STOR 767 Spring 2019 Hw5

Due on 03/18/2019 in Class

Zhenghan Fang

Remark. This homework focuses on additive models, model assessment and selection, and support vector machines.

Instruction.

- **Theoretical Part and Computational Part are respectively credited 60 points. At most 100 points in total will be accounted for this homework.**
- 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 or modified from the textbook [Friedman et al., 2009].

Theoretical Part

1. (*10 pt*) (Naive Bayes and Logistic GAM, Textbook Ex. 6.9) What's the differences between the naive Bayes model and a generalized additive Logistic regression model in terms of (a) model assumptions, and (b) estimation? If all the variables are discrete, what can you say about the Logistic GAM?

   Model assumptions: The naive Bayes model assumes that features X_k's are independent for a given class. The Logistic GAM assumes that the log-posterior odds are additive functions of the features X_k's.

   Estimation: Class-conditional marginal densities in naive Bayes are estimated separately by one-dimensional density estimates. The logistic GAM is estimated iteratively by backfitting algorithm.

2. (*15 pt*) (Optimism, Textbook Ex. 7.4, 7.5) Let $\mathcal{Y} = \{Y_i\}_{i=1}^n$ be a training sample, $\mathcal{Y}^{\text{new}} = \{Y_i^{\text{new}}\}_{i=1}^n \overset{iid}{=} \mathcal{Y}$ be an independent copy of $\mathcal{Y}$, $\{\widehat{Y}_i\}_{i=1}^n$ be the in-sample prediction based on $\mathcal{Y}$. Recall the in-sample prediction error and its training estimate

$$\text{Err}_{\text{in}} := \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathcal{Y},\mathcal{Y}^{\text{new}}} \ell(Y_i^{\text{new}}, \widehat{Y}_i), \quad \overline{\text{err}} := \frac{1}{n}\sum_{i=1}^n \ell(Y_i, \widehat{Y}_i),$$

   and the optimism

$$\text{op} := \text{Err}_{\text{in}} - \mathbb{E}_{\mathcal{Y}}\overline{\text{err}}.$$

   Consider the squared-error loss $\ell(y, \widehat{y}) := (y - \widehat{y})^2$.

(I) Show that

$$\text{op} = \frac{2}{n}\sum_{i=1}^{n} \mathbf{Cov}_{\mathcal{Y}}(\widehat{Y}_i, Y_i).$$

$$\begin{aligned}
\text{op} &= \text{Err}_{\text{in}} - \mathbb{E}_{\mathcal{Y}}\overline{\text{err}} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{Y},\mathcal{Y}^{\text{new}}}\ell(Y_i^{\text{new}}, \widehat{Y}_i) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{Y}}\ell(Y_i, \widehat{Y}_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{Y},\mathcal{Y}^{\text{new}}}\left[\ell(Y_i^{\text{new}}, \widehat{Y}_i) - \ell(Y_i, \widehat{Y}_i)\right]
\end{aligned}$$

Plug in $\ell(y, \widehat{y}) := (y - \widehat{y})^2$,

$$\begin{aligned}
\text{op} &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{Y},\mathcal{Y}^{\text{new}}}\left[(Y_i^{\text{new}} - \widehat{Y}_i)^2 - (Y_i - \widehat{Y}_i)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{Y},\mathcal{Y}^{\text{new}}}\left[Y_i^{\text{new}2} - 2Y_i^{\text{new}}\widehat{Y}_i - Y_i^2 + 2Y_i\widehat{Y}_i\right]
\end{aligned}$$

Because

$$\mathcal{Y}^{\text{new}} \perp\!\!\!\perp \mathcal{Y},$$

$$\mathbb{E}_{\mathcal{Y}^{\text{new}}} Y_i^{\text{new}2} = \mathbb{E}_{\mathcal{Y}} Y_i^2,$$

$$\mathbb{E}_{\mathcal{Y}^{\text{new}}} Y_i^{\text{new}} = \mathbb{E}_{\mathcal{Y}} Y_i,$$

we get

$$\begin{aligned}
\text{op} &= \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{Y}} Y_i\widehat{Y}_i - \mathbb{E}_{\mathcal{Y}} Y_i \mathbb{E}_{\mathcal{Y}}\widehat{Y}_i \\
&= \frac{2}{n}\sum_{i=1}^{n} \mathbf{Cov}_{\mathcal{Y}}(\widehat{Y}_i, Y_i).
\end{aligned}$$

