vrije Universiteit *amsterdam*

School of Business and Economics

Exam:	Data Science Methods
Code:	E_EOR2_DSM
Examinator:	Paolo Gorgi
Co-reader:	Charles Bos
Date:	-
Time:	-
Duration:	2 hours
Calculator allowed:	Yes
Graphical calculator allowed:	No
Number of questions:	4
Type of questions:	Open
Answer in:	English

Remarks:

- Read carefully the questions before answering.
- Provide clear and complete answers to the questions with concise explanations.
- The question sheet should be handed back at end of the exam.

Credit score:	100 credits counts for a 10
Grades:	-
Inspection:	-
Number of pages:	6 (including front page)

Good luck!

(*This page is intentionally left blank.*)

Question 1 [20 points] Non-parametric density estimation

Consider a sample of iid observation generated by an unknown density function $f(x)$. The kernel density estimator $\hat{f}_h(x)$ of the unknown density $f(x)$ is given by

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where $K(\cdot)$ is a kernel function.

(a) Discuss which properties the kernel function $K(\cdot)$ needs to satisfy in order to ensure that the kernel density estimator $\hat{f}_h(x)$ is a density function. Justify your answer.

Answer:
The kernel density estimator is a density function if $\hat{f}_h(x) \geq 0$, for any $x \in \mathbb{R}$, and $\int_{-\infty}^{+\infty} \hat{f}_h(x)dx = 1$. In order to ensure that these properties are satisfied $K(x)$ has to be a density function itself, that is $K(x) \geq 0$, for any $x \in \mathbb{R}$, and $\int_{-\infty}^{+\infty} K(x)du = 1$. Below we formally show that this is the case. First, we note that $\hat{f}_h(x) \geq 0$ is satisfied if $K(x) \geq 0$ since $K\left(\frac{x-X_i}{h}\right) \geq 0$ given that $K(x) \geq 0$ for any $x \in \mathbb{R}$. As concerns $\int_{-\infty}^{+\infty} \hat{f}_h(x)dx = 1$, we have that

$$\int_{-\infty}^{+\infty} \hat{f}_h(x)dx = \frac{1}{nh}\sum_{i=1}^{n} \int_{-\infty}^{+\infty} K\left(\frac{x - X_i}{h}\right) dx. \tag{1}$$

We can solve the integral $\int_{-\infty}^{+\infty} K\left(\frac{x-X_i}{h}\right) dx$ by substitution. In particular, we set $u = (x - X_i)/h$, which means $x = hu + X_i$, and obtain

$$\begin{aligned}\int_{-\infty}^{+\infty} K\left(\frac{x - X_i}{h}\right) dx &= \int_{-\infty}^{+\infty} K(u)\frac{\partial x}{\partial u}du \\ &= h\int_{-\infty}^{+\infty} K(u)\,du = h,\end{aligned}$$

where $\int_{-\infty}^{+\infty} K(u)\,du = 1$ since $K(u)$ is a density function. Therefore, using this result together with equation (1) we obtain that

$$\int_{-\infty}^{+\infty} \hat{f}_h(x)dx = \frac{1}{nh}\sum_{i=1}^{n} h = 1.$$

(b) We know that approximate expressions for the *Variance* and *Bias* of $\hat{f}_h(x)$ are

$$Var\big(\hat{f}_h(x)\big) \approx \frac{1}{nh}\|K\|_2^2 f(x), \qquad Bias\big(\hat{f}_h(x)\big) \approx \frac{h^2}{2}f''(x)\mu_2(K),$$

where $\|K\|_2^2 = \int_{-\infty}^{\infty} K^2(u)du$, $\mu_2(K) = \int_{-\infty}^{\infty} u^2K(u)du$ and $f''(x) = \frac{\partial^2 f(x)}{\partial x^2}$. Explain the trade-off between *Variance* and *Bias* in the selection of the bandwidth parameter h.

Answer:
The mean squared error MSE of $\hat{f}_h(x)$ is

$$MSE\big(\hat{f}_h(x)\big) = Var\big(\hat{f}_h(x)\big) + Bias\big(\hat{f}_h(x)\big)^2.$$

For a fixed n, we wish to select h that minimizes the MSE. Here we notice that there is a trade-off between bias and variance in the selection of h. This trade-off is illustrated as follows. The expression of the bias shows that, for a fixed n, $Bias\big(\hat{f}_h(x)\big)^2$ decreases as h decreases. This makes sense since a small value of h means that only data points closed to x are used and therefore this will lead to a small bias. However, on the contrary, we see that $Var\big(\hat{f}_h(x)\big)$ decreases as h increases. This is also intuitive since the variance will be smaller when h is large since more observations are used and therefore there will be less sampling uncertainty in the local estimate. This trade-off leads to a selection of h that compromises between bias and variance.

Question 2 [40 points] Univariate non-parametric regression

Consider a univariate regression of Y_i on X_i of the form

$$Y_i = m(X_i) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $m(x)$ is a non-parametric unknown function and ϵ_i is an error term such that $E(\epsilon_i|X_i = x) = 0$ and $Var(\epsilon_i|X_i = x) = \sigma^2$.

