CSCC11 Assignment 1

Jefferson Li (jeffersonli.li@mail.utoronto.ca)

February 1, 2021

Question 1.

- a) Given training data denoted by

$$\{(x_i, y_i)_{i=1}^N\}$$

let $\mathbf{w} = [w_0, w_1, \dots, w_k]^T$

let $B = \begin{bmatrix} 1 & b_1(x_1) & b_2(x_1) & \dots & b_k(x_1) \\ 1 & b_1(x_2) & b_2(x_2) & \dots & b_k(x_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & b_1(x_N) & b_2(x_N) & \dots & b_k(x_N) \end{bmatrix}$

let $\mathbf{y} = [y_1, \dots, y_N]^T$

Then the LS objective can be formulated as such

$$y = f(x) = w_0 + \Sigma_{k=1}^K w_k b_k(x) \tag{1}$$
$$E(\mathbf{w}) = \Sigma_{i=1}^N (y_i - f(x_i))^2 \quad \text{Definition of LS objective} \tag{2}$$
$$E(\mathbf{w}) = \Sigma_{i=1}^N (y_i - w_0 - \Sigma_{k=1}^K w_k b_k(x_i))^2 \tag{3}$$
$$E(\mathbf{w}) = ||\mathbf{y} - B\mathbf{w}||_2^2 \quad \text{Replace with matrix norm representation} \tag{4}$$
$$E(\mathbf{w}) = (\mathbf{y} - B\mathbf{w})^T(\mathbf{y} - B\mathbf{w}) \quad \text{Property of Euclidean norm} \tag{5}$$
$$E(\mathbf{w}) = (\mathbf{y}^T - \mathbf{w}^T B^T)(\mathbf{y} - B\mathbf{w}) \quad \text{Properties of Transpose} \tag{6}$$
$$E(\mathbf{w}) = (\mathbf{y}^T\mathbf{y} - \mathbf{y}^T B\mathbf{w} - \mathbf{w}^T B^T\mathbf{y} + \mathbf{w}^T B^T B\mathbf{w}) \tag{7}$$
$$E(\mathbf{w}) = (\mathbf{y}^T\mathbf{y} - 2\mathbf{w}^T B^T\mathbf{y} + \mathbf{w}^T B^T B\mathbf{w}) \quad y^T B w = w^T B^T y \text{ as they are both } [y_i B_{ij} x_j] \tag{8}$$

- b)

$$E(\mathbf{w}) = (\mathbf{y}^T\mathbf{y} - 2\mathbf{w}^T B^T\mathbf{y} + \mathbf{w}^T B^T B\mathbf{w}) \tag{9}$$
$$\text{Gradient of E} = \frac{\partial E}{\partial \mathbf{w}} = (0 - 2B^T\mathbf{y} + (B^T B + B^T B)\mathbf{w}) \quad \text{From matrix identities 5a, 5b} \tag{10}$$
$$\frac{\partial E}{\partial \mathbf{w}} = -2B^T\mathbf{y} + 2B^T B\mathbf{w} \tag{11}$$

- c) The optimal weight vector $\mathbf{w}$ is obtained by solving for $\mathbf{w}$ when

$$\frac{\partial E}{\partial \mathbf{w}} = 0$$

$$-2B^T\mathbf{y} + 2B^T B\mathbf{w} = 0 \tag{12}$$
$$B^T B\mathbf{w} = B^T\mathbf{y} \tag{13}$$
$$(B^T B)^{-1} B^T B\mathbf{w} = (B^T B)^{-1} B^T\mathbf{y} \tag{14}$$
$$\mathbf{w}^* = (B^T B)^{-1} B^T\mathbf{y} = [w_0^*, w_1^*, \dots, w_k^*] \tag{15}$$

$$\implies \text{The optimal weight vector w is } \mathbf{w}^*$$

Question 2.

- a) Q1(c) will not be unique if the columns of B are not linearly independent.