(II) Assume $\mathbf{Var}(Y_i) = \sigma^2$ $(1 \leqslant i \leqslant n)$. Write $\mathcal{Y}$ in vector form $\boldsymbol{Y} \in \mathbb{R}^n$. Let $\mathbf{S} \in \mathbb{R}^{n\times n}$ be a (fixed) smoother matrix, $\widehat{\boldsymbol{Y}} := \mathbf{S}\boldsymbol{Y}$ be the linear-smoother in-sample prediction vector. Show that

$$\text{op} = \frac{2}{n}\text{Tr}(\mathbf{S})\sigma^2.$$

Let S_{ij} be the element on the ith row and jth column of $\mathbf{S}$.

$$\begin{aligned}
\mathbf{Cov}_{\mathcal{Y}}(\widehat{Y}_i, Y_i) &= \mathbf{Cov}_{\mathcal{Y}}(\sum_{j=1}^{n} S_{ij}Y_j, Y_i) \\
&= \sum_{j=1}^{n} \mathbf{Cov}_{\mathcal{Y}}(S_{ij}Y_j, Y_i) \\
&= \mathbf{Cov}_{\mathcal{Y}}(S_{ii}Y_i, Y_i) \quad (Y_i \perp\!\!\!\perp Y_j, i \neq j) \\
&= S_{ii}\sigma^2.
\end{aligned}$$

Therefore,

$$\text{op} = \frac{2}{n}\sum_{i=1}^{n} \mathbf{Cov}_{\mathcal{Y}}(\widehat{Y}_i, Y_i) = \frac{2}{n}\text{Tr}(\mathbf{S})\sigma^2.$$

3. (*15 pt*) (Bootstrap Prediction Error) Suppose $\mathcal{Y} := \{Y_1 = 1, Y_2 = 2, Y_3 = 6\}$ where $n = 3$. Consider a linear model

$$Y_i = \theta + \epsilon_i \quad (i = 1, 2, 3)$$

with $\epsilon_1, \epsilon_2, \epsilon_3 \overset{iid}{\sim} (0, \sigma^2)$ and squared-error loss $\ell(y, \hat{y}) := (y - \hat{y})^2$.

(a) Consider Bootstrap on $\mathcal{Y}$. Enumerate all possible unordered **Bootstrap bags**[1] and their Bootstrap probabilities. For example, $\{1, 1, 2\}$ is a possible Bootstrap bag with probability 3/27. Indicate the **out-of-bag (OOB)** sample points[2] for each unordered Bootstrap bag.

Bootstrap bag	Probability	OOB
$\{1,1,1\}$	1/27	$\{2,6\}$
$\{2,2,2\}$	1/27	$\{1,6\}$
$\{6,6,6\}$	1/27	$\{1,2\}$
$\{1,1,2\}$	3/27	$\{6\}$
$\{1,1,6\}$	3/27	$\{2\}$
$\{2,2,1\}$	3/27	$\{6\}$
$\{2,2,6\}$	3/27	$\{1\}$
$\{6,6,1\}$	3/27	$\{2\}$
$\{6,6,2\}$	3/27	$\{1\}$
$\{1,2,6\}$	6/27	$\varnothing$

(b) For each Bootstrap sample $\mathcal{Y}_b^*$, derive the least-square prediction rule and its prediction on $\mathcal{Y}$ as $\{\widehat{Y}_{bi}^*\}_{i=1}^n$. Compare the training error $\overline{\text{err}}$, Bootstrap prediction error estimate

$$\text{err}_b^* := \sum_{i=1}^n \ell(Y_i, \widehat{Y}_{bi}^*), \quad \widehat{\text{Err}}_{\text{boot}} := \lim_{B \to +\infty} \frac{1}{nB} \sum_{b=1}^B \text{err}_b^*,$$

and the OOB prediction error estimate[3]

$$\text{err}_{\text{oob},b}^* := \sum_{Y_i \in \mathcal{Y} \setminus \mathcal{Y}_b^*} \ell(Y_i, \widehat{Y}_{bi}^*), \quad p_{\text{oob},n} := \left(1 - \frac{1}{n}\right)^n, \quad \widehat{\text{Err}}_{\text{oob}} := \lim_{B \to +\infty} \frac{1}{n p_{\text{oob},n} B} \sum_{b=1}^B \overline{\text{err}}_{\text{oob},b}^*.$$

