Examination for Statistical Learning (MT7038 - VT19)

Date: 9:00am - 14:00pm, May 16, 2019

Examiner: 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**.

- The exam consists of 5 problems and each with 10 credit points. Logical explanation and steps leading to the final solution must be clearly shown in order to receive full marks.

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

Problem 1 (Linear Methods for Regression)

a) For the ridge regression problem, one has to solve

$$\hat{\beta}^{\text{ridge}} = \text{argmin}_{\beta}\left[\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right].$$

Find the relations between the parameters β_0, β_j's and $\beta_0^c,\ \beta_j^c$'s such that the above optimization problem can be stated equivalently as

$$\hat{\beta}^{\text{c}} = \text{argmin}_{\beta^c}\left[\sum_{i=1}^{N}(y_i - \beta_0^c - \sum_{j=1}^{p}(x_{ij} - \bar{x}_j)\beta_j^c)^2 + \lambda\sum_{j=1}^{p}(\beta_j^c)^2\right].$$ (3 pts)

b) Consider the case of an orthonormal $N \times p$ input matrix X and let $\hat{\beta}_j\ (j = 1,\cdots,p)$ be the least square estimators of the parameters. Derive the estimators in Table 3.4 in the course book:

TABLE 3.4. *Estimators of β_j in the case of orthonormal columns of $\mathbf{X}$. M and λ are constants chosen by the corresponding techniques;* sign *denotes the sign of its argument (± 1), and x_+ denotes "positive part" of x. Below the table, estimators are shown by broken red lines. The 45° line in gray shows the unrestricted estimate for reference.*

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

for the best subset with size M (2 pts), ridge regression (2 pts), and Lasso (3 pts).

Hint: Part a) and part b) are unrelated.

Problem 2 (Linear Methods for Classification)

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. (4 pts)

b) Consider Fig. 4.6 in the course book:

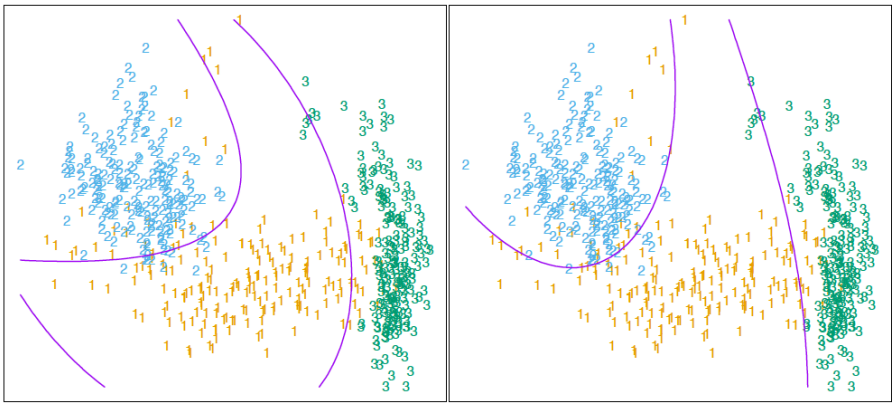

FIGURE 4.6. *Two methods for fitting quadratic boundaries. The left plot shows the quadratic decision boundaries for the data in Figure 4.1 (obtained using LDA in the five-dimensional space $X_1, X_2, X_1X_2, X_1^2, X_2^2$). The right plot shows the quadratic decision boundaries found by QDA. The differences are small, as is usually the case.*

Both decision boundaries in the left and right panel are quadratic but they look slightly different. Explain what cause(s) them to be different. Discuss which one (LDA in the 5-dimensional space or QDA) is more appropriate to use and justify your answer. (3 pts)

c) In finding the separation hyperplane using Rosenblatt's perceptron learning algorithm, the cost function, $D(\beta, \beta_0) = -\sum_{i \in M} y_i(x_i^T \beta + \beta_0)$, is minimized,

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 solution (i.e., separation hyperplane) when the data are separable. Consider minimizing another cost function,

$D^*(\beta, \beta_0) = -\sum_{i=1}^{N} y_i(x_i^T \beta + \beta_0)$ subject to the constraint $\|\beta\| = 1$, where all observations are summed. 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. (3 pts)

Problem 3 (Basis Expansion & Regularization)

