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.

(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 $\|K\|_2^2 = \int_{-\infty}^{\infty} K^2(u)du$, $\mu_2(K) = \int_{-\infty}^{\infty} u^2 K(u)du$ 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.

Question 2 [40 points] Univariate non-parametric regression

Consider a univariate regression of Y_i on X_i of the form

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

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. $\lim_{h \to 0} \hat{m}_h(X_k) = Y_k$.

(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.

(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.

(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*.

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.

(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.

Question 4 [20 points] Shrinkage methods

Consider the following multivariate linear regression model written in vector form

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

where $Y = (Y_1, \dots, Y_n)^\top$, X is a $n \times d$ matrix where each column contains one of the regressors and $\epsilon = (\epsilon_1, \dots, \epsilon_n)^\top$. 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.

(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 ($d > n$).