

BSCB 694: Statistical Machine Learning  
**Homework 1**  
**Due: Tuesday, September 18**

1. Text, Ex 2.4

(Parenthetical remark: note that when  $X$  is distributed as a standard normal on  $d$ -space,  $\|X\|$  is the square root of a  $\chi_d^2$  random variable. It can be shown that as  $d \rightarrow \infty$ , this is distributed with mean  $\sqrt{d-1/2}$  and variance  $1/2$ , meaning that nearly all the mass in the distribution is contained in a shell of width  $2, \sqrt{d}$  units away from the mean. )

2. The Dimension of a Data Set<sup>1</sup>

Suppose that we have measured a set of features  $x_i \in \mathbb{R}^n$ , but we suspect that there may be complex dependencies in the data, so that all the  $x_i$  lie on a lower dimensional manifold. We want to estimate how many "real" dimensions there are in the data.

In  $\mathbb{R}^d$  a sphere of radius  $r$  has volume proportional to  $r^d$ . By analogy<sup>2</sup> we define the dimension of a manifold at a point  $x$  to be

$$d = \lim_{r \rightarrow 0} \frac{\ln(S(r, x))}{\ln(r)}$$

where  $S(r, x)$  is a measure of the volume of the manifold that falls in sphere of radius  $r$  centered at  $x$ . This definition is given so long as this limit exists.

When we are using a data set  $x_1, \dots, x_N$ , we can approximate  $S(r, x)$  by

$$S_N(r, x) = \frac{1}{N} \sum_{i=1}^N I(\|x_i - x\| \leq r)$$

since  $x$  is required to be on the manifold, we will average this over  $x$  to obtain

$$\bar{S}_N(r) = \frac{1}{N^2} \sum_{i,j=1}^N I(\|x_i - x_j\| \leq r)$$

and get an estimate

---

<sup>1</sup>The methods proposed here are due to Camastra and Vinciarelli, 2002. Almost the same methods are advocated in Guckenheimer and Clewely, 2007, for estimating the dimension of an attractor manifold.

<sup>2</sup>This can be formalized in terms of box-counting dimensions.

$$\hat{d} = \lim_{r \rightarrow r_0} \frac{\ln(\bar{S}_N(r))}{\ln(r)}$$

where the limit is estimated by the slope of  $\ln(\bar{S}_N(r))$  plotted against  $\ln(r)$  as  $r$  becomes small. The limit  $r_0$  is used since the estimate will be zero for  $r$  small enough.

Observe that  $\bar{S}_N(r)$  may be calculated efficiently by first calculating all the inter-point distances and then sorting them.

- (a) The Boston Housing Data is linked on the class web-site. Use the technique above to estimate the intrinsic dimensionality of its continuous feature variables (ie, not the Charles River dummy variable or the response).

Note that  $\bar{S}_N(r)$  will only be a "reasonable" estimate for an interval of  $r$  values. If  $r$  is too large, almost all the data will be inside each sphere; when  $r$  gets too small  $\bar{S}_N(r)$  will become 1. Justify the set of values you used to calculate  $\hat{r}$ ; this may include exploring the results over the values you considered.

- (b) A good estimate  $\hat{d}$  may require  $N$  to be very large. To check the accuracy of your estimator, generate 506 uniformly distributed points on the 5-dimensional unit cube. Use these to estimate the dimensionality of this data set. Repeat 20 times to provide a mean and variance of the estimator.

### 3. Ridge Regression

- (a) Text, Ex 3.10, Ex 3.17
- (b) Consider making a prediction at a new point  $x^*$  based on a ridge-regression with smoothing parameter  $\lambda$ :  $\hat{y} = x^* \beta_\lambda^{ridge}$ .
- Derive explicit expressions for the bias and variance of  $\hat{y}$  as a function of  $\lambda$ .
  - Set  $MSE(\lambda) = bias(\lambda) + var(\lambda)$  from above, show that

$$\left. \frac{d}{d\lambda} MSE(\lambda) \right|_{\lambda=0} < 0$$

hence prediction error is always improved by some regularization. You may find the following identity helpful: if  $A(\theta)$  is a matrix-valued function of a univariate parameter  $\theta$ , then

$$\frac{d}{d\theta} [A(\theta)]^{-1} = -[A(\theta)]^{-1} \frac{dA(\theta)}{d\theta} [A(\theta)]^{-1}$$

4. Text, Ex 7.4, 7.5

## 5. Bias, Variance and Variable Selection

It is an easy see from the calculations in Question 4 that ridge regression always reduces the variance of the prediction  $x\beta_\lambda^{ridge}$ , here we examine variable selection.

Consider a two feature problem  $x_i = (x_{1i}, x_{2i})$  in which we are interested in selecting one feature. For the sake of simplicity, assume that the experiment is designed so that  $X^T X$  is the identity. Also take  $\beta_1 = \beta_2$  and the error variance to be  $\sigma$ .

This has been set up carefully so that each variable is selected with probability  $1/2$ , so long as the selection criterion is independent of the feature index. Note that we are also not using an intercept term.

We define  $\beta^{select}$  as the coefficients estimated by least squares after variable selection has taken place. In other words, this is  $(\hat{\beta}_1, 0)$  if variable 1 was selected, otherwise  $(0, \hat{\beta}_2)$ .

- (a) Using the law of total variation, derive the variance of the prediction  $x^* \beta^{select}$ .
- (b) What is the variance of  $x^* \hat{\beta}$  when  $\hat{\beta}$  is the least squares estimate? Derive conditions in terms of  $\sigma$  and  $\beta$  under which the variance for the variable selection estimate is less than that for the least squares estimate.
- (c) Now consider total mean squared error. Under what circumstances (in terms of  $\beta$  and  $\sigma$ ) will the following be optimal:
  - i. Predict 0.
  - ii. Use the variable selection estimate  $x^* \beta^{select}$ .
  - iii. Use the least squares estimate  $x^* \hat{\beta}$ .

Use a fixed  $x^*$  throughout these calculations.