STOR 767 Spring 2019 Hw2

Due on 02/06/2019 in Class

Remark. This homework focuses on basics of statistical learning and linear regression methods.

Instruction.

- Submission of handwritten solution for the **Theoretical Part** of this homework is allowed.
- Please use **RMarkdown** to create a formatted report for the **Computational Part** of this homework.
- Some of the problems are selected from the textbook [3].
- At most 100 pt (out of 115) can be earned for Homework 2.

Theoretical Part

1. (*5 pt*) (Textbook Ex. 2.2) Read Secion 2.3 and 2.4. Show how to compute the Bayes decision boundary for the simulation example in Figure 2.5.

2. (*10 pt*) (Textbook Ex. 2.7) Suppose we have a sample of N pairs $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^N$ drawn *i.i.d.* from the distribution characterized as follows:

$$\begin{cases} \boldsymbol{X}_i \sim h(\boldsymbol{x}), & \text{the design density on } \mathbb{R}^p \\ Y_i = f(\boldsymbol{X}_i) + \epsilon_i, & f \text{ is the regression function} \\ \epsilon_i \sim (0, \sigma^2), & \text{mean zero, variance } \sigma^2 \\ \{\epsilon_i\}_{i=1}^N \perp\!\!\!\perp \{\boldsymbol{X}_i\}_{i=1}^N. \end{cases}$$

   We construct an estimator for f **linear** in the Y_i

$$\hat{f}(\boldsymbol{x}_0) := \sum_{i=1}^N \ell_i(\boldsymbol{x}_0; \mathcal{X}) Y_i,$$

   where the weights $\ell_i(\boldsymbol{x}_0; \mathcal{X})$ do not depend on the Y_i, but do depend on the entire training sequence of $\mathcal{X} := \{\boldsymbol{X}_j\}_{j=1}^N$. Fix $\boldsymbol{x}_0 \in \mathbb{R}^p$.

   (a) Show that linear regression and k-nearest-neighbor regression are members of this class of estimators. Describe explicitly the weights $\ell_i(\boldsymbol{x}_0; \mathcal{X})$ in each of these cases.

   (b) Decompose the conditional mean-square error (CMSE)

$$\mathbb{E}_{\mathcal{Y}|\mathcal{X}} \left(f(\boldsymbol{x}_0) - \hat{f}(\boldsymbol{x}_0) \right)^2$$

   into a conditional squared bias and a conditional variance component, where $\mathcal{X} = \{\boldsymbol{X}_i\}_{i=1}^N$ and $\mathcal{Y} = \{Y_i\}_{i=1}^N$ represent the entire training sequence of covariates and responses respectively.

(c) Decompose the (unconditional) mean-square error (MSE)

$$\mathbb{E}_{\mathcal{X},\mathcal{Y}}\left(f(\boldsymbol{x}_0)-\hat{f}(\boldsymbol{x}_0)\right)^2$$

into a squared bias and a variance component.

(d) Establish the relationships between the conditional and unconditional versions of squared biases, variances respectively. In particular, how do the unconditional versions compared to the expected conditional versions?

3. (*10 pt*) (Textbook Ex. 2.9) Consider a linear regression model with p covariates, fit by least squares to a set of training data $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^N$ drawn at random from a population. Let $\hat{\boldsymbol{\beta}}$ be the least squares estimate. Suppose we have some test data $\{(\tilde{\boldsymbol{X}}_i, \tilde{Y}_i)\}_{i=1}^M$ drawn at random from a population as the training data. If

$$\mathcal{R}_{\mathrm{tr}}(\boldsymbol{\beta}) := \frac{1}{N}\sum_{i=1}^{N}\left(Y_i-\boldsymbol{\beta}^T\boldsymbol{X}_i\right)^2$$

and

$$\mathcal{R}_{\mathrm{te}}(\boldsymbol{\beta}) := \frac{1}{M}\sum_{i=1}^{M}\left(\tilde{Y}_i-\boldsymbol{\beta}^T\tilde{\boldsymbol{X}}_i\right)^2,$$

prove that

$$\mathbb{E}[\mathcal{R}_{\mathrm{tr}}(\hat{\boldsymbol{\beta}})] \leqslant \mathbb{E}[\mathcal{R}_{\mathrm{te}}(\hat{\boldsymbol{\beta}})].$$

