 10 features in order to predict the area of the forest fire. We will also truncate the number of samples to 515 instead of 517.

   (a) First load the data, and generate the X matrix consisting of the features and $\vec{y}$ vector consisting of the observations of the area.
   We'll be using degree $D = 1$ for now. **Report the dimensions of X and $\vec{y}$.**

   (b) Now, split the data into training and test sets. First, make sure to shuffle the data. Then, we will do k-fold training-test split with $k = 5$. Report the dimensions of $X_{train}, X_{test}, \vec{y}_{train}$, and $\vec{y}_{test}$. Explain why we don't need a validation set for now.
   Now apply linear regression (OLS) to predict the weights for the features, to find the vector $\vec{w}$ which minimizes the error. **Also calculate the test error:** $\frac{1}{n}\Sigma_{i=1}^{n}||(\vec{y}_{test_i} - \vec{x_{test_i}}^T\vec{w})||_2^2$**.**
   From these folds, you should obtain 5 test errors. Report the average of the 5 test errors.

   (c) Now, rather than considering all 10 features, let us only consider the features relating to weather conditions (temp, RH, wind, and rain). Repeat parts a and b with just these 4 features, and keeping $D = 1$. **Report your test error.**
   We clearly see that the test error is lower if we only consider the weather conditions, rather than all the features. This is because some of the features in the dataset may be uncorrelated to the area of the wildfire, so we don't want to fit to these features. Now consider if we had n samples where $\lim_{x\to\infty}$. **Would we still have the same problem? Why or why not?**

   (d) Continuing with just using the weather condition variables, let us now perform ridge regression. Using grid search, determine which degree out of $1-5$, and which λ out of $\{0.01, 0.1, 1, 10, 100\}$ minimizes the test error. **Report which pair of hyperparameters produces the best result.**

   (e) Now we will add another hyperparameter: The number of features we use! Using the same D and λ values we got in the previous part, let's use validation to determine which set of features give the best model. We will always have the 4 weather conditions, but the hyperparameter will be how many other features we use out of the 6 that are remaining. **Determine what value for this hyperparameter minimizes the test error.**