MIDTERM EXAMINATION

E6690: Statistical Learning for Bio & Info Systems

Prof. P. R. Jelenković

October 30, 2018

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 $\hat{y}_i = \hat{\beta}_\lambda x_i$ with no intercept and Ridge penalty $\lambda\hat{\beta}_\lambda^2, \lambda \geq 0$, i.e., compute the optimal $\hat{\beta}_\lambda$, which minimizes the RSS with Ridge penalty

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

Answer: Since this a convex function we compute $\hat{\beta}_\lambda$ from

$$\frac{d}{d\hat{\beta}_\lambda}\left(\sum_{i=1}^{n}(y_i - \hat{\beta}_\lambda x_i)^2 + \lambda\hat{\beta}_\lambda^2\right) = -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_\lambda x_i) + 2\lambda\hat{\beta}_\lambda = 0,$$

which yields

$$\hat{\beta}_\lambda = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 + \lambda} \tag{1}$$

Next, in (b, c, d), assume that the preceding observations satisfy $y_i = \beta x_i + \epsilon_i$, 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 $\hat{\beta}_\lambda$ from (a), show that

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

(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.)

Answer: $\mathbb{E}\hat{\beta}_\lambda$ follows from (1) and $\mathbb{E}y_i = \beta x_i$. Variance also follows from (1) and the hint

$$\mathrm{Var}(\hat{\beta}_\lambda) = \mathrm{Var}\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 + \lambda}\right) = \sum_{i=1}^{n}\left(\frac{x_i}{\sum_{i=1}^{n} x_i^2 + \lambda}\right)^2 \mathrm{Var}(y_i)$$

(c) Now, we can test the the model on a new sample $x_0, y_0 = \beta x_0 + \epsilon_0$, 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

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

where $(\beta x_0 - \mathbb{E}(\hat{\beta}_\lambda)x_0)^2$ 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).

Answer: Bias^2 increases from 0 to $x_0^2\beta^2$, as $\lambda \uparrow \infty$.
Variance, on the other hand, decreases from $\sigma^2/(\sum_{i=1}^{n} x_i^2)$ to 0 as $\lambda \uparrow \infty$.
Note that λ represents the flexibility of the model: as λ increases, the flexibility decreases. Hence, the less the flexibility in the model, the smaller the variance, but the bigger the bias. In general, we trade a bit of a bias increase, for a hopefully even bigger decline in variance.

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

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

Answer:

$$\begin{aligned}
\mathbb{E}(\beta x_0 + \epsilon_0 - \hat{\beta}_\lambda x_0)^2 &= x_0^2 \mathbb{E}(\beta - \hat{\beta}_\lambda)^2 + \sigma^2 \\
&= x_0^2 \mathbb{E}\left(\beta - \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2 + \lambda}\right)^2 + \sigma^2 \\
&= x_0^2 \mathbb{E}\left(\beta - \frac{\sum_{i=1}^n x_i(\beta x_i + \epsilon_i)}{\sum_{i=1}^n x_i^2 + \lambda}\right)^2 + \sigma^2 \\
&= x_0^2 \mathbb{E}\left(\frac{\beta\lambda + \sum_{i=1}^n x_i \epsilon_i}{\sum_{i=1}^n x_i^2 + \lambda}\right)^2 + \sigma^2 \\
&= x_0^2 \frac{\beta^2\lambda^2 + \sigma^2 \sum_{i=1}^n x_i^2}{\left(\sum_{i=1}^n x_i^2 + \lambda\right)^2} + \sigma^2 \\
&= x_0^2 \frac{\beta^2\lambda^2 + \sigma^2 s}{(s+\lambda)^2} + \sigma^2.
\end{aligned}$$

where $s = \sum x_i^2$. Next, to find an optimal λ, we compute the derivative

$$\frac{d}{d\lambda}\frac{\beta^2\lambda^2 + \sigma^2 s}{(s+\lambda)^2} = \frac{2\beta^2\lambda}{(s+\lambda)^2} - 2\frac{\beta^2\lambda^2 + \sigma^2 s}{(s+\lambda)^3} = 2\frac{\beta^2\lambda(s+\lambda) - (\beta^2\lambda^2 + \sigma^2 s)}{(s+\lambda)^3} = 2s\frac{\beta^2\lambda - \sigma^2}{(s+\lambda)^3} = 0$$

