MIDTERM EXAMINATION

E6690: Statistical Learning for Bio & Info Systems
Prof. P. R. Jelenković
November 9, 2021

Exam duration: 2 1/2 hours; closed book; no calculator/computer; one sheet of paper (both sides) with **handwritten** formulas is allowed. All problems/subproblems carry equal points, although the difficulty of various parts can be different. Many parts can be answered out of order. Please read all problems carefully:

P1. Consider a set of observations $(y_1, x_1), \dots, (y_n, x_n), n \ge 1$.

(a) Fit these observations with a simple linear function $\hat{y}(x_i) = \hat{\beta} x_i$ by minimizing the $\mathrm{RSS}(\hat{\beta}) = \sum_{i=1}^n (y_i - \hat{y}(x_i))^2 = \sum_{i=1}^n (y_i - \hat{\beta} x_i)^2$. Find an expression for $\hat{\beta}$ that minimizes $\mathrm{RSS}(\hat{\beta})$.

Next, in (b, c, d), assume that 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. In (b,d), the observations come from a population model $y_i = \beta x_i + \epsilon_i$.

(b) Under the preceding assumptions, for the optimal $\hat{\beta}$ from (a), compute explicit formulas for $\mathbb{E}\hat{\beta}$ and $\mathsf{Var}(\hat{\beta})$ in terms of x_i and σ^2. Is this a biased or unbiased estimator?
(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) Bias-variance principle. Suppose $\hat{y}(x) = \hat{f}(x)$ is an approximation for $y_i = f(x_i) + \epsilon_i, 1 \le i \le n$, which is tested at a new point $y_0 = f(x_0) + \epsilon_0$. We can show that the model variance, $\mathsf{Var}(\hat{f}(x_0)$, and its squared bias, $\mathsf{Bias}^2 = (f(x_0) - \mathbb{E}\hat{f}(x_0))^2$, satisfy

$$\mathbb{E}\left(y_0 - \hat{f}(x_0)\right)^2 = \sigma^2 + \left(f(x_0) - \mathbb{E}\hat{f}(x_0)\right)^2 + \mathsf{Var}(\hat{f}(x_0)).$$

Provide a detailed derivation of the preceding equality, and describe the bias-variance principle.

(d) Example of bias-variance principle. Instead of the simple model in (a), consider a more complex model $\hat{y}_1(x_i) = \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2$. **Assuming that** $\sum_{i=1}^n x_i^3 = 0$, compute the expressions for the optimal values of $(\hat{\beta}_1, \hat{\beta}_2)$, which minimize the $\mathrm{RSS}(\hat{\beta}_1, \hat{\beta}_2) = \sum_{i=1}^n (y_i - \hat{y}_1(x_i))^2 = \sum_{i=1}^n (y_i - (\hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2))^2$. Show that this is also an unbiased estimator, i.e., $\mathbb{E}\hat{y}_1(x) = \beta x$, but with higher variance than the one in part (a), i.e., $\mathsf{Var}(\hat{y}_1(x)) > \mathsf{Var}(\hat{y}(x))$.

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 $n = 132$ observations with ESS $= 50$ and TSS $= 310$. 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) Suppose that in the preceding part, (a), the noise is not Gaussian. Still, to test the null hypothesis, $H_0 : \beta_1 = 0$, we can compute the F-statistic, but we cannot compute the p-value since we don't know the distribution of F. Describe briefly how bootstrap can be used to estimate the p-value.

(c) In shrinkage models, Ridge or Lasso, we obtain a family of models indexed by λ. Outside of very simple models, we cannot compute the best λ (model) analytically. What are the most common *direct* ways for selecting the best λ? How can we eliminate randomness from the testing procedure?

(d) Given $n = 100$ data points with $p = 10$ features, $\boldsymbol{x}_i = (x_{i,1}, \ldots x_{i,10})$, consider 9th degree polynomial basis expansion $\boldsymbol{x}_i \to \boldsymbol{\phi}(\boldsymbol{x}_i)$, such that the new basis vector $\boldsymbol{\phi}(\boldsymbol{x}_i)$ contains all the terms of the form $x_{i,1}^{\alpha_1} x_{i,2}^{\alpha_2} \cdots x_{i_{10}}^{\alpha_{10}}$ with $0 \le \alpha_k \le 9$, $k = 1, 2, \ldots, 10$. Compute the dimension of the new basis vector $\boldsymbol{\phi}(\boldsymbol{x}_i)$. If we were to fit a Ridge regression model to the new basis vectors $\boldsymbol{\phi}(\boldsymbol{x}_i)$, is it better to solve the problem in primal or dual (dot product) formulation? Compare the orders of computational complexities of these solutions.