Let a Bootstrap sample $\mathcal{Y}_b^* = \{Y_{bi}^*\}_{i=1}^n$. The least-square estimate of θ is

$$\begin{aligned} \hat{\theta}_b &= \arg\min_\theta \sum_{i=1}^n (Y_{bi}^* - \theta)^2 \\ &= \frac{1}{n} \sum_{i=1}^n Y_{bi}^* \end{aligned}$$

The least-square rule's prediction $\widehat{Y}_{bi}^* = \hat{\theta}_b$ for all the samples.

[1] We call a size-n sample with replacement from $\mathcal{Y}$ as a Bootstrap bag.

[2] Let $\mathcal{Y}^*$ be a Bootstrap bag from $\mathcal{Y}$, then $\mathcal{Y} \setminus \mathcal{Y}^*$ is the OOB sample.

[3] $\widehat{\text{Err}}_{\text{oob}}$ is the same as the leave-one-out Bootstrap estimate $\widehat{\text{Err}}^{(1)}$ introduced in Friedman et al. [2009, Equation (7.56)], where they only differ in the order of summation (summing over Bootstrap bags and sample points).

Bootstrap bag	Probability	OOB	$\hat{\theta}_b$	Training error, err_b^*	OOB error, $\mathrm{err}_{\mathrm{oob},b}^*$
$\{1,1,1\}$	1/27	$\{2,6\}$	1	26	26
$\{2,2,2\}$	1/27	$\{1,6\}$	2	17	17
$\{6,6,6\}$	1/27	$\{1,2\}$	6	41	41
$\{1,1,2\}$	3/27	$\{6\}$	4/3	67/3	1176/9
$\{1,1,6\}$	3/27	$\{2\}$	8/3	43/3	4/9
$\{2,2,1\}$	3/27	$\{6\}$	5/3	58/3	169/9
$\{2,2,6\}$	3/27	$\{1\}$	10/3	43/3	49/9
$\{6,6,1\}$	3/27	$\{2\}$	13/3	58/3	49/9
$\{6,6,2\}$	3/27	$\{1\}$	14/3	67/3	121/9
$\{1,2,6\}$	6/27	$\varnothing$	3	14	0

$$\begin{aligned}\widehat{\mathrm{Err}}_{\mathrm{boot}} &:= \lim_{B\to+\infty} \frac{1}{nB}\sum_{b=1}^{B} \mathrm{err}_b^* \\ &= \frac{1}{n}\sum_b \Pr(b)\mathrm{err}_b^* \\ &= 56/3\end{aligned}$$

$$\begin{aligned}\widehat{\mathrm{Err}}_{\mathrm{oob}} &:= \lim_{B\to+\infty} \frac{1}{np_{\mathrm{oob},n}B}\sum_{b=1}^{B} \overline{\mathrm{err}}_{\mathrm{oob},b}^* \\ &= \frac{1}{np_{\mathrm{oob},n}}\sum_b \Pr(b)\mathrm{err}_{\mathrm{oob},b}^* \\ &= 455/18\end{aligned}$$

4. (*20 pt*) (SVM) Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^p \times \{\pm 1\}$ be a training sample. Consider the large-margin linear classification problem

$$\begin{aligned}\max_{(\boldsymbol{w},b)\in\mathbb{R}^{p+1}} \quad & \gamma \\ \text{s.t.} \quad & y_i(b + \boldsymbol{w}^T\boldsymbol{x}_i) \geqslant \gamma \quad (1 \leqslant i \leqslant n) \\ & \|\boldsymbol{w}\|_2 = 1\end{aligned} \tag{1}$$