4. (*10 pt*) (Textbook Ex. 3.7) Assume independently

$$Y_i \sim \mathcal{N}(\beta_0+\boldsymbol{X}_i^T\boldsymbol{\beta},\sigma^2), \quad (i=1,2,\cdots,N)$$

where $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \cdots, X_{ip})^T$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$, and the parameters $\{\beta_j\}_{j=1}^p$ are each distributed as $\mathcal{N}(0,\tau^2)$, independently of one another. Assuming β_0, σ^2 and τ^2 are known, show that the (minus) log-posterior density of $\boldsymbol{\beta}$ is proportional to

$$\sum_{i=1}^{N}\left(Y_i-\beta_0-\sum_{j=1}^{p}X_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

where $\lambda = \sigma^2/\tau^2$.

5. (*15 pt*) (Textbook Ex. 3.15) Let $\mathbf{X} = [X_{ij}]_{n\times p}$ be a standardized covariate matrix (*i.e.* $\sum_{i=1}^n X_{ij} = 0$, $\sum_{i=1}^n X_{ij}^2 = 1$) with columns $\{\boldsymbol{X}_{\cdot j}\}_{j=1}^p$, $\boldsymbol{Y} \in \mathbb{R}^n$ be the response vector. Fix $1 \leqslant m \leqslant p$. Recall that the m-th PLS variable $\boldsymbol{Z}_m = \mathbf{X}\hat{\boldsymbol{\phi}}_m$ is obtained by Algorithm 3.3 in Textbook. Taking the following steps to verify that $\hat{\boldsymbol{\phi}}_m$ solves

$$\begin{aligned}\max_{\boldsymbol{\phi}} \quad & \widehat{\mathbf{Cor}}(\mathbf{X}\boldsymbol{\phi},\boldsymbol{Y})^2\widehat{\mathbf{Var}}(\mathbf{X}\boldsymbol{\phi})\\ \text{s.t.} \quad & \boldsymbol{\phi}^T\widehat{\mathbf{Cov}}(\mathbf{X})\hat{\boldsymbol{\phi}}_l = 0 \qquad (1\leqslant l\leqslant m-1)\\ & \|\boldsymbol{\phi}\|_2 = 1\end{aligned} \tag{1}$$

where $\widehat{\mathbf{Cov}}(\mathbf{X}\boldsymbol{\phi},\boldsymbol{Y})$, $\widehat{\mathbf{Var}}(\mathbf{X}\boldsymbol{\phi})$ and $\widehat{\mathbf{Cov}}(\mathbf{X})$ are the sample correlation, sample variance and sample covariance matrix of $(\mathbf{X}\boldsymbol{\phi},\boldsymbol{Y})$, $\mathbf{X}\boldsymbol{\phi}$ and $\mathbf{X}$ respectively.

(a) Show that (1) can be rewritten in inner-product form

$$\begin{aligned} \max_{\boldsymbol{\phi}} \quad & \langle \boldsymbol{Z}_m, \boldsymbol{Y} \rangle \\ \text{s.t.} \quad & \boldsymbol{Z}_m = \mathbf{X}\boldsymbol{\phi} \\ & \langle \boldsymbol{Z}_m, \boldsymbol{Z}_l \rangle = 0 \quad (1 \leqslant l \leqslant m-1) \\ & \|\boldsymbol{\phi}\|_2 = 1. \end{aligned} \tag{2}$$

(b) Inductively argue that orthogonality constraints in (2) can be successively replaced by Gram-Schmidt orthogonalization:

$$\begin{aligned} \text{s.t.} \quad & \widetilde{\boldsymbol{Z}}_m = \mathbf{X}\boldsymbol{\varphi} \\ & \boldsymbol{Z}_m = \widetilde{\boldsymbol{Z}}_m - \textstyle\sum_{l=1}^{m-1} \frac{\langle \widetilde{\boldsymbol{Z}}_m, \boldsymbol{Z}_l \rangle}{\langle \boldsymbol{Z}_l, \boldsymbol{Z}_l \rangle} \boldsymbol{Z}_l. \end{aligned} \tag{3}$$

(c) Define $\mathbf{X}^{(m-1)}$ so that (3) reduces to

$$\boldsymbol{Z}_m = \mathbf{X}^{(m-1)}\boldsymbol{\varphi}.$$

Find $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{p\times p}$ such that $\mathbf{X}^{(m-1)} = \mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{B}$. What can you tell from $\mathbf{X}^{(m-1)}$?