Hence,

$$\lambda^* = \frac{\sigma^2}{\beta^2}$$

This λ^* is the minimum since for $\lambda \le 3\sigma^2/(2\beta^2)$, the second derivative, i.e.,

$$\frac{d}{d\lambda}\frac{\beta^2\lambda - \sigma^2}{(s+\lambda)^3} = \frac{\beta^2(s+\lambda) - 3(\beta^2\lambda - \sigma^2)}{(s+\lambda)^4} = \frac{\beta^2 s - 2\beta^2\lambda + 3\sigma^2}{(s+\lambda)^4} \ge \frac{\beta^2 s}{(s+\lambda)^4} > 0.$$

P2. Recall $\mathsf{TSS} = \sum(y_i - \bar{y})^2$ is the total sum of squares and $\mathsf{ESS} = \sum(\hat{y}_i - \bar{y})^2$ 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 $n = 152$ observations with $\mathsf{ESS} = 50$ and $\mathsf{TSS} = 350$. 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?

Answer: $\mathsf{RSS} = \mathsf{TSS} - \mathsf{ESS} = 350 - 50 = 300$. Hence

$$F = \frac{\mathsf{TSS} - \mathsf{RSS}}{\frac{\mathsf{RSS}}{n-2}} = \frac{50}{\frac{300}{152-2}} = 25$$

and therefore, $\mathbb{P}[\mathcal{F}_{1,d} > F] < 2^{-0.7\cdot 25}/\sqrt{\pi 25/2} < 2^{-17}/4 = 2^{-19} \approx 2 \times 10^{-6}$ is small and we reject the H_0 hypothesis.

(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.

Answer: Draw bootstraps $(x_i^{*b}, y_i^{*b}), 1 \leq b \leq B$. For each bootstrap fit the data and compute the corresponding F^{*b}. Then, estimate the p-value as

$$p - \text{value} = \mathbb{P}[\mathcal{F}_{1,d} > F] \approx \frac{\sum_{b=1}^{B} 1_{\{F^{*b} > F\}}}{B}.$$

(c) In shrinkage models, Ridge or Lasso, we obtain a family of models indexed by λ. Outside of very simple models, e.g., P1. (d), we cannot compute the best λ/model analytically. What are the most common *direct* ways for select the best λ/model?

Answer: Cross-validation, either K-fold or LOOCV.

(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?

Answer: Description in the lecture notes or section 5.1 in ISL book. Pros: no randomness; cons: too much computation. Yes.

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?

Answer: Description in the lecture notes or p. 205 in ISL book.
Drawback: too much computational complexity - need to check 2^p models.
Resolution: Greedy approach: either forward or backward subset selection.

(b) Write the main optimization equations for Ridge and Lasso regression, and compare them in terms of: analytical tractability, model simplicity, interpretability and accuracy. Explain.

Answer: Equations: lecture notes or the ISL book.
Analytical tractability: Ridge better since it has explicit formulas, see equation (3.47), p. 66 in ESL.
Simplicity/interpretability: Lasso better since it has more $\hat{\beta}_i = 0$.
Accuracy: could go either way. If the actual problem depends only on a substet of features, then Lasso might be better. But, if it depends on all features, then Ridge might be better.

(c) Compare the tree-based methods versus other regression (or classification) techniques, e.g. Ridge/Lasso, in terms of interpretability and accuracy.

Answer: Interpretability: tree-based win (by far);
Accuracy: other methods better, e.g. Ridge/Lasso, since the fit is optimized.
(see sec. 8.1.4, p. 315 in ISL)

(d) Compare the 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.

Answer: Similarity: both have linear decision boundaries.
Difference: the way the parameters are fit (see p. 151 in ISL)

- for regression, the parameters $\hat{\beta}_i$ are optimized.

- for LDA, we estimate the means and variance from data.

Simplicity: about the same.
Interpretability: LDA is more interpretable since mean and variance have a meaning.
Accuracy: could go either way: if data is actually Gaussian, then LDA could work better

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.