  Proof. let B have dimensions m by n
  Suppose Columns of B are not linearly independent

$$\implies rank(B) < min(m, n) \tag{16}$$
$$\implies rank(B^TB) < min(m, n) \quad \text{Since } rank(B) = rank(B^TB) \text{ by (2f)} \tag{17}$$
$$\implies B^TB \text{ is not invertible, and therefor has linearly dependent cols} \tag{18}$$
$$\implies \text{for } (B^TB)\mathbf{w} = B^T\mathbf{y},\ \mathbf{w} \text{ has infinite solutions} \tag{19}$$

  □

  An example is if
$$K = 2, b_1 = x^2, b_2 = 2x^2$$
  Note that column 1 and 2 of B are linearly dependent as column 2 of B is simply twice of column 1

  Suppose the optimal weights are
$$w_0 = 1, w_1 = 2, w_2 = 3$$

  So the optimal model is as follows
$$f(x) = 1 + 2b_1(x) + 3b_2(x) = 1 + 2x^2 + 6x^2 = 1 + 8x^2$$
  This exact model can also be generated with the weights
$$w_0 = 1, w_1 = 8, w_2 = 0$$
  as
$$f(x) = 1 + 8b_1(x) + 0b_2(x) = 1 + 8x^2 + 0x^2 = 1 + 8x^2$$
  as we can see, if columns of B are not linearly independent, there is an infinite number of weights that correspond to the same model.

- b)Using the same $\mathbf{w}, B, \mathbf{y}$ definitions from Q1(a)

$$y = f(x) = w_0 + \Sigma_{k=1}^{K} w_k b_k(x) \tag{20}$$

$$E(\mathbf{w}) = ||\mathbf{y} - B\mathbf{w}||_2^2 + \lambda||\mathbf{w}|_2^2 \quad \text{Definition of Regularized LS objective} \tag{21}$$

$$E(\mathbf{w}) = (\mathbf{y}^T\mathbf{y} - 2\mathbf{w}^T B^T \mathbf{y} + \mathbf{w}^T B^T B\mathbf{w} + \lambda \mathbf{w}^T\mathbf{w}) \quad \text{From the result of Q1(a)} \tag{22}$$

$$E(\mathbf{w}) = (\mathbf{y}^T\mathbf{y} - 2\mathbf{w}^T B^T \mathbf{y} + \mathbf{w}^T [B^T B + \lambda I]\mathbf{w}) \tag{23}$$

$$\text{Gradient of E} = \frac{\partial E}{\partial \mathbf{w}} = (0 - 2B^T\mathbf{y} + ((B^T B + \lambda I) + (B^T B + \lambda I)^T)\mathbf{w}) \quad \text{From matrix identities 5a, 5b} \tag{24}$$

$$\frac{\partial E}{\partial \mathbf{w}} = (0 - 2B^T\mathbf{y} + (2B^T B + 2\lambda I)\mathbf{w}) \tag{25}$$

$$\frac{\partial E}{\partial \mathbf{w}} = 0 \implies 2B^T\mathbf{y} = (2B^T B + 2\lambda I)\mathbf{w} \tag{26}$$

$$\implies \mathbf{w}^* = (B^T B + \lambda I)^{-1} B^T \mathbf{y} \tag{27}$$

The regularization helps insure $\mathbf{w}^*$ is a unique value.
Note that it is given that $\lambda > 0$
Since $B^T B$ is and is positive semidefinite, its eigenvalues are greater than or equal to zero (from the Linear Algebra Review and Reference pdf). Therefor $B^T B + \lambda I$ has all greater than zero eigenvalues. Which implies $B^T B + \lambda I$ is invertible.

Finally, this means

$$\mathbf{w}^* = (B^T B + \lambda I)^{-1} B^T \mathbf{y}$$

has one unique solution.

- c) let $\hat{B} = \begin{bmatrix} 1 & b_1(x_1) & b_2(x_1) & \dots & b_k(x_1) \\ 1 & b_1(x_2) & b_2(x_2) & \dots & b_k(x_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & b_1(x_N) & b_2(x_N) & \dots & b_k(x_N) \\ \sqrt{\lambda} & & & & \\ & \sqrt{\lambda} & & 0 & \\ & & \ddots & & \\ & 0 & & \sqrt{\lambda} & \\ & & & & \sqrt{\lambda} \end{bmatrix}$ let $\hat{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ let $\mathbf{w} = \begin{bmatrix} w_0 \\ \vdots \\ w_k \end{bmatrix}$

