Homework 1

Dai Yuehao (1800010660@pku.edu.cn)

October 12, 2020

1 No free lunch

For a loss function $l(f,h)$, we hope to have

$$\begin{aligned}\sum_f E_{ote}(\mathscr{L}_a|X,f) &= \sum_f \sum_h \sum_{x\in\mathcal{X}-X} P(x) l(f(x),h(x)) P(h|X,\mathscr{L}_a) \\ &= \sum_{x\in\mathcal{X}-X} P(x) \sum_h P(h|X,\mathscr{L}_a) \sum_f l(f(x),h(x)) \\ &= \sum_f l(f(x),h(x)) \sum_{x\in\mathcal{X}-X} P(x) \sum_h P(h|X,\mathscr{L}_a) \\ &= \sum_f l(f(x),h(x)) \sum_{x\in\mathcal{X}-X} P(x),\end{aligned}$$

thus l need to satisfy that $\sum_f l(f(x),h(x))$ is the same for any given h. An assumption is that l has the form

$$l(f,h) = g(|f-h|), \quad \sup|g| < \infty,$$

under this circumstance

$$\sum_f l(f(x),h(x)) = \sum_f [g(1)\mathbb{I}(h(x) \neq f(x)) + g(0)\mathbb{I}(f(x) = g(x)) = 2^{|\mathcal{X}|-1}[g(0)+g(1)].$$

Otherwise, we can choose

$$h_1 = \arg\max \sum_f l(f(x),h(x)), \quad h_2 = \arg\min \sum_f l(f(x),h(x)),$$

let algorithm a yield h_1 with probability 1, algorithm b yield h_2 with probability 1, then we have

$$E_{ote}(\mathscr{L}_a|X,f) > E_{ote}(\mathscr{L}_b|X,f).$$

2 ESL 3.4

Let Z_i denote the vector obtained at the i^{th} step of the Gram-Schmidt procedure, then we have

$$Z_i = X_i - \sum_{k=0}^{i} \frac{X_i^T Z_k}{Z_k^T Z_k} Z_k,$$

thus Z_p is the only vector that contains X_p. Now from $Z_0 = X_0$, we regress y on $Z_0, Z_1, \cdots, Z_{i-1}$ and obtain the residual r_i, then regress r_i on Z_i and obtain the coefficient $\hat{\beta}_i'$. Eventually we have $\hat{\beta}_p' = \hat{\beta}_p$, here $\hat{\beta}_p$ is the coefficient of X_p. Now we reverse the procedure, we remove X_p from the combination of $\{Z_i\}$ and the rest of the combination of $\{Z_i\}$ must be the regression of y on $\{Z_0, \cdots, Z_{p-1}\}$(since they are orthogonal), hence the coefficient of Z_{p-1} is exact the coefficient of X_{p-1}. We continue this procedure until we get all the coefficients.

On the other hand, with the QR decomposition of X we have

$$\hat{\beta}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y = \boldsymbol{R}^{-1}\boldsymbol{Q}y,$$

thus

$$R\hat{\beta} = Q^T y, \tag{2.0.1}$$

equation (2.0.1) is easy to solve because $\boldsymbol{R}$ is upper triangular so that we can solve the equaiton by back-substitution, just the same as we show above.

3 ESL 3.6

In terms of Bayes formula we have

$$p(\beta|y) = \frac{p(y|\beta)p(\beta)}{p(y)} \propto p(y|\beta)p(\beta)$$

which has normal distribution, hence the mean of the posterior distribution is the maximal of the density function. Now we take the logarithm and we need to

$$\min_{\beta} \frac{(y - \boldsymbol{X}\beta)^T(y - \boldsymbol{X}\beta)}{2\sigma^2} + \frac{\beta^T\beta}{2\tau},$$

by taking the derivative of β and let it be zero we have

$$-\frac{\boldsymbol{X}^T\boldsymbol{X}\hat{\beta}}{\sigma^2} - \frac{\hat{\beta}}{\tau} + \frac{\boldsymbol{X}^T y}{\sigma^2} = 0,$$

hence we obtain

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T y, \quad \lambda = \frac{\sigma^2}{\tau}.$$

Thereby we conclude the proof, the relationship between the regularization parameter λ in the ridge formula, and the variances τ and σ^2 is the equation $\lambda = \sigma^2/\tau$.

4 ESL 3.8

We consider

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}, \quad \tilde{\boldsymbol{X}} = \boldsymbol{U}\Sigma\boldsymbol{V}^T$$

where $\boldsymbol{R}$ is upper triangular, hence we have

$$Q_1 = \left(\frac{1}{\sqrt{N}}, \cdots, \frac{1}{\sqrt{N}}\right)^T,$$

hence $\boldsymbol{Q}_2$ forms an orthogonal complement of $\mathbf{1}$ in the column space of $\boldsymbol{X}$. On the other hand we have

$$\mathbf{1}^T\tilde{X}_i = \mathbf{1}^T\left(X_{i+1} - \frac{1}{N}\mathbf{1}^T X_{i+1}\mathbf{1}\right) = 0,$$

since $\text{rank}(\tilde{\boldsymbol{X}}) = p$, we know that $\tilde{\boldsymbol{X}}$ forms an orthogonal complement of $\mathbf{1}$ in the column space of $\boldsymbol{X}$. In terms of the nature of SVD decomposition we know that the column space of $\boldsymbol{U}$ is the same as $\tilde{\boldsymbol{X}}$'s, thereby we conclude the proof.

