STOCKHOLM UNIVERSITY
DEPT. OF MATHEMATICS
Div. of Mathematical statistics

MT 7038
EXAMINATION
13 Jan 2020

Exam in Statistical Learning
13 Jan 2020, time 9-14

Examinator: Chun-Biu Li, cbli@math.su.se.
Permitted aids: Course textbook (Elements of statistical learning) and your own lecture notes. **Electronic devices and e-books are not allowed.**
Return of the exam: To be announced in the discussion forum in the course page.

NOTE: The exam consists of 5 problems and each with 10 points. Logical explanation and steps leading to the final solution must be clearly shown in order to receive full marks. Minimum points to receive a given grade are as follows:

A	B	C	D	E
45	40	35	30	25

NOTE: The mathematical notations in this exam are the same as those in the course book.

NOTE: For those parts require explanation in words, your writing must be to the point, **redundant writing irrelevant to the solution can result in point deduction**.

Problem 1

a) One of the manifestations of the curse of dimensionality is the sparseness of data points. Show that the sampling density is roughly proportional to $N^{1/p}$, where p is the input space dimension and N is the sample size (see Section 2.5 of the course book). (2p)

b) Consider the case of an orthonormal $N \times p$ input matrix $\mathbf{X}$. Let $\hat{\beta}_j \quad (j = 1, \cdots, p)$ be the least square estimators of the parameters. Derive the estimators in Table 3.4 in the course book for the best subset with size M (3p), ridge regression (2p), and Lasso (3p).

Problem 2

a) Show that the degree-of-freedom of quadratic discriminant analysis equals to $(K-1) \times \left[\frac{p(p+3)}{2} + 1\right]$, where K is the number of classes and p is the dimension of the predictor variables. (4p)

b) In finding the separation hyperplane using Rosenblatt's perception learning algorithm, one minimizes the cost function $D(\beta, \beta_0) = -\sum_{i \in M} y_i(x_i^\top \beta + \beta_0)$, where $y_i = -1$ or 1, and M is the set of misclassified points. One problem of this cost function is that there is no unique separation hyperplane when the data is separable. Consider minimizing another cost function, $D_1(\beta, \beta_0) = -\sum_{i=1}^{N} y_i(x_i^\top \beta + \beta_0)$ subject to the constraint $\|\beta\| = 1$, where N is the number of observations. Describe this criterion clearly **in words** in terms of the signed distance and explain if this new cost function solves the uniqueness problem in the separable case. (3p)

c) Discuss one drawback of using the cost function $D_1(\beta, \beta_0)$ with constraint $\|\beta\| = 1$ in part b), then propose a possible solution to it and justify your answers. Hint: You may consider drawing a figure to help your explanation. (3p)

Problem 3

a) Consider the basis expansion of a function $f(X)$ using the cubic splines with K interior knots: $f(X) = \sum_{j=0}^{3} \beta_j X^j + \sum_{k=1}^{K} \alpha_k (X - \xi_k)_+^3$, where ξ_k are the positions of the knots. Show that $f(X)$ has continuous first and second derivatives at the knots. (2p)

b) Now taking into account the additional boundary conditions imposed by the natural cubic spline, show that this implies $\beta_2 = 0$, $\beta_3 = 0$, $\sum_{k=1}^{K} \alpha_k = 0$, $\sum_{k=1}^{K} \alpha_k \xi_k = 0$. (3p)

c) Finally, show that the results in b) lead to the basis functions of the natural cubic spline (i.e., Eq. 5.4 and 5.5 in the course book). (3p)

d) Let $\hat{\mathbf{f}}$ denote the N-vector fitted values $\hat{f}(x_i)$ at the training predictors x_i in the smoothing spline. One has the relation $\hat{\mathbf{f}} = \mathbf{S}_\lambda \mathbf{y}$, where $\mathbf{S}_\lambda$ is the smoother matrix with regularization parameter λ. The effective defree-of-freedom (dof) is given by the trace of $\mathbf{S}_\lambda$. Explain concisely **in words** why the rank of $\mathbf{S}_\lambda$ is not a good choice for the dof. You can cite the corresponding equations and properties in the course book to support your answer. (2p)

Problem 4

a) Consider the local linear regression at a target point x_0 as a weighted least square estimation: $\min_{\alpha(x_0),\beta(x_0)} \sum_{i=1}^{N} K_\lambda(x_0, x_i) \left[y_i - \alpha(x_0) - \beta(x_0)x_i\right]^2$, with kernel $K_\lambda(x_0, x_i)$. Show that the estimate is given by

$$\hat{f}(x_0) = b(x_0)^\top \left(\mathbf{B}^\top \mathbf{W}(x_0)\mathbf{B}\right)^{-1} \mathbf{B}^\top \mathbf{W}(x_0)\mathbf{y},$$

where $b(x)^\top = (1, x)$, $\mathbf{B}$ is the $N \times 2$ matrix with the i-th row given by $b(x_0)^\top$, and $\mathbf{W}(x_0)$ is the $N \times N$ diagonal matrix with the i-th diagonal element given by $K_\lambda(x_0, x_i)$. (5p)

b) Now let $\hat{f}(x_0) = \sum_{i=1}^{N} l_i(x_0) y_i$, show that $\sum_{i=1}^{N} l_i(x_0) = 1$ (2p) and $\sum_{i=1}^{N} (x_i - x_0) l_i(x_0)$ (3p).

Problem 5

For parts a) to c) below, suppose that the data is generated from the model $Y = f(X) + \epsilon$, with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$.

a) If $\hat{f}_k(x_0)$ is the k-nearest neighbor regression fit and assume that the values of x_i in the sample are fixed (i.e., non-random), show that the expected prediction error at x_0 is given by

$$E\left[\left(Y - \hat{f}_k(x_0)\right)^2 |X = x_0\right] = \sigma^2 + \left[f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f(x_{(l)})\right]^2 + \sigma^2/k,$$

where the subscript (l) indicates the l-th nearest neighbor to x_0. (4p)

b) Describe the meaning of each of the 3 terms in part a) and give an intuitive explanation why the last term (i.e. σ^2/k) is inversely proportional to k. (3p)

c) Now consider the linear model fit $\hat{f}_p(x) = x^\top\hat{\beta}$, where the parameter vector β with p components is fit by the least squares. Show that the variance in the expected prediction error at x_0, $E\left[\left(Y - \hat{f}_p(x_0)\right)^2 |X = x_0\right]$, is given by $Var[\hat{f}_p(x_0)] = \|\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}x_0\|^2\sigma^2$. (3p)

Good Luck!