

# 10-702 Statistical Machine Learning: Assignment 4

Due Friday, March 19

Complete three of the following four problems. Hand in to Sharon Cavlovich, GHC (Gates Hillman Center) 8215 by 3:00. Use R for all numerical computations.

## 1. (Nonparametric density estimation)

In this exercise, you are going to analyze a kernel estimator of the  $s$ -th derivative  $p^{(s)}$  of a density  $p \in \Sigma(\beta, L)$  (density  $p$  belongs to the Hölder class  $(\beta, L)$ ),  $s < \beta$ , which is defined as

$$\hat{p}_{n,s}(x) = \frac{1}{nh^{s+1}} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right).$$

The bandwidth  $h > 0$  and  $K : \mathbb{R} \mapsto \mathbb{R}$  is a bounded kernel on a compact support  $[-1, 1]$  and  $X_i \in \mathbb{R}$  is a one dimensional random variable with the density  $p$ . Assume that the kernel satisfies

$$\begin{aligned} \int u^j K(u) du &= 0, \quad j = 0, 1, \dots, s-1, s+1, \dots, \lfloor \beta \rfloor \\ \int u^s K(u) du &= s! \end{aligned} \tag{1}$$

(a) Given a point  $x_0 \in \mathbb{R}$ , prove that

$$\sup_{p \in \Sigma(\beta, L)} \mathbb{E}[\hat{p}_{n,s}(x_0) - p^{(s)}(x_0)] \leq ch^{\beta-s},$$

where  $c > 0$  is an appropriate constant.

(b) Given a point  $x_0 \in \mathbb{R}$ , prove that

$$\sup_{p \in \Sigma(\beta, L)} \mathbb{E}[(\hat{p}_{n,s}(x_0) - \mathbb{E}[\hat{p}_{n,s}(x_0)])^2] \leq c'(nh^{2s+1})^{-1},$$

where  $c' > 0$  is an appropriate constant.

(c) Combine results from (a) and (b) to show that

$$\sup_{x_0 \in \mathbb{R}} \sup_{p \in \Sigma(\beta, L)} \mathbb{E}[(\hat{p}_{n,s}(x_0) - p^{(s)}(x_0))^2] \leq Cn^{-\frac{2(\beta-s)}{2\beta+1}} \text{ as } n \rightarrow \infty,$$

if the bandwidth  $h = h_n$  is chosen optimally.

(d) Let  $\{\phi_m\}_{m=0}^{\infty}$  be the orthonormal Legendre basis on  $[-1, 1]$ . Show that the kernel

$$K(u) = \sum_{m=0}^{\lfloor \beta \rfloor} \phi_m^{(s)}(0) \phi_m(u) \mathbb{I}\{|u| \leq 1\}$$

satisfies conditions (1).

The orthonormal Legendre basis  $\{\phi_m\}_{m=0}^\infty$  on  $[-1, 1]$  are defined as:

$$\phi_0(x) \equiv \frac{1}{\sqrt{2}}, \quad \phi_m(x) = \sqrt{\frac{2m+1}{2}} \frac{1}{2^m m!} \frac{d^m}{dx^m} [(x^2 - 1)^m], \quad m = 1, 2, \dots$$

for  $x \in [-1, 1]$ . The following property will be useful

$$\int_{-1}^1 \phi_j(x) \phi_k(x) dx = \delta_{jk},$$

where  $\delta_{jk}$  is the Kronecker delta, i.e.,  $\delta_{jk} = 1$  if  $j = k$  and  $\delta_{jk} = 0$  otherwise.

