MIDTERM EXAMINATION

E6690: Statistical Learning for Bio & Info Systems
Prof. P. R. Jelenković
Nov 10, 2020

Exam duration: 2 1/2 hours; closed book; no calculator/computer; one sheet of paper (both sides) with formulas is allowed. All problems/subproblems carry equal points. Please read all problems carefully:

P1. Consider a set of observations $(y_1, x_1), \ldots, (y_n, x_n), x_i \geq 0, n \geq i \geq 1$.

(a) Fit these observations with a function $\hat{y}_i = \hat{\beta}_\lambda \sqrt{x_i}$ and Ridge penalty $\lambda \hat{\beta}_\lambda^2, \lambda > 0$, i.e., compute the optimal $\hat{\beta}_\lambda$, which minimizes the RSS with Ridge penalty

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_\lambda \sqrt{x_i})^2 + \lambda \hat{\beta}_\lambda^2.$$

Next, in (b, c, d), assume that the preceding observations satisfy $y_i = \beta\sqrt{x_i} + \epsilon_i$, where ϵ_i-s are i.i.d. random variables with normal/Gaussian distribution $\mathcal{N}(0, \sigma^2)$; ϵ_i-s are the only source of randomness.

(b) Under the preceding assumptions, for the optimal $\hat{\beta}_\lambda$ from (a), show that

$$\mathbb{E}\hat{\beta}_\lambda = \beta \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i + \lambda} \qquad \text{and} \qquad \mathsf{Var}(\hat{\beta}_\lambda) = \frac{\sigma^2 \sum_{i=1}^n x_i}{(\sum_{i=1}^n x_i + \lambda)^2}.$$

(Hint: $\mathsf{Var}(\sum c_i Z_i) = \sum c_i^2 \mathsf{Var}(Z_i)$, where c_i-s are constants and Z_i-s are independent random variables.)

(c) Now, we can test the the model on a new sample $x_0, y_0 = \beta\sqrt{x_0} + \epsilon_0$, where ϵ_0 is $\mathcal{N}(0, \sigma^2)$ and independent of $\epsilon_i, i \geq 1$. The mean square test error can be decomposed in terms of bias and variance as

$$\mathsf{MSE} = \mathbb{E}\,(y_0 - \hat{y}_0)^2 = \mathbb{E}(\beta\sqrt{x_0} + \epsilon_0 - \hat{\beta}_\lambda\sqrt{x_0})^2 = \sigma^2 + (\beta\sqrt{x_0} - \mathbb{E}(\hat{\beta}_\lambda)\sqrt{x_0})^2 + \mathsf{Var}(\hat{\beta}_\lambda\sqrt{x_0}),$$

where $(\beta\sqrt{x_0} - \mathbb{E}(\hat{\beta}_\lambda)\sqrt{x_0})^2$ is the squared bias. Discuss the bias-variance tradeoff in terms of the explicit expressions for squared bias and variance, which follow from formulas in part (b).

(d) Assume that $\lambda \leq 3\sigma^2/(2\beta^2)$ and compute the optimal λ^* which minimizes the

$$\mathsf{MSE} = \mathbb{E}(y_0 - \hat{y}_0)^2 = \mathbb{E}(\beta\sqrt{x_0} + \epsilon_0 - \hat{\beta}_\lambda\sqrt{x_0})^2.$$

P2. Recall TSS $= \sum(y_i - \bar{y})^2$ is the total sum of squares and ESS $= \sum(\hat{y}_i - \bar{y})^2$ is the explained sum of squares, where $\bar{y} = (\sum y_i)/n$.

(a) Simple linear regression model $\hat{y} = \beta_0 + \beta_1 x$ is fitted to $n = 162$ observations with ESS $= 70$ and TSS $= 390$. Compute the F-value for the null hypothesis $H_0 : \beta_1 = 0$

$$\frac{\mathsf{TSS} - \mathsf{RSS}}{\frac{\mathsf{RSS}}{n-2}}$$

and then, compute the corresponding p-value using this simple bound of the F-distribution $\mathbb{P}[\mathcal{F}_{1,d} > F] \approx e^{-F/2}/\sqrt{\pi F/2} < 2^{-0.7F}/\sqrt{\pi F/2}$, where $\mathcal{F}_{1,d}$ is F variable with $(1, d)$ degrees of freedom and d is large. Based on this estimate of p, should you accept or reject H_0?

(b) Consider data points $y_1, \ldots, y_n$, and assume no information about the distribution of y_i. Describe briefly how bootstrap can be used to estimate the variance of a sample mean $\bar{y} = \sum_{i=1}^{n} y_i/n$.

(c) You are optimizing the SVM classifier by trying different values of the slack budget C ($\sum_i \epsilon_i \leq C$). What are the most common *direct* ways for selecting the best value of C?

(d) Describe briefly K-fold cross validation, and its extreme case leave-one-out cross validation (LOOCV). What are pros and cons of LOOCV? Can these approaches be used for nonlinear models and without the Gaussian assumptions?

