Statistical Learning HW 1

Johnson Lin

September 23 2019

1 Problem 1

Prove the No Free Lunch theorem for any loss function $\ell(h, f)$

We know that,

$$\sum_f E_{ote}(\mathcal{L}_a|X, f) = \sum_f \sum_h \sum_{x\in\chi-X} P(x)\ell(h, f)P(h|X, \mathcal{L}_a)$$

In order to prove the No Free Lunch theorem, we need to show that the loss function isn't a factor after simplifying the right side of the equation. We can use Fubini's theorem to rearrange the summations,

$$= \sum_{x\in\chi-X} P(x) \sum_h P(h|X, \mathcal{L}_a) \sum_f \ell(h, f)$$

In this case, let us use the Means Squared Error as our loss function so

$$\ell(h, f) = \frac{1}{n}\sum_{x=1}^{n}(h(x) - f(x))^2$$

where n is the total number of data points.

Our equation becomes

$$= \sum_{x\in\chi-X} P(x) \sum_h P(h|X, \mathcal{L}_a) \sum_f \frac{1}{n}\sum_{x=1}^{n}(h(x) - f(x))^2$$

$$= \sum_{x\in\chi-X} P(x) \sum_h P(h|X, \mathcal{L}_a) n * \frac{1}{n}\sum_{x=1}^{n}(h(x) - f(x))^2$$

$$= \sum_{x\in\chi-X} P(x) \sum_h P(h|X, \mathcal{L}_a) \sum_{x=1}^{n}(h(x) - f(x))^2$$

$$= \sum_{x\in\chi-X} P(x) * 1 * \sum_{x=1}^{n}(h(x) - f(x))^2$$

$$= \sum_{x \in \chi - X} P(x)(\chi - X)^2$$

This concludes the proof. Intuitively, it makes sense that it works for any loss function. The No Free Lunch theorem states

$$\sum_f P(d_m^y | f, m, a_1) = \sum_f P(d_m^y | f, m, a_2)$$

Essentially, when summed over all elements, no single learning algorithm will outperform another. Thus, if all algorithms perform the same then it doesn't matter what loss function you use since loss functions are simply a method to measure how well an algorithm performed.

2 Problem 2

Exercise 3.4

Given equations (3.30) and (3.31) we know that

$$X = Z\Gamma$$
$$X = ZD^{-1}D\Gamma$$
$$X = QR$$

We can plug this in to the equation

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$
$$(X^T X)\hat{\beta} = X^T Y$$
$$X\hat{\beta} = Y$$
$$QR\hat{\beta} = Y$$
$$R\hat{\beta} = Q^{-1}Y$$

Because Q is an orthogonal matrix,

$$R\hat{\beta} = Q^T Y$$

Because R is an upper triangular matrix,

$$R_{pp}\hat{\beta}_p = \langle q_p, y \rangle$$
$$||z_p||\hat{\beta}_p = ||z_p||^{-1}\langle z_p, y \rangle$$
$$\hat{\beta}_p = \frac{\langle z_p, y \rangle}{||z||^2}$$

Now that we have found $\hat{\beta}$, we can find all β_j. Thus we can obtain all least square coefficients through a single pass of the Gram-Schimdt procedure

3 Problem 3

Exercise 3.11

In order to find the solution to equation (3.39), we can take the partial derivative with respect to β of equation (3.38) and set that equal to zero.

$$RSS(\beta) = tr[(Y - X\beta)^T(Y - X\beta)]$$
$$RSS(\beta) = tr[(Y - X\beta)(Y - X\beta)^T]$$
$$\frac{d}{d\beta} = tr[(XA)^T(Y - X\beta)] = 0$$

where A is a $(p + 1 \times k)$ matrix. Let A be equal to the zero matrix except a 1 in row j and column k.