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. (3 pts)

b) Now taking into account the additional boundary conditions imposed by the natural cubic splines, show that this implies

$$\beta_2 = 0, \quad \beta_3 = 0, \quad \sum_{k=1}^{K} \alpha_k = 0, \quad \sum_{k=1}^{K} \alpha_k \xi_k = 0.$$

(3 pts)

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). (2 pts)

d) Denote by $\hat{\boldsymbol{f}}$ the N-vector fitted values $\hat{f}(x_i)$ at the training predictors x_i in the smoothing spline, one has the relation $\hat{\boldsymbol{f}} = \boldsymbol{S}_\lambda \boldsymbol{y}$, where $\boldsymbol{S}_\lambda$ is the smoother matrix with regularization parameter λ. The effective degree-of-freedom (dof) is given by trace($\boldsymbol{S}_\lambda$). Explain **in words** why rank($\boldsymbol{S}_\lambda$) is not a good choice for the dof. You can cite the corresponding equations and properties in the course book to support your answer. Hint: Part d) is unrelated to parts a) to c). (2 pts)

Problem 4 (Kernel Smoothing Methods)

a) Consider the local linear regression at a target point x_0 as a weighted least square problem: $\min_{\alpha(x_0),\beta(x_0)} \sum_{i=1}^{N} K_\lambda(x_0, x_i)\left[y_i - \alpha(x_0) - \beta(x_0)x_i\right]^2$, with kernel $K_\lambda(x_0, x_i)$. Show that the estimate $\hat{f}(x_0)$ is given by $\hat{f}(x_0) = b(x_0)^T\left(B^T W(x_0) B\right)^{-1} B^T W(x_0) y$, where $b(x)^T = (1, x)$, B is the $N \times 2$ matrix with the i-th row equal to $b(x_i)^T$, and $W(x_0)$ is a $N \times N$ diagonal matrix with the i-th diagonal element equal to $K_\lambda(x_0, x_i)$. (4 pts)

b) Let $\hat{f}(x_0) = b(x_0)^T\left(B^T W(x_0) B\right)^{-1} B^T W(x_0) y = \sum_{i=1}^{N} l_i(x_0) y_i$, show that $\sum_{i=1}^{N} l_i(x_0) = 1$ and $\sum_{i=1}^{N} (x_i - x_0) l_i(x_0) = 0$.(4 pts)

c) In part a) and b), suppose that $K_\lambda(x_0, x_i)$ is a Gaussian kernel with λ the standard deviation. Discuss clearly **in words** the bias-variance tradeoff when λ varies from small to big values. (2 pts)

Problem 5 (Model Assessment & Selection)

Consider data generated from the model $Y = f(X) + \varepsilon$, with $\mathsf{E}(\varepsilon) = 0$, $\mathsf{Var}(\varepsilon) = \sigma^2$. Further assume that the values of x_i in the sample are fixed (i.e., nonrandom).

a) If $\hat{f}_k(x_0)$ is the *k*-nearest neighbor regression fit, show that the expected prediction error at x_0:

$$\mathsf{E}\left[(Y - \hat{f}_k(x_0))^2 \,|\, X = x_0\right] = \sigma^2 + \left[f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f(x_{(l)})\right]^2 + \frac{\sigma^2}{k},$$ where the subscript (l) indicates the *l*-th nearest neighbor to x_0. (4 pts)

b) Discuss the meaning of each of the three terms in the expected prediction error in part a) and discuss the bias-variance tradeoff as a function of *k*. (2 pts)

c) For the same setup as in part a) but now $\hat{f}_\lambda(x_0)$ is the ridge regression fit with $\lambda \geq 0$ the complexity parameter controlling the amount of shrinkage, show that the expected prediction error at x_0 equals to

$$\mathsf{E}\left[(Y - \hat{f}_\lambda(x_0))^2 \,|\, X = x_0\right] = \sigma^2 + \left[f(x_0) - \mathsf{E}\hat{f}_\lambda(x_0)\right]^2 + \left\|\boldsymbol{h}(x_0)\right\|^2\sigma^2,$$

where $\boldsymbol{h}(x_0) = X(X^T X + \lambda I)^{-1} x_0$ is the *N*-vector linear weights. (4 pts)

~ Good Luck ~