(a) Show that the Nadaraya-Watson estimator $\hat{m}_h(x)$, given by

$$\hat{m}_h(x) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{s=1}^n K\left(\frac{x-X_s}{h}\right)},$$

interpolates all data points as $h \to 0$, i.e. $\lim_{h\to 0} \hat{m}_h(X_k) = Y_k$.

Answer:
First we note that the limit for $h \to 0$ of $K\left(\frac{x-X_i}{h}\right)$ is

$$\lim_{h\to 0} K\left(\frac{x - X_i}{h}\right) = \begin{cases} 0 & \text{if } x \neq X_i \\ K(0) & \text{if } x = X_i. \end{cases}$$

Therefore, assuming that $x = X_k$ and $h \to 0$, the Nadaraya-Watson estimator $\hat{m}_h(X_k)$ becomes

$$\begin{aligned} \lim_{h\to 0} \hat{m}_h(X_k) &= \lim_{h\to 0} \frac{\sum_{i=1}^n K\left(\frac{X_k-X_i}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{X_k-X_i}{h}\right)} \\ &= \frac{K(0)Y_k}{K(0)} = Y_k. \end{aligned}$$

This shows that the Nadaraya-Watson regression curve $\hat{m}_h(x)$ interpolates all data points as $h \to 0$.

(b) Explain why minimizing the Residuals Sum of Squares (RSS)

$$RSS(h) = \sum_{i=1}^n \left(Y_i - \hat{m}_h(X_i)\right)^2$$

with respect to h is not a viable way to select the optimal bandwidth. Discuss how the optimal bandwidth h can be selected.

Answer:
As shown in the previous question, we have that $m_h(X_i) \to Y_i$ as $h \to 0$. Therefore, we obtain that $RSS(h) \to 0$ as $h \to 0$. This means that minimizing $RSS(h)$ will always lead to $h = 0$ and therefore overfitting. The problem is that in the derivation of the RSS we are using Y_i to predict itself, since Y_i is used in the estimation of $\hat{m}_h(x)$. We can solve this problem by leave-one-out cross validation. The idea of

cross validation is to derive the prediction of Y_i by leaving out Y_i in the estimation of $m(X_i)$. In particular, the leave-one-out estimator is

$$\hat{m}_{h,-i}(X_i) = \frac{\sum_{j \neq i} K_h(X_i - X_j) Y_j}{\sum_{j \neq i} K_h(X_i - X_j)},$$

where ith observation is left out from the summations. Given the leave-one-out estimate, we can obtain the optimal h minimizing the following cross validation criterion

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \hat{m}_{h,-i}(X_i)\big)^2.$$

(c) Consider the following R code to derive the leave-one-out cross validation criterion for the Nadaraya-Watson estimator, which can be obtained using the R function `ksmooth()`. Note that, in the code below, the vector `x` contains the regressor and `y` the variable of interest.

```
+ mcv <- rep(0,n)
+ for(i in 1:n){
+ mcv[i] <- ksmooth(x[-i], y[-i], kernel="normal", bandwidth=h, x.points=x[i])$y
+ }
+ cv <- mean((y-mcv)^2)
```

Explain what the R code is doing. How would you use the code given above to derive the optimal bandwidth h? Your answer may contain a pseudo R code to explain the last part.

Answer:
The code creates a vector `mcv` of length n where the leave-one-our cross validation predictions $\hat{m}_{h,-i}(X_i)$ will be stored. The *for loop* computes $\hat{m}_{h,-i}(X_i)$ for $i = 1, \ldots, n$ using the R function `ksmooth()` for a given value of the bandwidth parameter `h`. In particular, $\hat{m}_{h,-i}(X_i)$ is computed leaving out the ith observation `x[-i]` and `y[-i]` from the dataset, computing the prediction at $x = X_i$, `x.points=x[i]`, and considering a Gaussian kernel. Finally, the last line of code computes the cross-validation criterion $CV(h)$.
The code given above can be used to create an R function that computes the cross validation criterion $CV(h)$. This R function can then be minimized with respect to h to obtain the optimal h. The minimization can be done using a numerical optimizer, such as `optim()`, or through a grid search. We write pseudo R code to create the function

```
cv_fun <- function(h,x,y){
n <- length(y)
''code given above''
return(cv)
}
```

and to optimize it

```
h_ini <- 0.2
h_min <- optim(h_ini, function(h) cv_fun(h,x,y), ...)
```

(d) Assume now that we are interested in estimating the regression model by *regression splines*. Discuss the difference between *cubic splines (truncated power basis)* and *natural cubic splines*.