If $\boldsymbol{Q}_2 = \boldsymbol{U}$, we know that $\boldsymbol{Q}_2$ is the Gran-Schmidt orthogonal of $\tilde{\boldsymbol{X}}$ with regularization, hence we have $\tilde{\boldsymbol{X}} = \boldsymbol{Q}_2\boldsymbol{R}_2$ and $\boldsymbol{R}_2 = \boldsymbol{\Sigma}\boldsymbol{V}^T$, thus $\boldsymbol{R}_2$ is both upper triangular and orthogonal, which forces $\boldsymbol{R}_2$ to be diagonal. Thereby we claim that $\boldsymbol{Q}_2$ and $\boldsymbol{U}$ are the same when $\tilde{\boldsymbol{X}}$ is orthogonal.

5 ESL 3.9

At this time we know that r is orthogonal with the column space spaned by $\boldsymbol{X}_1$ hence the variable that will reduce the RSS the most is the one whose vertical component of the column space spaned by $\boldsymbol{X}_1$, denoted by u, is the most parellel with r, hence we need to maximize

$$\frac{|u_j^T r|}{||u_j||_2}, \quad q < j \leq p.$$

Now suppose we have $\boldsymbol{X}_1 = \boldsymbol{Q}\boldsymbol{R}$, let z_i denote the i^{th} column vector of $\boldsymbol{Q}$ and we denote

$$u_j = x_j - \sum_{i=1}^{q}(z_i^T x_j)z_i, \quad q < j \leq p,$$

then

$$v_j = \frac{u_j}{||u_j||},$$

then the reduce of RSS is $||(y^T v_j)v_j||^2 = (y^T v_j)^2$, hence we can solve the equivalent problem

$$j^* = \arg\max_j (y^T v_j)^2,$$

then x_{j^*} is the variable we want.

6 ESL 3.11

We derive the results from minimizing

$$\text{RSS}(\boldsymbol{B}) = \text{tr}[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})] = \text{tr}[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})^T].$$

This is a convex optimization problem, in terms of the one order condition and since $\boldsymbol{\Sigma}^{-1}$ is positive definite we can write $\boldsymbol{\Sigma}^{-1} = \boldsymbol{S}^T\boldsymbol{S}$, now we take the derivative of $\boldsymbol{B}\boldsymbol{S}^T$ and let

$$(\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{B}}\boldsymbol{S}^T - \boldsymbol{X}^T\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{Y}\boldsymbol{S}^T = 0$$

yielding that

$$\hat{\boldsymbol{B}}\boldsymbol{S}^T = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{S}^T,$$

since $\boldsymbol{S}^T$ is non-singular we have

$$\hat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

If the covarience matrices are different, we need to minimize

$$\text{RSS}(\boldsymbol{B}) = \sum_{i=1}^{N}(y_i - x_i\boldsymbol{B})^T\boldsymbol{\Sigma}_i^{-1}(y_i - x_i\boldsymbol{B}),$$

however this time we can no longer write a closed-form solution by taking the derivative of $\boldsymbol{X}\boldsymbol{S}^T$ as before since $\boldsymbol{\Sigma}_i^{-1}$ varies. In spite of this, we can still take the derivative of $\boldsymbol{B}$ and obtain

$$\sum_{i=1}^{N}(x_i^T\boldsymbol{S}_i^T\boldsymbol{S}_i x_i)\boldsymbol{B} - \sum_{i=1}^{N}\boldsymbol{X}^T\boldsymbol{S}_i^T\boldsymbol{S}_i y_i = 0$$

and use some numerical methods to solve this equation.

7 ESL 3.12

Then we have

$$\tilde{\boldsymbol{X}} = \begin{pmatrix}\boldsymbol{X}\\ \sqrt{\lambda}\boldsymbol{I}_p\end{pmatrix}, \quad \tilde{y} = \begin{pmatrix} y \\ \boldsymbol{0}\end{pmatrix}$$

thus the least square solution is

$$\hat{\beta} = (\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{y} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T y.$$

8 ESL 3.29

For a single variable X and response y, the result is

$$a = (X^TX + \lambda)^{-1}X^Ty = \frac{\sum_{i=1}^N x_i y_i}{\lambda + \sum_{i=1}^N x_i^2}.$$

Now we include a copy of X and

$$a_{(2)} = (\boldsymbol{X}_{(2)}^T\boldsymbol{X}_{(2)} + \lambda\boldsymbol{I}_2)^{-1}\boldsymbol{X}_{(2)}^T y, \quad \boldsymbol{X}_{(2)} = \begin{pmatrix} X & X\end{pmatrix}$$

With some calculation we have

$$a_{(2)} = \begin{pmatrix} \frac{\sum_{i=1}^N x_i y_i}{\lambda + 2\sum_{i=1}^N x_i^2} \\ \\ \frac{\sum_{i=1}^N x_i y_i}{\lambda + 2\sum_{i=1}^N x_i^2}\end{pmatrix}$$

hence both coefficients are identical.

In general if there is m copies of the variable X, we also have

$$a_{(m)} = (\boldsymbol{X}_{(m)}^T\boldsymbol{X}_{(m)} + \lambda\boldsymbol{I}_2)^{-1}\boldsymbol{X}_{(m)}^T y,$$

and it is very easy to solve the equation and obtain that each element of $a_{(m)}$ is

$$\frac{\sum_{i=1}^N x_i y_i}{\lambda + m\sum_{i=1}^N x_i^2}.$$