$$\sum_i x_{ij}(y_{ik} - \sum S x_{is} b_{sk}) = 0$$

$$X^T(Y - X\hat{\beta}) = 0$$
$$(X^TX)^{-1}X^T(Y - X\hat{\beta}) = (X^TX)^{-1}0$$
$$(X^TX)^{-1}(X^TY - X^TX\hat{\beta}) = 0$$
$$(X^TX)^{-1}X^TY - (X^TX)^{-1}X^TX\hat{\beta} = 0$$
$$(X^TX)^{-1}X^TY - \hat{\beta} = 0$$
$$(X^TX)^{-1}X^TY = \hat{\beta}$$
$$RSS(\beta, \Sigma) = tr[(Y - X\beta)\Sigma^{-1}(Y - X\beta)^T]$$

Since Σ is the covariance, we know that it is a positive definite symmetric matrix. Because it is a positive definitive symmetric matrix, there exists some $k \times k$ positive definite symmetric square root matrix. We can call that matrix S.

$$\Sigma^{-1} = S^{1/2}$$

We can replace Y with YS and B with BS, so

$$= tr[(YS - X\beta S)S^{1/2}(YS - X\beta S)^T]$$
$$= tr[S^2S^{1/2}(Y - X\beta)(Y - X\beta)^T]$$
$$= tr[S(Y - X\beta)(Y - X\beta)^T]$$

$$\hat{\beta}S = (X^TX)^{-1}X^TYS$$
$$\hat{\beta} = (X^TX)^{-1}X^TY$$

This concludes the proof. If Σ_i varies, then this is no longer applicable. However, as long as correlations are known, then the argmin can be found.

4 Problem 4

Exercise 3.5 We are given the equation

$$\hat{\beta}^c = argmin_{\beta^c}\{\sum_{i=1}^{N}[y_i - \beta_0^c - \sum_{j=1}^{p}(x_{ij} - \bar{x}_j\beta)j^c]^2 + \lambda\sum_{j=1}^{p}\beta_j^{c2}\}$$

$$= argmin_{\beta^c}\{\sum_{i=1}^{N}[y_i - \beta_0^c - \sum_{j=1}^{p}x_{ij}\beta j^c - \sum_{j=1}^{p}\bar{x}_j\beta j^c]^2 + \lambda\sum_{j=1}^{p}\beta_j^{c2}\}$$

Equation (3.41) is

$$\hat{\beta}^{ridge} = argmin_{\beta}\{\sum_{i=1}^{N}[y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta j]^2 + \lambda\sum_{j=1}^{p}\beta_j^2\}$$

Thus, in order for the two equations to be equivalent, the correspondence between β_0 and β_0^c must be

$$\beta_0 = \beta_0^c + \sum_{j=1}^{p}\bar{x}_j\beta j^c$$

This means that β_0 and β_0^c are equivalent, just that β_0^c is shifted. Thus all of the β^c have the same slope as their β counterpart. Because the lasso and ridge regression's only difference is absolute value of β and β squared respectively, this holds true for the lasso as well.

5 Problem 5

Exercise 3.6

Bayesian statistics states that the posterior distribution is proportional to the prior multiplied by the likelihood. Our prior and likelihood functions are

$$p(\beta) \sim N(0, \tau I)$$

$$p(D|\beta) \sim N(y - X\beta, \sigma^2 I)$$

The log transformation of the posterior distribution is

$$log(p(\beta|D)) = log(p(D|\beta)) + log(p(\beta))$$

$$= \frac{1}{2}\frac{(y - X\beta)^T(y - XB)}{\sigma^2} + \frac{1}{2}\frac{\beta^T\beta}{\tau}$$

In this case, if we multiply the whole thing by $2\sigma^2$

$$= (y - X\beta)^T(y - XB) + \frac{2\sigma^2}{\tau}\beta^T\beta$$

we get equation (3.43) where $\lambda = \frac{2\sigma^2}{\tau}$. Thus proving the equivalence to ridge regression. Our posterior distribution is a normal distribution since we used a normal prior and normal likelihood. In a normal distribution, the mean is equivalent to the mode. Thus the ridge regression estimate is both the mean and mode of the posterior distribution.

