Solution to Homework 3b

Name: Chen Shen

NetID: cs5236

1.

The problem with (a): There is no case for test, so that we would never know the performance of the model we trained. There may also be some risk of overfitting.

The problem with (b): The test error could vary significantly depending on samples selected. Only use limited number of samples for training. Problems particularly bad for data with limited number of samples.

The problem with (c): Need more computing capacity. An accurate result requires K fits of parameters.

$$P(X_i = x|x \text{ in list}) = \frac{1}{n} \times 50\%$$
$$E = \sum_{i=1}^{n} iP(X_i = x|x \text{ in list}) = \sum_{i=1}^{n} i\frac{1}{2n} = \frac{1+n}{4}$$

2.

$$\begin{aligned}
y^k &= \beta^k \boldsymbol{x}^{\mathrm{T}} \\
\Rightarrow \frac{1}{K}\sum_{k=1}^{K} y^k &= \frac{1}{K}\sum_{k=1}^{K}(\beta^k \boldsymbol{x}^{\mathrm{T}}) \\
&= \frac{1}{K}\sum_{k=1}^{K}(\beta^k)\boldsymbol{x}^{\mathrm{T}} \\
&= \bar{\boldsymbol{\beta}}\boldsymbol{x}^{\mathrm{T}}
\end{aligned}$$

3.

4.

Filtering method, wrapper method, embedded method, forward stepwise algorithm.

5.

(a) The mean is

$$\begin{aligned}
\frac{1}{J}\sum_i y_i &= \frac{1}{J}\sum_i \frac{y_i^r - \bar{y}}{\sigma_y} \\
&= \frac{1}{J}\frac{\sum_i (y_i^r - \bar{y})}{\sigma_y} \\
&= \frac{1}{J}\frac{\sum_i y_i^r - \sum_i \bar{y}}{\sigma_y} \\
&= 0
\end{aligned}$$

The variance is

$$\begin{aligned}
\frac{1}{J}\sum_i (y_i - 0)^2 &= \frac{1}{J}\sum_i y_i^2 \\
&= \frac{1}{J}\sum_i [\frac{y_i^r - \bar{y}}{\sigma_y}]^2 \\
&= \frac{1}{J}\sum_i \frac{(y_i^r - \bar{y})^2}{\sigma_y^2} \\
&= \frac{1}{J}\frac{\sum_i (y_i^r - \bar{y})^2}{\sigma_y^2} \\
&= \frac{1}{J}\frac{J \cdot \sigma_y^2}{\sigma_y^2} \\
&= 1
\end{aligned}$$

(b) Since both the target y and the features x_j are normalized, we have

$$\sum \hat{y} = 0.$$

So

$$\sum \beta_0 + \sum \beta_1 x_1 + \cdots + \sum \beta_J x_J = 0.$$

Because

$$\sum x_j = 0,$$

there should be

$$\sum \beta_0 = 0,$$

i.e.

$$\beta_0 = 0.$$

(c) The formula should be

$$\beta_j^r = (\beta_j \cdot \sigma_j) + x_{i,j}^r,$$

which is the inverse of normalization.

6.

Without regularization, large positive and negative coefficients cancel each other for correlated features, resulting in high variance of the resulting models.

7.

Ridge Regression is raised to solve muticolinearity. It uses L2 norm to simplify the calculation. However, it cannot shrink parameters to zero. So it can not be used to do feature selection.

LASSO Regression is raised to do feature selection. It uses L1 norm. So the calculation is more complex and there is no analytical solution. However, it can shrink some parameters to zero so as to select some certain features.

8.

First, we do differentiating with respect to β. Then set the result to be zero.

$$
\begin{aligned}
\frac{\partial J(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= 2\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\beta} - 2\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} + 2\alpha\boldsymbol{\beta} \\
&= \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} + \alpha\boldsymbol{\beta} \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
&\Rightarrow (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} + \alpha\boldsymbol{I})\boldsymbol{\beta}_{\mathrm{opt}} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} \\
&\Rightarrow \boldsymbol{\beta}_{\mathrm{opt}} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} + \alpha\boldsymbol{I})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}
\end{aligned}
$$