MIDTERM EXAMINATION

E6690: Statistical Learning for Bio & Info Systems

Prof. P. R. Jelenković

October 29, 2019

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 $(y_1, x_1), \ldots, (y_n, x_n), n \geq 1$.

(a) Fit these observations with a simple linear function $\hat{y}_i = \alpha + \hat{\beta}_\alpha x_i$ with a fixed intercept α (α is not optimized). Compute the optimal $\hat{\beta}_\alpha$, which minimizes the $\mathrm{RSS}(\hat{\beta}_\alpha) = \sum_{i=1}^n (y_i - \alpha - \hat{\beta}_\alpha x_i)^2$.

Next, in (b, c, d), assume that the preceding observations satisfy $y_i = \beta_0 + \beta x_i + \epsilon_i$, 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 $\hat{\beta}_\alpha$ from (a), compute $\mathbb{E}\hat{\beta}_\alpha$ and $\mathrm{Var}(\hat{\beta}_\alpha)$.
(Hint: $\mathrm{Var}(\sum c_i Z_i) = \sum c_i^2 \mathrm{Var}(Z_i)$, where c_i-s are constants and Z_i-s are independent random variables.)

(c) Now, instead of the simple model in (a) with fixed intercept α, consider a more flexible model $\hat{y}_i = \hat{\beta}_0 + \hat{\beta} x_i$, where both $\hat{\beta}_0$ and $\hat{\beta}$ are fit to data. For the optimal choice of $\hat{\beta}$, which minimize the RSS, we showed in class that

$$\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

where $\bar{x} = \sum x_i / n$. Compare the preceding expression with the expression from (b) and show that $\mathrm{Var}(\hat{\beta}_\alpha) < \mathrm{Var}(\hat{\beta})$ if $\bar{x} \neq 0$. For which value of α, in terms of the population model parameters (β_0, β), is $\mathbb{E}\hat{\beta}_\alpha = \beta$, i.e., $\hat{\beta}_\alpha$ is an unbiased estimator of β?

(d) Part (c) suggests that fitting first $\hat{\beta}_\alpha$ for fixed α, and then optimizing $(\alpha, \hat{\beta}_\alpha)$ leads to a better model in terms of bias and variance than fitting $\hat{\beta}_0$ and $\hat{\beta}$ at once. Explain this seeming contradiction. In practice, we don't know the population model parameters (β_0, β). How would you implement the preceding procedure of first fitting $\hat{\beta}_\alpha$ for fixed α, and then optimizing the choice of $(\alpha, \hat{\beta}_\alpha)$ using only data?

P2. Recall $\mathrm{TSS} = \sum (y_i - \bar{y})^2$ is the total sum of squares and $\mathrm{ESS} = \sum (\hat{y}_i - \bar{y})^2$ 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 $n = 142$ observations with $\mathrm{ESS} = 60$ and $\mathrm{TSS} = 340$. Compute the F-value for the null hypothesis $H_0 : \beta_1 = 0$

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

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?

(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. 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.

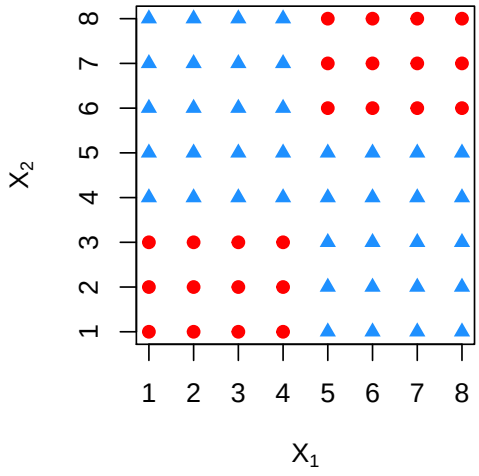
(c) For two classes, $k = 0, 1$, and one feature, x, assume that the conditional densities are given by

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

If $\lambda_0 = \lambda_1 = \lambda$, compute the region of x where class 1 is selected, i.e., $p_1(x) \geq p_0(x)$. How does this case compare to LDA?

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

P5. (a) Consider a tree-based method for classification in 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 $(1, 3), (2, 3), (3, 5), (5, 8)$, and 4 red points with coordinates $(4, 1), (6, 1), (7, 3), (8, 2)$. 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!