Answer: In practice, we don't know $\mathbb{P}[Y = k|\boldsymbol{X} = \boldsymbol{x}]$ or $f_k(\boldsymbol{x})$, and these are difficult to estimate when there are a lot of features. Hence, we assume a specific shape for $f_k(\boldsymbol{x})$: Logistic or Gaussian (QDA/LDA).

(b) Consider a problem with two classes, $k = 0, 1$, and two features ($p = 2$), (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 $\hat{\beta}_0 = \ln(3), \hat{\beta}_1 = \ln(4/3), \hat{\beta}_2 = \ln(2)$. 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.

Answer: For given $\hat{\beta}_i$ and point $(1, -1)$

$$\hat{\beta}_0 + x_1\hat{\beta}_1 + x_2\hat{\beta}_2 = \ln(3) + \ln(4/3) - \ln(2) = \ln(2).$$

Therefore, $p(1, -1) = e^{\ln(2)}/(1 + e^{\ln(2)}) = 2/3 > 1/2$, implying $(1, -1)$ belongs to class 1.

(c) Consider QDA with two classes, $k = 0, 1$, and one feature, x ($p = 1$). Assume that $f_0(x)$ has standard normal density, $\mu_0 = 0, \sigma_0 = 1$, $f_1(x)$ is normal with $\mu_1 = 4, \sigma_1 = 3$ and prior $\pi_0 = 1/4$. Compute the region where x is assigned to class 1, i.e., $p_1(x) \geq p_0(x)$.

Answer: $p_1(x) \geq p_0(x)$ is equivalent to $\pi_1 f_1(x) \geq \pi_0 f_0(x)$, or

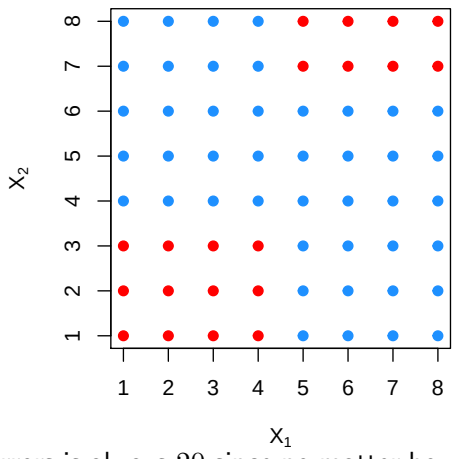
$$\frac{3}{4}\frac{1}{3\sqrt{2\pi}}e^{-\frac{(x-4)^2}{2\cdot 3^2}} \geq \frac{1}{4}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \Leftrightarrow \frac{x^2}{2} - \frac{(x-4)^2}{2\cdot 3^2} \geq 0,$$

which is true if $x \geq 1$ or $x \leq -2$.

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

Answer: Marginal hyperplane (line): $x_2 = 1 + x_1$
Support vectors: $(2, 1), (5, 4), (1, 4)$ and margin $M = \sqrt{2}$.

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 $x_2 < 6.5$, 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.



Answer: The total number of errors is always 20 since no matter how we separate the points by a vertical or horizontal line, they are all classified as blue.
Perfect tree with zero errors:

If $x_1 < 4.5, x_2 < 3.5$, then Red
If $x_1 < 4.5, x_2 \geq 3.5$, then Blue
If $x_1 > 4.5, x_2 < 6.5$, then Blue
If $x_1 > 4.5, x_2 \geq 6.5$, then Red

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

Answer: Motivation: in (a) there is no good way to find a root node of the tree since the number of errors is always 20.
Hence, the idea of tree pruning is to first build a big tree and then prune it to find the best subtree.
See pages 307-309 in ESL.

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

Answer: Description: see sec 8.2 in ISL
Similarity: they both use bootstrap to build a lot of trees, and then average them out.
Difference: random forests optimize in each step over a random subset, $m, m \approx \sqrt{p}$, of features, which help decorrelate the individual bootstrap trees, and in this way improve the accuracy.

(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 $\beta_0 + \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle = 0$; $\|\boldsymbol{\beta}\|^2 = \beta_1^2 + \cdots + \beta_p^2$)

Answer: See sec 9.2 and equations (9.12)-(9.15) on page 346 in ISL book.

GOOD LUCK!