6 Problem 6

Exercise 3.12

By augmenting the matrix X and Y, we have two new matrices which we will call X' and Y'

$$X' = \begin{bmatrix} X \\ \sqrt{\lambda} I_{p\times p} \end{bmatrix}$$

$$Y' = \begin{bmatrix} Y \\ 0_{p\times 1} \end{bmatrix}$$

The least square solution is

$$\hat{\beta}_{LS} = (X'^T X')^{-1} X'^T Y$$

$$= (\begin{bmatrix} X^T & \sqrt{\lambda} I_{p\times p} \end{bmatrix} \begin{bmatrix} X \\ \sqrt{\lambda} I_{p\times p} \end{bmatrix})^{-1} \begin{bmatrix} X^T & \sqrt{\lambda} I_{p\times p} \end{bmatrix} \begin{bmatrix} Y \\ 0_{p\times 1} \end{bmatrix}$$

$$= (X^T X + \lambda I_{p\times p})^{-1} X^T Y$$

which is the regularized least squares equation.

7 Problem 7

Exercise 3.16

In the orthogonal case, $X^T X = 1$ which means that the least squares coefficient becomes

$$\hat{\beta} = (X^T X)^{-1} X^T y = X^T y$$

Because best-subset selection takes the model with predictors that had the best least-squares regression score, we can say that

$$\hat{\beta}^{bs} = \hat{\beta}(M = p) = X^T y$$

This means the coefficients are identical, even if $M \leq p$.

For ridge regression,

$$\begin{aligned}\hat{\beta}^{ridge} &= (X^TX + \lambda I)^{-1}X^Ty \\ &= (I + \lambda I)^{-1}X^Ty \\ &= \frac{\hat{\beta}}{(1+\lambda)}\end{aligned}$$

For lasso,

$$\hat{\beta}^{lasso} = (y - X\beta)^T(y - X\beta) + \lambda|\beta|$$

In order to find the estimator, we take the partial derivative with respect to β,

$$\frac{d}{d\beta} = -X^Ty + X^TX\beta + \lambda \times sign(\beta)$$

$$= -X^Ty + \beta + \lambda \times sign(\beta)$$

and then set it equal to 0 and solve for β

$$\begin{aligned}0 = -X^Ty + X^TX\beta + \lambda \times sign(\beta) \\ X^Ty - \lambda \times sign(\beta) = \beta \\ = sign\beta(|X^Ty| - \lambda)\end{aligned}$$

8 Problem 8

Exercise 3.28

Equation (3.51) states

$$\hat{\beta}^{lasso} = argmin_{\beta}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j)^2$$

$$\text{subject to}\sum_{j=1}^{[}|\beta_j| \leq t$$

We are given that for some variable X_j, the corresponding lasso coefficient $\hat{\beta}_j = a$. In this case, we single out this variable from the summation. Let $p - j$ denote all elements excluding the corresponding variables to j.

$$\hat{\beta}^{lasso} = argmin_{\beta}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}x_{ip-j}\beta_{p-j} - x_j\beta_j)^2$$

$$\text{subject to}\sum_{j=1}^{[}|\beta_{p-j}| + |\beta_j| \leq t$$

Now we have an identical copy $X_j^* = X_j$ with a corresponding β_j^*.

$$\hat{\beta}^{lasso} = argmin_\beta \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ip-j}\beta_{p-j} - x_j\beta_j - x_j^*\beta_j^*)^2$$

$$\text{subject} to \sum_{j=1}^{[} |\beta_{p-j}| + |\beta_j| + |\beta_j^*| \leq t$$

We know that the sum of two units absolute valued must be equal to or greater than the absolute value of the sum of two units. So we can use this inequality

$$\sum_{j=1}^{[} |\beta_{p-j}| + |\beta_j + \beta_j^*| \leq t$$

Because we are using the same t, we know that

$$\beta_j + \beta_j^* \leq a$$