Answer:
A spline function is constructed by splitting the range of values of x using K knots $\xi_1, \ldots, \xi_K$, which are such that $\xi_1 < \xi_2 < \cdots < \xi_K$. A *cubic spline (truncated power basis)* is a picewise cubic polynomial that is continuous at each knot. A Similarly, also a *natural cubic spline* is a picewise cubic polynomial that is continuous at each knot. However, *natural cubic spline* is imposed to be linear beyond the boundary knots ξ_1 and ξ_K. In practice, *natural cubic splines* can be useful to avoid overfitting near the boundaries of the sample space where there are fewer data points.

Question 3 [20 points] Multivariate non-parametric regression

Consider the following multivariate regression model

$$Y_i = m(X_{1,i}, \ldots, X_{d,i}) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $m(\cdot)$ is a d-variate unknown non-parametric function.

(a) Write the specification of the multivariate function $m(X_{1,i}, \ldots, X_{d,i})$ in the *additive model*. What is the identification condition of the additive model? Discuss one advantage and one disadvantage of the additive model.

Answer:
The additive model specifies $m(\cdot)$ as

$$m(X_{1,i}, \ldots, X_{d,i}) = c + \sum_{j=1}^{d} m_j(X_{j,i}),$$

where each $m_j(\cdot)$ is a univariate non-parametric unknown function. The identification condition of the additive model is $E\big(m_j(X_{1,i})\big) = 0$ for any $j = 1, \ldots, d$. This implies that c is the unconditional mean of Y_i, i.e. $E(Y_i) = c$. One advantage of the additive model is that it breaks the curse of dimensionality through the additive structure. In particular, estimates of the additive model have the same rate of convergence as univariate non-parametric models. One disadvantage of the additive model is that it imposes no interaction effects between the regressors. This can be a very restrictive assumption that can lead to misleading results as well as poor predictions.

(b) A colleague of yours claims the following: *"The additive model is always better than the parametric linear regression model because it is more flexible and it can capture non-linear relationships"*. Do you agree with this statement? Explain your reasoning.

Answer:
The additive model is indeed more flexible than the parametric linear regression model. Therefore, the statement is correct in this respect. However, this does not imply that it is better. The additive model is a non-parametric method and therefore estimation accuracy is lower compared to the linear regression model. As a result, in practice, the parametric linear regression model will be better than the additive model when the true relationship between the variables is linear (or close to linear).

Question 4 [20 points] Shrinkage methods

Consider the following multivariate linear regression model written in vector form

$$Y = X\beta + \epsilon,$$

where $Y = (Y_1, \ldots, Y_n)^\top$, X is a $n \times d$ matrix where each column contains one of the regressors and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$. Assume furthermore that Y and all the regressors are standardized, i.e. sample mean 0 and sample variance 1.

(a) Assume orthonormal regressors, i.e. $X^\top X = I_d$. Show that the Ridge estimator $\hat{\beta}_{Ridge}$, given by

$$\hat{\beta}_{Ridge} = (X^\top X + \lambda I_d)^{-1} X^\top Y,$$

is a biased estimator of the true parameter vector β.

Answer:
First, we can rewrite the Ridge estimator $\hat{\beta}_{Ridge}$ as a function of the OLS estimator $\hat{\beta}_{OLS}$, that is,

$$\begin{aligned}\hat{\beta}_{Ridge} &= (X^\top X + \lambda I_d)^{-1} X^\top Y \\ &= (X^\top X + \lambda I_d)^{-1}(X^\top X)(X^\top X)^{-1} X^\top Y \\ &= (X^\top X + \lambda I_d)^{-1}(X^\top X)\hat{\beta}_{OLS}.\end{aligned}$$

Therefore, accounting the assumption that $X^\top X = I_d$, we obtain

$$\begin{aligned}\hat{\beta}_{Ridge} &= (I_d + \lambda I_d)^{-1} I_d \hat{\beta}_{OLS} \\ &= \frac{1}{1+\lambda}\hat{\beta}_{OLS}.\end{aligned}$$

Since the OLS estimator is unbiased, $E(\hat{\beta}_{OLS}) = \beta$, we obtain that

$$E(\hat{\beta}_{Ridge}) = \frac{1}{1+\lambda}\hat{E}(\beta_{OLS}) = \frac{1}{1+\lambda}\beta.$$

Therefore the bias is

$$Bias(\hat{\beta}_{Ridge}) = \frac{1}{1+\lambda}\beta - \beta = -\frac{\lambda}{1+\lambda}\beta.$$

(b) Explain why the Ridge estimator, unlike the OLS estimator, can also be used in situations where the number of regressors is larger than the number of observations ($d > n$).

Answer:
The OLS estimator, $\hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top Y$, cannot be used when the number of regressors is smaller than the sample size. This is the case because $X^\top X$ is a singular matrix and therefore the inverse $(X^\top X)^{-1}$ is not defined. Instead, in the Ridge estimator, the inverse $(X^\top X)^{-1}$ is replaced by $(X^\top X + \lambda I_d)^{-1}$. The matrix $X^\top X + \lambda I_d$ is positive definite for any $\lambda > 0$ even if $X^\top X$ is singular. Therefore, the Ridge estimator can be implemented also when $d > n$.