

## B555: Homework 4

1. The standard regression model is

$$y = Xw + e$$

where  $X$  is an  $n \times d$  matrix of predictor variables,  $y$  is an  $n$ -vector of response variables, and  $e \sim N(0, \sigma^2 I)$  is an  $n$ -vector of model errors. Suppose also in this case that  $w \sim N(0, \rho^2 I)$

- (a) What is the density of  $y$  in terms of  $X, w, \sigma^2$ ?
- (b) Write down the joint density function  $p(w, y)$
- (c) Show that

$$\hat{w}_{\text{MAP}} = \arg \min_w \frac{\|w\|^2}{\rho^2} + \frac{\|y - Xw\|^2}{\sigma^2}$$

- (d) How does  $\hat{w}_{\text{MAP}}$  compare with ridge regression?

2. Consider the regression model with  $e_1, e_2, \dots, e_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$ . Derive an unbiased estimate for  $\mu$  in this model.
3. The class website has two data sets for this homework assignment. The first has an  $n \times d = 1000 \times 50$  data matrix ( $X$ ) “pred1.dat” with a  $1000 \times 1$  response vector ( $y$ ) “resp1.dat.” The second has a  $1000 \times 500$  data matrix “pred2.dat” with a response vector “resp2.dat.” These data sets were generated according to the standard linear regression model, as described above.
- (a) For each data set use the first half of the data (observations  $i = 1, \dots, n/2$ , all  $d$  predictors) to get the minimum variances unbiased estimate of  $w, \hat{w}$ .
  - (b) For the first data set using only the first half of the data compute and report an unbiased estimate of  $\sigma^2$ . Describe any difficulties you would have in performing the same task for the 2nd data set.
  - (c) For each data set, use your estimate of  $w$  on the 2nd half of the data set ( $n/2 + 1, \dots, n$ ), to get your estimated response variables,  $\hat{y}$  and compute and report your total squared error:

$$SSE = \sum_{i=n/2+1}^n (\hat{y}_i - y_i)^2$$

4. In *variable selection* we iteratively grow a collection of variables that are used predictor variables. We begin by finding the single predictor variable that gives the smallest value of  $SSE = \sum (y_i - \hat{y}_i)^2$ , making this our first predictor variable. We then add another predictor variable (column of our data matrix) by again choosing the variable that, when combined with our first variable, gives the smallest value of  $SSE$ . In our simple implementation we implement this greedy search until we have have a fixed number of variables.
- (a) Implement variable selection on the first half of the first data set to identify the 3 best predictor variables. Report the three variables you get and the three decreasing values of  $SSE$  they produce.
  - (b) Compute the value for  $SSE$  on the 2nd half of the data set using the model you have learned. Compare this SSE with that obtained using all predictor variables. Which approach gives a better SSE and why?
5. Use the first half of the 2nd data set to perform ridge regression on  $w$  using a parameter of  $\lambda = 20$  to get a new  $\hat{w}$ .
- (a) Using your new  $\hat{w}$ , compute the estimated response,  $\hat{y}$ , for the 2nd half of the dataset.
  - (b) Compare the resulting sum of squared errors on the 2nd half of the dataset, using both ridge regression and plain regression. If one of the two methods performs better, explain why.
  - (c) In general, we do not know what the best choice of the smoothing parameter,  $\lambda$  will be. One way to choose the parameter would be to try a variety of values estimated using the first half of the dataset, choosing the value that gives the best performance on the 2nd half of the dataset. This is known as cross validation. Use this idea to estimate your best choice of smoothing parameter,  $\lambda$ .

6. A time series is a sequence of observations taken over time, usually with constant time between measurements. The data on the website “time\_series.dat” was taken from a time series model

$$x_i = \alpha_1 x_{i-1} + \alpha_2 x_{i-2} + e_i$$

where the  $\{e_i\}$  are modeled as independent  $N(0, \sigma^2)$  random variables. Estimate the parameters  $\alpha_1, \alpha_2, \sigma^2$  from these data.