(a) Show that (1) is equivalent to

$$\begin{aligned}\min_{(\boldsymbol{w},b)\in\mathbb{R}^{p+1}} \quad & \tfrac{1}{2}\|\boldsymbol{w}\|_2^2 \\ \text{s.t.} \quad & y_i(b + \boldsymbol{w}^T\boldsymbol{x}_i) \geqslant 1 \quad (1 \leqslant i \leqslant n)\end{aligned} \tag{2}$$

Let $\boldsymbol{v} = \boldsymbol{w}/\gamma$. (1) is equivalent to

$$\begin{aligned}\max_{(\boldsymbol{v},b)\in\mathbb{R}^{p+1}} \quad & \gamma \\ \text{s.t.} \quad & y_i(\tfrac{b}{\gamma} + \boldsymbol{v}\boldsymbol{x}_i) \geqslant 1 \quad (1 \leqslant i \leqslant n) \\ & \|\boldsymbol{v}\|_2 = \tfrac{1}{\gamma}\end{aligned}$$

equivalent to

$$\begin{aligned} \max_{(\boldsymbol{v},b)\in\mathbb{R}^{p+1}} \quad & \frac{1}{\|\boldsymbol{v}\|_2} \\ \text{s.t.} \quad & y_i(\frac{b}{\gamma} + \boldsymbol{v}\boldsymbol{x}_i) \geqslant 1 \quad (1 \leqslant i \leqslant n) \end{aligned}$$

equivalent to (2).

(b) Introduce Lagrangian variables $\boldsymbol{\alpha} \in \mathbb{R}^n_+$ to inequality constraints in (2) and write down the Lagrangian function $L(\boldsymbol{w}, b; \boldsymbol{\alpha})$ [see Boyd and Vandenberghe, 2004, Chapter 5]. Use strong duality

$$\begin{aligned} & \min_{(\boldsymbol{w},b)\in\mathbb{R}^{p+1}} \max_{\boldsymbol{\alpha}\in\mathbb{R}^n_+} L(\boldsymbol{w}, b; \boldsymbol{\alpha}) = \min_{(\boldsymbol{w},b)\in\mathbb{R}^{p+1}} L_{\mathcal{P}}(\boldsymbol{w}, b) \quad \text{(primal problem (2))} \\ = & \max_{\boldsymbol{\alpha}\in\mathbb{R}^n_+} \min_{(\boldsymbol{w},b)\in\mathbb{R}^{p+1}} L(\boldsymbol{w}, b; \boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}\in\mathbb{R}^n_+} L_{\mathcal{D}}(\boldsymbol{\alpha}) \quad \text{(dual problem (3))} \end{aligned}$$

to derive the Lagrangian dual problem

$$\begin{aligned} \max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \quad & L_{\mathcal{D}}(\boldsymbol{\alpha}) = \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \\ \text{s.t.} \quad & \alpha_i \geqslant 0 \qquad (1 \leqslant i \leqslant n) \\ & \sum_{i=1}^n \alpha_i y_i = 0 \end{aligned} \tag{3}$$

where the primal problem solution given dual optima $\boldsymbol{\alpha}^*$ is

$$\boldsymbol{w}^* = \sum_{i=1}^n \alpha_i^* y_i \boldsymbol{x}_i.$$

The Lagrangian function is

$$L(\boldsymbol{w}, b; \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \sum_{i=1}^n \alpha_i[y_i(b + \boldsymbol{w}^T\boldsymbol{x}_i) - 1].$$

For any fixed α,

$$\frac{\partial L(\boldsymbol{w}, b; \boldsymbol{\alpha})}{\partial w_j} = 0, j = 1, 2, ..., p \implies \boldsymbol{w} = \sum_{i=1}^n \alpha_i y_i \boldsymbol{x}_i$$

$$\frac{\partial L(\boldsymbol{w}, b; \boldsymbol{\alpha})}{\partial b} = 0 \implies \sum_{i=1}^n \alpha_i y_i = 0$$

Plug $\boldsymbol{w} = \sum_{i=1}^n \alpha_i y_i \boldsymbol{x}_i$ and $\sum_{i=1}^n \alpha_i y_i = 0$ into $L(\boldsymbol{w}, b; \boldsymbol{\alpha})$, then we get $L_{\mathcal{D}}(\boldsymbol{\alpha})$.