P3. Answer the following:

(a) Describe briefly ”best subset” and ”forward stepwise” model selection algorithms. What are their pros and cons?

(b) Let n be the number of samples and p the number of features. For a high dimensional problem with $p \gg n$, is it better to use Ridge or Lasso regression? Explain.

(c) Compare the LDA and SVC classification. What is their main similarity and difference? How do these models compare in terms of model simplicity, interpretability and accuracy? Explain your reasoning.

(d) Consider $n = 100$ samples, each having $p = 10,000$ features, that you are fitting with linear Ridge regression. Is it better to solve the ridge optimization problem as primal or dual? What are the orders of their computational complexities, respectively?

P4. The optimal Bayes classifier assigns an observation $\boldsymbol{x}$ to a class k for which the posterior probability $p_k(\boldsymbol{x}) = \mathbb{P}[Y = k|\boldsymbol{X} = \boldsymbol{x}]$ is the largest. Using Bayes' theorem, $p_k(\boldsymbol{x})$ is often conveniently represented in terms of priors, $\pi_k = \mathbb{P}[Y = k]$, and conditional densities $f_k(\boldsymbol{x})d\boldsymbol{x} = \mathbb{P}[\boldsymbol{X} \in (\boldsymbol{x} + d\boldsymbol{x})|Y = k]$.

(a) What is the problem of using the optimal Bayes classifier in practice and give two approaches of how this problem can be resolved.

(b) Consider a problem with two classes, $k = 0, 1$, and two features (x_1, x_2). Logistic regression $p(x_1, x_2) \equiv p_1(x_1, x_2)$ is fitted to training data with coefficients $\hat{\beta}_0 = -\ln(5), \hat{\beta}_1 = \ln(4/5), \hat{\beta}_2 = \ln(2)$. We assign an observation (x_1, x_2) to class $k = 1$ if $p(x_1, x_2) \geq 1/2$. For point $(-1, 1)$, compute $p(-1, 1)$ and decide to which class it belongs.

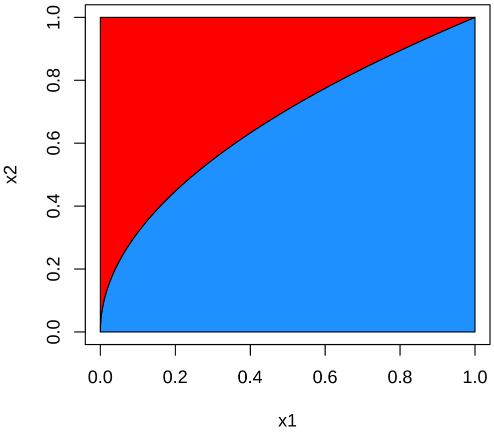
(c) Now, assume that there are two classes, $k = 0, 1$, and one feature, x ($p = 1$). For LDA with $\sigma = 1$, $\mu_1 = -\mu_0 = 2$ and $\pi_1 = e^3/(1 + e^3)$, compute the decision boundary x^* for which $p_1(x) = p_0(x)$.

(d) Again, for two classes, $k = 0, 1$, and one feature, x, assume that

$$f_k(x) = \frac{1}{\pi_k \gamma_k (1 + ((x - m_k)/\gamma_k)^2)}$$

with $m_0 = 0, m_1 = 2, \gamma_0 = 3, \gamma_1 = 1$. Compute the region of x where class 1 is selected, i.e., $p_1(x) \geq p_0(x)$.

P5. (a) Consider a data set illustrated on the figure below with two features (X_1, X_2) taking values in $[0,1]^2$, and belonging to two classes labeled with values "red" and "blue". The "boundary" between the classes is defined by $X_2 = \sqrt{X_1}$. Suppose that a classification tree is to be built. The number of observations is extremely high, and the density of observations in the unit square is uniform, i.e., the misclassification error is proportional to the area of the misclassified region. What should be the first split at the root of the tree, if the objective is to minimize the misclassification error? (Recall that $\int_a^b x^r = (b^{r+1} - a^{r+1})/(r+1)$.)



(b) Describe briefly and compare Bagging and Random Forest procedures. What is the main difference/improvement of Random Forest relative to Bagging?

(c) Consider 4 blue points with (x_1, x_2) coordinates $(1,7), (2,2), (3,6), (4,6)$, and 4 red points with coordinates $(2,-2), (4,1), (5,4), (8,3)$. Compute or draw clearly the maximal marginal line (hyperplane) that separates the red from blue points, identify the supporting vectors and compute the margin. (Recall, the height in the right triangle is given by $h = ab/c$, where a and b are the sides and c is the hypotenuse.)

(d) Describe the quadratic (polynomial of degree 2) basis expansion and what can this method accomplishes for SVM classifiers. Then, write and explain all the optimization equations for quadratic SVM, and in particular the meaning of slack variables, ϵ_i, when $\epsilon_i = 0, 0 < \epsilon_i < 1$ and $\epsilon_i > 1$.

GOOD LUCK!