(d) Give the reparametrization relationship of $\boldsymbol{\phi}$ in (2) and $\boldsymbol{\varphi}$ in (3) by $\mathbf{X}\boldsymbol{\phi} = \mathbf{X}^{(m-1)}\boldsymbol{\varphi}$. Prove that

$$\boldsymbol{Z}_m = \mathbf{X}^{(m-1)}\boldsymbol{\phi}.$$

Hint. Studying $\mathbf{X}^{(m-1)}\widehat{\boldsymbol{\phi}}_l$ for $1 \leqslant l \leqslant m-1$ might be useful.

(e) Show that (2) reduces to a projection of $\boldsymbol{Y}$ on the column space of $\mathbf{X}^{(m-1)}$.

(f) Give an updating formula from $\boldsymbol{X}^{(m-1)}$ to $\boldsymbol{Z}_m$. Compare it with Algorithm 3.3 in Textbook.

(g) Give an updating formula from $\mathbf{X}^{(m-1)}$ to $\mathbf{X}^{(m)}$ which might depend on $\boldsymbol{Z}_m$. Compare it with Algorithm 3.3 in Textbook.

6. (*15 pt*) (Textbook Ex. 3.27) Consider the LASSO problem in Lagrange multiplier form: with

$$L(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{N}\left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2,$$

we minimize

$$L(\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}|\beta_j| \tag{4}$$

for fixed $\lambda \geqslant 0$.

(a) Setting $\beta_j = \beta_j^+ - \beta_j^-$ with $\beta_j^+, \beta_j^- \geqslant 0$, expression (4) becomes

$$L(\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}(\beta_j^+ + \beta_j^-) \tag{5}$$

with non-negativity constraints on $\{\beta_j^{\pm}\}_{j=1}^p$. Show that the Lagrange dual function to (5) is

$$L(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} (\beta_j^+ + \beta_j^-) - \sum_{j=1}^{p} \lambda_j^+ \beta_j^+ - \sum_{j=1}^{p} \lambda_j^- \beta_j^-$$

and the the Karush–Kuhn–Tucker (KKT) optimality conditions are

$$\begin{aligned} \nabla_j L(\boldsymbol{\beta}) + \lambda - \lambda_j^+ &= 0 \\ -\nabla_j L(\boldsymbol{\beta}) + \lambda - \lambda_j^- &= 0 \\ \lambda_j^+ \beta_j^+ &= 0 \\ \lambda_j^- \beta_j^- &= 0 \end{aligned}$$

along with the non-negativity constraints on the parameters $\{\beta_j^{\pm}\}_{j=1}^p$ and all the Lagrange multipliers $\{\lambda_j^{\pm}\}_{j=1}^p$.

Hint. More information on the KKT conditions can be found in [1].

(b) Show that $|\nabla_j L(\boldsymbol{\beta})| \leqslant \lambda$ $(\forall 1 \leqslant j \leqslant p)$, and that the KKT conditions imply one of the following three scenarios:

$$\begin{aligned} \lambda = 0 \quad &\Rightarrow \quad \nabla_j L(\boldsymbol{\beta}) = 0 \quad (\forall 1 \leqslant j \leqslant p) \\ \beta_j^+ > 0,\ \lambda > 0 \quad &\Rightarrow \quad \lambda_j^+ = 0,\ \nabla_j L(\boldsymbol{\beta}) = -\lambda < 0,\ \beta_j^- = 0 \\ \beta_j^- > 0,\ \lambda > 0 \quad &\Rightarrow \quad \lambda_j^- = 0,\ \nabla_j L(\boldsymbol{\beta}) = +\lambda > 0,\ \beta_j^- = 0. \end{aligned}$$

Hence show that for any "active" predictor having $\beta_j \neq 0$, we must have $\nabla_j L(\boldsymbol{\beta}) = -\lambda$ if $\beta_j > 0$, and $\nabla_j L(\boldsymbol{\beta}) = \lambda$ if $\beta_j < 0$. Assuming the predictors are standardized (*i.e.* $\sum_{j=1}^p X_{ij} = 0$, $\sum_{j=1}^p X_{ij}^2 = 1$), relate λ to the correlation between the j-th predictor and the current residuals.

