Math 189r
Homework 1
September 19, 2016

There are 8 problems in this set. You must complete 6 (doing more will get no credit - work on your project!) 3 of the problems (you choose) are due on September 12, and the rest of the problems you complete are due on September 19. Feel free to work with other students, but make sure you write up the homework and code on your own (no copying homework *or* code; no pair programming). Feel free to ask students or instructors for help debugging code or whatever else, though. When implementing algorithms you may not use any library (such as `sklearn`) that already implements the algorithms but you may use any other library for data cleaning and numeric purposes (`numpy` or `pandas`). Use common sense. Problems are in no specific order.

1 (**regression**). Download the data at `https://math189r.github.io/hw/data/online_news_popularity/online_news_popularity.csv` and the info file at `https://math189r.github.io/hw/data/online_news_popularity/online_news_popularity.txt`. Read the info file. Split the csv file into a training and test set with the first two thirds of the data in the training set and the rest for testing. Of the testing data, split the first quarter into a 'validation set' (used to optimize hyperparameters while leaving your testing data pristine) and the remaining 3/4 as your test set. We will use this data for the remainder of the problem. The goal of this data is to predict the number of shares a news article will have given the other parameters.

(a) (**implementation**) A K-Nearest Neighbor classifier/regressor takes the k 'closest' points to a test point x according to some distance metric (often the L-2 norm) and predicts that the label/value at x to be the mean of the labels of the k closest points. Using the Euclidean distance as your metric, write a KNN classifier in the language of your choosing. Try a bunch of different values of k between 1 and 100, record the root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2}$$

on the validation set, and choose that value of k as optimal. What is the RMSE on the test set?

(b) (**exploration**) Find a new distance metric that performs better on this dataset (using the same process as above for choosing k). What is the new metric you used (as well as others you tried that may have not worked better) and what is your new RMSE? Why should this metric work better?

(c) (**implementation**) Attempt the same problem using ridge regression. Find the optimal regularization parameter λ from the validation set. What is the RMSE on the

validation set for all values of λ you tried and what is the final RMSE on the test set with the optimal $\lambda^\star$?

(d) (**math**) See section 7.6 of Murphy (you should have read this already) on Bayesian Linear Regression. Prove/derive everything in 7.6.1 and 7.6.2:

(a) $\mathbb{P}(\mathbf{y}|X, \mathbf{w}, \sigma^2) \propto \exp(-\frac{1}{2\sigma^2}||\mathbf{y} - \overline{\mathbf{y}}\mathbf{1}_N - X\mathbf{w}||_2^2)$

(b) $\mathbb{P}(\mathbf{x}|X, \mathbf{y}, \sigma^2) \propto \mathcal{N}(\mathbf{w}|\mathbf{w}_0, V_0)\mathcal{N}(\mathbf{y}|X\mathbf{w}, \sigma^2 I_N) = \mathcal{N}(\mathbf{w}|\mathbf{w}_N, V_N)$ where

$$\begin{aligned}
\mathbf{w}_N &= V_n V_0^{-1}\mathbf{w}_0 + \frac{1}{\sigma^2}V_N X^\top \mathbf{y} \\
V_N^{-1} &= V_0^{-1} + \frac{1}{\sigma^2}X^\top X, \text{ and} \\
V_N &= \sigma^2(\sigma^2 V_0^{-1} + X^\top X)^{-1}.
\end{aligned}$$

(c) Show that when we let $\mathbf{w}_0 = \mathbf{0}$ and $V_0 = \tau^2 I$, the posterior mean becomes the ridge regression estimate with $\lambda = \frac{\sigma^2}{\tau^2}$.

(d) Show

$$\begin{aligned}
\mathbb{P}(y|\mathbf{x}, \mathcal{D}, \sigma^2) &= \int \mathcal{N}(y|\mathbf{x}^\top \mathbf{w}, \sigma^2)\mathcal{N}(\mathbf{w}, \mathbf{w}_N, V_N)d\mathbf{w} \\
&= \mathcal{N}(y|\mathbf{x}_N^\top \mathbf{x}, \sigma_N(\mathbf{x})) \\
\sigma_N^2(\mathbf{x}) &= \sigma^2 + \mathbf{x}^\top V_N \mathbf{x}.
\end{aligned}$$