(c) Use KKT conditions to argue that $\text{supp}(\boldsymbol{\alpha}^*) := \{1 \leqslant i \leqslant n : \alpha_i^* \neq 0\}$ indicates the support vectors. Show how to solve for b^*. What's the support hyperplanes and margin?

KKT conditions include

$$\hat{\alpha}_i[y_i(\hat{b} + \hat{\boldsymbol{w}}^T\boldsymbol{x}_i) - 1] = 0, i = 1, 2, ..., n$$

These imply

$$\text{if } \alpha_i \neq 0, \text{ then } y_i(\hat{b} + \hat{\boldsymbol{w}}^T\boldsymbol{x}_i) = 1,$$

i.e., $\text{supp}(\boldsymbol{\alpha}^*) := \{1 \leqslant i \leqslant n : \alpha_i^* \neq 0\}$ indicates the support vectors.

If $\boldsymbol{x}_i$ is a support vector, then $b^* = 1/y_i - \boldsymbol{w}^{*T}\boldsymbol{x}_i$.

(d) (Kernel Trick) Let K be a positive semidefinite (PSD) kernel on $\mathbb{R}^p$ generating a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$, admitting eigen expansion

$$K(\boldsymbol{x}, \boldsymbol{x}') = \sum_{j=1}^{\infty} \gamma_j \phi_j(\boldsymbol{x}) \phi_j(\boldsymbol{x}'). \quad (\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^p)$$

For any $f \in \mathcal{H}_K$, there exists $\{\theta_j\}_{j=1}^{\infty} \subseteq \mathbb{R}$ such that

$$f(\boldsymbol{x}) = \sum_{j=1}^{\infty} \theta_j \phi_j(\boldsymbol{x}), \quad \|f\|_{\mathcal{H}_K}^2 = \sum_{j=1}^{\infty} \frac{\theta_j^2}{\gamma_j} < +\infty.$$

Consider an RKHS analog to (2)

$$\begin{aligned} \min_{f \in \mathcal{H}_K, b \in \mathbb{R}} \quad & \tfrac{1}{2}\|f\|_{\mathcal{H}_K}^2 \\ \text{s.t.} \quad & y_i[b + f(\boldsymbol{x}_i)] \geqslant 1 \quad (1 \leqslant i \leqslant n) \end{aligned} \tag{4}$$

Show that the Lagrangian dual problem now becomes[4]

$$\begin{aligned} \max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \quad & L_{\mathcal{D}}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ \text{s.t.} \quad & \alpha_i \geqslant 0 \qquad (1 \leqslant i \leqslant n) \\ & \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \tag{5}$$

where the primal problem solution given dual optima $\boldsymbol{\alpha}^*$ is[5]

$$f^*(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i^* y_i K(\boldsymbol{x}_i, \boldsymbol{x}). \quad (\boldsymbol{x} \in \mathbb{R}^p)$$

The Lagrangian function is

$$L(\boldsymbol{w}, b; \boldsymbol{\alpha}) = \frac{1}{2}\|f\|_{\mathcal{H}_K}^2 - \sum_{i=1}^{n} \alpha_i \{y_i[b + f(\boldsymbol{x}_i)] - 1\}.$$

For any fixed α,

$$\frac{\partial L(\boldsymbol{w}, b; \boldsymbol{\alpha})}{\partial \theta_j} = 0, j = 1, 2, \ldots \implies \theta_j = \sum_{i=1}^{n} \alpha_i y_i \gamma_j \phi_j(\boldsymbol{x}_i) \tag{d1}$$

$$\frac{\partial L(\boldsymbol{w}, b; \boldsymbol{\alpha})}{\partial b} = 0 \implies \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{d2}$$

Plug (d1) and (d2) into the Lagrangian function $L(\boldsymbol{w}, b; \boldsymbol{\alpha})$, then we get $L_{\mathcal{D}}(\boldsymbol{\alpha})$.