(c) Suppose that the set of active predictors is unchanged for $\lambda_0 \geqslant \lambda \geqslant \lambda_1$. Show that there is a vector $\boldsymbol{\gamma}_0 \in \mathbb{R}^p$ such that

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}}(\lambda) = \hat{\boldsymbol{\beta}}_{\text{LASSO}}(\lambda_0) - (\lambda - \lambda_0)\boldsymbol{\gamma}_0.$$

Thus the LASSO solution path is linear as λ ranges from λ_0 to λ_1 ([2, Efron *et al.*, 2004]; [4, Rosset and Zhu, 2007]).

[0]Continue with **Computational Part** in next page.

Computational Part

1. (*10 pt*) **Replicating Analysis on Prostate Cancer Data**

   Compare the performance of least squares (LS), best subset selection, ridge regression, LASSO, principal components regression (PCR) and partial least squares (PLS) on the *Prostate Cancer Data*[1]. More information can be found from Section 3.2.1 and 3.3.4 in Textbook. In particular, follow the same procedure on data preparation and replicate the results. Extend the analysis with your own comments.

2. (*15 pt*) **KNN Classification**

   (a) Write a function (from scratch) to perform K-Nearest Neighbors Classifier using various distance functions.

   (b) Apply this function to the digits data *zip.train.gz*.[2] For simplicity, you will need to parse the 3's and 8's from the data. Be sure to split the data into training (60%) set and test set before you apply your function to the data.

   (c) Compare the training and test errors for various K's and various distance functions, and the results of linear regression.

3. (*25 pt*) **Analysis of Baseball Data**

   Install the `ISLR` package in **R** and load the baseball dataset by running the command `data(Hitters)`. The response variable for this problem is `Salary`.

   (a) **Variable Selection:** Fit and visualize regularization paths for the following methods: LASSO, elastic net, adaptive LASSO, SCAD. What are the top predictors selected by each method? Are they different? If so, why?

   (b) **Prediction:** Compare the averaged prediction MSE on the test set for the following methods: least squares, ridge regression, best subset selection, LASSO, elastic net, adaptive LASSO, SCAD. Which types of methods give the best prediction error? Why do these methods perform well? Do any methods seem to overfit the training set? If so, why? Do all the methods choose the same subset of variables? Explain and expand on your discussions.

   Note.

   - As the distribution of `Salary` is skewed, you may need to take the log transform of `Salary`.
   - Remember to remove rows with missing data.
   - Remember to account for the intercept by centering covariates or adding a column of **1**'s whenever necessary. **R** function `model.matrix` may be useful.

[1] Available in https://web.stanford.edu/~hastie/ElemStatLearn/datasets/prostate.data

[2] Available in https://web.stanford.edu/~hastie/ElemStatLearn/datasets/zip.train.gz. Unzip the file to get `zip.train`.

- **R** functions `regsubsets` (from package `leaps`), `glmnet` (with options `alpha`, `penalty.factor`), and `ncvreg` might be useful.
- Clearly report the way in which you call key functions in **R** or the procedures/algorithms by which you realize the key steps. Implicit processing of data (*e.g.* standardization) within given functions should be taken care of.
- Tuning parameters should be determined by cross-validation or performing prediction on a hold-out dataset (validation set) from training set. **k-fold cross-validation procedure:** Fix a tuning parameter/model.
  - Randomly partition the training set into k subsets.
  - For each hold-out subset, use the rest of training set to train the desired model and perform prediction on the hold-out subset to obtain a prediction error.
  - Average the prediction errors for all hold-out subsets as the assessment for the tuning parameter/model.
- Prediction comparisons should be performed on an independent dataset (test set). To avoid randomness from the way you partition the dataset (into training and test sets), you might repeat by multiple times and report the aggregated results.
- Use appropriate tables and plots to organize your discussion.

References

[1] Stephen Boyd and Lieven Vandenberghe. KKT optimality conditions. In *Convex optimization*, chapter 5.5.3, pages 243–246. Cambridge University Press, http://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf, 2004. 4

[2] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression (with discussion. *The Annals of Statistics*, 32(2):407–499, 2004. 4

[3] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Springer-Verlag, https://web.stanford.edu/~hastie/Papers/ESLII.pdf, second edition, 2009. 1

[4] Saharon Rosset and Ji Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, 35(3):1012–1030, 2007. 4