P3. In general, it is desirable to find the simplest, interpretable models with good accuracy.

(a) Describe briefly the "best subset" selection algorithm. What is its main drawback and how can it be resolved?

(b) Write the main optimization equations for Ridge and Lasso regression, and compare them in terms of: analytical tractability, model simplicity, interpretability and accuracy. How can "bias-variance tradeoff" be used to justify Ridge and Lasso regression?

(c) Compare the tree-based regression methods versus Ridge/Lasso in terms of interpretability and accuracy.

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

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' formula, $p_k(\boldsymbol{x})$ 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 that we covered in class of how this problem can be resolved.

(b) Make a detailed comparison between Logistic and LDA classification. What is their main similarity and difference? How do these models compare in terms of model simplicity, interpretability and accuracy? Explain your reasoning.

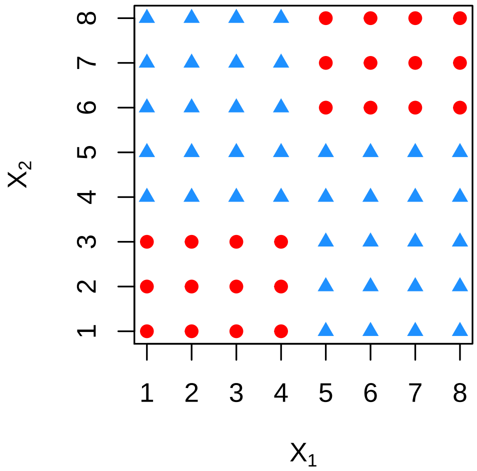
(c) For two classes, $k = 0, 1$, and one feature, x, assume that the conditional densities are given by

$$f_k(x) = \frac{\lambda_k}{2} e^{-\lambda_k |x - \mu_k|}.$$

If $\lambda_0 = \lambda_1 = \lambda > 0, \pi_1 = \pi_0 e^{-\lambda}, \mu_1 - \mu_0 > 1$, compute the region of x where class 1 is selected, i.e., $p_1(x) \ge p_0(x)$.

(d) Repeat the preceding question with $\pi_0 = \pi_1, \lambda_0 = 1, \lambda_1 = e, (\mu_1 - \mu_0) e \ge 1$.

P5. (a) Consider a tree-based classification method with 2 classes, red and blue, depicted in the figure below. To select the first node (root) of the tree, we consider splitting the features $x_k, k = 1, 2$ in two regions along points $x_k = i + 1/2, i = 0, 1, 2, \ldots, 7, k = 1, 2$. After a split in 2 regions (say $x_1 < 3.5$), we assign each region to a class according to the majority vote: if (# of blue points)$\geq$(# of red points), then the region is classified as blue; otherwise, it is red. What is the number of errors in each of these splits? Next, if the root node is selected to be $x_1 < 4.5$, draw and label a perfect classification tree that makes no errors.



(b) Use the preceding part, (a), to motivate and then explain the tree pruning method.

(c) Consider 4 blue points with (x_1, x_2) coordinates $(1,3), (3,6), (4,4), (5,9)$, and 4 red points with coordinates $(4,1), (6,1), (7,4), (9,2)$. Compute or draw clearly the maximal marginal hyperplane (line) that separates the red from blue points, identify the supporting vectors and compute the margin.

(d) When points/classes are not separable, the Support Vector Classifier (SVC) resolves the problem. Write and explain all the optimization equations for SVC. What is the meaning of slack variables, ϵ_i, and in particular, explain the meaning of $\epsilon_i = 0, 0 < \epsilon_i < 1$ and $\epsilon_i > 1$.
(Hint: Use the fact that $(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})/\|\boldsymbol{\beta}\|$ represents the signed distance of point $\boldsymbol{x}_i$ to the hyperplane $\beta_0 + \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle = 0$; $\|\boldsymbol{\beta}\|^2 = \beta_1^2 + \cdots + \beta_p^2$)

GOOD LUCK!