Plug (d1) into $f(\boldsymbol{x})$, then we get

$$f^*(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i^* y_i K(\boldsymbol{x}_i, \boldsymbol{x}). \quad (\boldsymbol{x} \in \mathbb{R}^p)$$

[4] It greatly reduces the nonlinear problem to simply replace $[\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle]_{n \times n}$ by the kernel matrix $[K(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{n \times n}$ and motivates

[5] It hence also shows that $f^* \in \text{span}\{K(\boldsymbol{x}_i, \cdot)\}_{i=1}^n$, which is a generic result for loss minimization over RKHS [Wahba, 1990, Friedman et al., 2009, Ex 5.15].

Computational Part

1. (*20 pt*) **Backfitting and Coordinate Descent in LASSO** [**Wu and Lange, 2008, Friedman et al., 2010**]

   Recall that the univariate LASSO regression $\{Y_i\}_{i=1}^n$ on standardized regressor $\{X_i\}_{i=1}^n$ with $\sum_{i=1}^n X_i = 0$, $\frac{1}{n}\sum_{i=1}^n X_i^2 = 1$[6] is soft-thresholding

   $$\operatorname*{argmin}_{\alpha,\beta\in\mathbb{R}} \frac{1}{2n}\sum_{i=1}^n (Y_i - \alpha - \beta X_i)^2 + \lambda|\beta| = \left(\bar{Y}, \mathcal{S}(\widehat{\beta}_{\mathrm{LS}}; \lambda)\right)$$

   where $\widehat{\beta}_{\mathrm{LS}} = \frac{1}{n}\sum_{i=1}^n X_i(Y_i - \bar{Y})$ is the ordinary least-square estimate, $\mathcal{S}$ is a soft-thresholding operator

   $$\mathcal{S}(z;\lambda) := \mathrm{sign}(z)(|z|-\lambda)_+ = \begin{cases} z-\lambda, & z > \lambda \\ 0, & -\gamma < z \leqslant \lambda \\ z+\lambda, & z \leqslant -\lambda \end{cases}$$

   Derive the cyclic backfitting algorithm to solve multivariate LASSO regression given a standardized covariate matrix $\mathbf{X} = [X_{ij}]_{n\times p} \in \mathbb{R}^{n\times p}$ with $\sum_{i=1}^n X_{ij} = 0$, $\frac{1}{n}\sum_{i=1}^n X_{ij}^2 = 1$, response vector $\boldsymbol{Y} \in \mathbb{R}^n$ and ℓ^1-regularization parameter $\lambda > 0$. Write an **R** function `lasso` and compare it with `glmnet` on your simulated

   $$n = p = 100, \quad \{X_{ij}\}_{i,j=1}^{100}, \{Y_i\}_{i=1}^{100} \overset{iid}{\sim} \mathcal{N}(0,1), \quad \lambda = 1/10.$$

   Hint. The algorithm derived above is implemented in `glmnet` [Friedman et al., 2010]. In order to get exactly the same result from `glmnet`, standardize the data on your own to avoid internal scaling since `glmnet` would report coefficients in the original scale. Specify `lambda = 1/10` in `glmnet` to avoid internal generated λ sequence. Set the `thresh` option to `1e-20` to get an accurate fit.

2. (*20 pt*) (Textbook Ex. 7.9) **Prostate Cancer Data**

   Carry out a best-subset regression analysis on the *Prostate Cancer Data* as Hw2 has done, while using AIC, BIC, 5-fold and 10-fold CVs, and Bootstrap .632 estimates of prediction error to tune the best size of subsets. Discuss the results.

3. (*20 pt*) **South African Heart Disease Data**

   Perform Support Vector Machine analysis on the *South African Heart Disease Data* with various kernels and compare the prediction performance with the results using LDA, QDA, and Logistic regression in Hw3. Remember to tune the bandwidth parameters in nonlinear kernels using cross-validation.

References

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, http://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf, 2004. 5

[6] It admits with the internal standardization of `glmnet`. Note that `scale` function scales as $\frac{1}{n-1}\sum_{i=1}^n X_i^2 = 1$.

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, 3, 6

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010. 7

Grace Wahba. *Spline models for observational data*, volume 59. Siam, 1990. 6

Tong Tong Wu and Kenneth Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244, 2008. 7