## 2. (Orthogonal series estimators)

Consider the following nonparametric regression model

$$Y_i = f(X_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where  $f$  is a mapping  $[0, 1] \mapsto \mathbb{R}$ . The random variables  $\epsilon_i$  are independent, with  $\mathbb{E}[\epsilon_i] = 0$  and  $\mathbb{E}[\epsilon_i^2] = \sigma^2 < \infty$ , and  $X_i = \frac{i}{n}$ . Suppose  $f \in L_2[0, 1]$  (recall the definition from Section 27.9) and let  $\theta_j$  be the Fourier coefficients of  $f$  with respect to the trigonometric basis  $\{\varphi_j\}_{j=1}^\infty$ :

$$\theta_j = \int_0^1 f(x) \varphi_j(x) dx.$$

Recall that the trigonometric basis forms an orthonormal basis for  $L_2[0, 1]$  and are given as

$$\begin{aligned} \varphi_1(x) &\equiv 1 \\ \varphi_{2k}(x) &= \sqrt{2} \cos(2\pi kx), \\ \varphi_{2k+1}(x) &= \sqrt{2} \sin(2\pi kx), \quad k = 1, 2, \dots \end{aligned}$$

Assume that  $f$  can be represented as  $f(x) = \sum_{j=1}^\infty \theta_j \varphi_j(x)$ , where the Fourier coefficients satisfy

$$\sum_{j=1}^\infty |\theta_j| < \infty.$$

Note that the above implies that the series  $\sum_{j=1}^\infty \theta_j \varphi_j(x)$  is absolutely convergent on  $x \in [0, 1]$ , so that the pointwise representation of  $f$  holds.

The orthogonal series estimator approximates  $f$  with  $f_N(x) = \sum_{j=1}^N \theta_j \varphi_j(x)$ , where the coefficients  $\theta_j$  are estimated with marginal regression as

$$\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^n Y_i \varphi_j(X_i).$$

and the resulting estimator is  $\hat{f}_{nN}(x) = \sum_{j=1}^N \hat{\theta}_j \varphi_j(x)$ . In this exercise, you will analyze some properties of  $\hat{f}_{nN}$ .

(a) Show that for the trigonometric basis  $\{\varphi_j\}_{j=1}^\infty$  the following holds

$$\frac{1}{n} \sum_{s=1}^n \varphi_j(s/n) \varphi_k(s/n) = \delta_{jk}, \quad 1 \leq j, k \leq n-1.$$

Define the remaining error

$$\alpha_j = \frac{1}{n} \sum_{i=1}^n f(i/n) \varphi_j(i/n) - \int_0^1 f(x) \varphi_j(x) dx, \quad j \geq 1,$$

which arises from approximation of the integral with the sum.

(b) Show that  $\mathbb{E}[\widehat{\theta}_j - \theta_j] = \alpha_j$ .

(c) Show that  $\mathbb{E}[(\widehat{\theta}_j - \mathbb{E}[\widehat{\theta}_j])^2] = \frac{\sigma^2}{n}$ . You may use the result from (a) even if you did not prove it.

(d) Show that  $\mathbb{E}[(\widehat{\theta}_j - \theta_j)^2] = \frac{\sigma^2}{n} + \alpha_j^2$ .

(e) Show that the risk of the orthogonal series estimator can be expressed as

$$\mathbb{E} \left[ \int_0^1 (\widehat{f}_{nN}(x) - f(x))^2 dx \right] = \frac{\sigma^2 N}{n} + \sum_{j=1}^N \alpha_j^2 + \sum_{j=N+1}^{\infty} \theta_j^2.$$

You may use the result from (a)-(d) even if you did not prove them.

Next, we you will show how to upper bound the remaining error  $\alpha_j$ .

(f) Show that  $\max_{1 \leq j \leq n-1} |\alpha_j| \leq 2 \sum_{m=n}^{\infty} |\theta_m|$ , for all  $n \geq 2$ . You may use the result from (a) even if you did not prove it.

Finally, you will find an upper bound on the risk when the function  $f$  is sufficiently smooth. Define the set

$$\Theta(\beta, Q) := \left\{ \theta = (\theta_1, \dots, \theta_i, \dots) : \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq Q \right\},$$

where  $a_j = j^\beta$ , for even  $j$ , and  $a_j = (j-1)^\beta$ , for odd  $j$ . You will consider functions  $f = \sum_j \theta_j \varphi_j$  for which  $\theta \in \Theta(\beta, Q)$ .

(g) Show that if  $\theta \in \Theta(\beta, Q)$ ,  $\beta > 1/2$ , then  $\max_{1 \leq j \leq n-1} |\alpha_j| \leq C n^{-\beta+1/2}$  for all  $n \geq 2$ , where  $C < \infty$  is a constant that depends only on  $\beta$  and  $Q$ . Use result from (f).

(h) Show that if  $N = \lfloor \alpha n^{\frac{1}{2\beta+1}} \rfloor$ , for  $\alpha > 0$ , then the orthogonal series estimator satisfies

$$\sup_{\theta \in \Theta(\beta, Q)} \mathbb{E} \left[ \int_0^1 (\widehat{f}_{nN}(x) - f(x))^2 dx \right] \leq C n^{-\frac{2\beta}{2\beta+1}},$$

where  $C < \infty$  is a constant that depends only on  $\beta, Q$  and  $\alpha$ . To show this, you will need to upper bound  $\sum_{j=1}^N \alpha_j^2$  and  $\sum_{j=N+1}^{\infty} \theta_j^2$  from part (e).

### 3. (Bagging)

- (a) In class, we noted that bagging does not improve linear procedures. Here is an example. Let  $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ . Let  $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ . Let  $Y_1^*, \dots, Y_n^*$  and  $Z_1^*, \dots, Z_n^*$  be two bootstrap samples. (Thus,  $Y_1^*, \dots, Y_n^*$  are drawn randomly from  $X_1, \dots, X_n$  with replacement and similarly for  $Z_1^*, \dots, Z_n^*$ .) Let  $\bar{Y}^* = n^{-1} \sum_{i=1}^n Y_i^*$  and  $\bar{Z}^* = n^{-1} \sum_{i=1}^n Z_i^*$ . Finally, define  $\hat{\mu}_{\text{bag}} = (\bar{Y}^* + \bar{Z}^*)/2$ . Find the mean and variance of  $\hat{\mu}_{\text{bag}}$ . Compare the variance of  $\hat{\mu}_{\text{bag}}$  to the variance of  $\bar{X}$ .
- (b) To see how bagging can reduce the variance for nonlinear procedures consider the following example. Let  $X_1, \dots, X_n \sim N(\mu, 1)$ . Define

$$g(x) = I_{(-\infty, \bar{X}]}(x)$$

where  $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ .

- (1) Find the limit of the mean and variance of  $g(\mu)$  as  $n \rightarrow \infty$ .
- (2) Now define  $G(x) = B^{-1} \sum_{j=1}^B g_j^*(x)$  where  $g_j^*(x) = I_{(-\infty, \bar{X}_j^*]}(x)$  and  $\bar{X}_j^*$  is the mean of the  $j^{\text{th}}$  bootstrap sample. Compute the limiting mean and variance of  $G(\mu)$ . You may use the fact (from bootstrap theory) that  $G(x) = \Phi(\sqrt{n}(x - \bar{X})) + o_P(1)$ .

### 4. Consider the following model from Mease and Wyner (2008):

$$\mathbb{P}(Y = 1|X = x) = q + (1 - 2q) I\left(\sum_{j=1}^J x_j > (j/2)\right)$$

where  $0 \leq q \leq 1/2$ ,  $J \leq d$ ,  $X \sim \text{Uniform}[0, 1]^d$ .

- (a) Find the risk of the Bayes rule.
- (b) Generate  $n = 200$  training data points and  $n = 200$  test points from the model. Use  $d = 100$ ,  $J = 2$  and  $q = 1/4$ . Compare the prediction error of the following methods: a classification tree, naive Bayes, the plugin classifier using kernel regression, the plugin classifier using an additive model, a random forest. An example of how to fit a random forest in R is as follows:

```
install.packages("randomForest")
library(randomForest)
data(iris)
help(iris)
help(randomForest)
out = randomForest(Species~., data=iris, importance=TRUE, proximity=TRUE)
print(out)
round(importance(out), 2)
tmp = cmdscale(1 - out$proximity, eig=TRUE)
op = par(pty="s")
```

```
pairs(cbind(iris[,1:4], tmp$points), cex=0.6, gap=0,  
      col=c("red", "green", "blue")[as.numeric(iris$Species)],  
      main="Iris Data: Predictors and MDS of Proximity Based on RandomForest")  
treesize(out)  
varImpPlot(out)
```