Proof. Show $E(\mathbf{w}) = ||\hat{y} - \hat{B}\mathbf{w}||_2^2$ is equivalent to $E(\mathbf{w}) = ||\mathbf{y} - B\mathbf{w}||_2^2 + \lambda||\mathbf{w}||_2^2$

$$E(\mathbf{w}) = ||\hat{y} - \hat{B}\mathbf{w}||_2^2 \tag{28}$$

$$E(\mathbf{w}) = || \begin{bmatrix} y_1 \\ \vdots \\ y_N \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} 1 & b_1(x_1) & b_2(x_1) & \dots & b_k(x_1) \\ 1 & b_1(x_2) & b_2(x_2) & \dots & b_k(x_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & b_1(x_N) & b_2(x_N) & \dots & b_k(x_N) \\ \sqrt{\lambda} & & & & \\ & \sqrt{\lambda} & & 0 & \\ & & \ddots & & \\ & 0 & & \sqrt{\lambda} & \\ & & & & \sqrt{\lambda} \end{bmatrix} \begin{bmatrix} w_0 \\ \vdots \\ w_k \end{bmatrix} ||_2^2 \tag{29}$$

$$E(\mathbf{w}) = \Sigma_{i=1}^{N}(y_i - w_0 - \Sigma_{k=1}^{K} w_k b_k(x_i))^2 + \Sigma_{k=0}^{K}(0 - \sqrt{\lambda} w_k)^2 \tag{30}$$

$$E(\mathbf{w}) = ||\mathbf{y} - B\mathbf{w}||_2^2 + \Sigma_{k=0}^{K}(0 - \sqrt{\lambda} w_k)^2 \quad \text{By definition of LS Objective} \tag{31}$$

$$E(\mathbf{w}) = ||\mathbf{y} - B\mathbf{w}||_2^2 + \lambda \Sigma_{k=0}^{K}(w_k)^2 \tag{32}$$

$$E(\mathbf{w}) = ||\mathbf{y} - B\mathbf{w}||_2^2 + \lambda ||(\mathbf{w})||_2^2 \tag{33}$$

as wanted. □

Since I've shown that $E(\mathbf{w}) = ||\hat{y} - \hat{B}\mathbf{w}||_2^2$ is equivalent to $E(\mathbf{w}) = ||\mathbf{y} - B\mathbf{w}||_2^2 + \lambda||\mathbf{w}||_2^2$, the solution for regularized regression can be obtained using either function

Question 3.

- a)
  let $\mathbf{w}^T\mathbf{b}(x_i) = f(x_i)$

$$E(w) = P(y_{1:N}|x_{1:N}, \mathbf{w}) \quad \text{Definition of Maximum Likelihood (ML) objective} \tag{34}$$

$$= \Pi_{i=1}^{N} P(y_i|x_i, \mathbf{w}) \tag{35}$$

$$= \Pi_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} exp\{\frac{(-y_i - \mathbf{w}^T\mathbf{b}(\mathbf{x_i}))^2}{2\sigma^2}\} \quad \text{Since } y \sim N(f(x), \sigma^2) \text{ and } \mathbf{w}^T\mathbf{b}(x_i)) = f(x_i) \tag{36}$$

$$= (\frac{1}{\sqrt{2\pi}\sigma})^N exp\{\Sigma_{i=1}^{N} \frac{(y_i - \mathbf{w}^T\mathbf{b}(\mathbf{x_i}))^2}{-2\sigma^2}\} \tag{37}$$

$$= (\frac{1}{2\pi\sigma^2})^{N/2} exp\{\frac{-1}{2\sigma^2}\Sigma_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{b}(\mathbf{x_i}))^2\} \tag{38}$$

- b)

$$E(w) = (\frac{1}{2\pi\sigma^2})^{N/2} exp\{\frac{-1}{2\sigma^2}\Sigma_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{b}(\mathbf{x_i}))^2\} \tag{39}$$

$$-ln(L(w)) = -ln((\frac{1}{2\pi\sigma^2})^{N/2} exp\{\frac{-1}{2\sigma^2}\Sigma_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{b}(\mathbf{x_i}))^2\}) \tag{40}$$

$$= -(\frac{N}{2}(ln(1) - ln(2\pi\sigma^2)) + \frac{-1}{2\sigma^2}\Sigma_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{b}(\mathbf{x_i}))^2 ln(e)) \tag{41}$$

$$= \frac{N}{2}ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\Sigma_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{b}(\mathbf{x_i}))^2 \tag{42}$$

$$= \frac{1}{2\sigma^2}\Sigma_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{b}(\mathbf{x_i}))^2 \quad \text{Removing unnecessary constants} \tag{43}$$

$$= \Sigma_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{b}(\mathbf{x_i}))^2 \quad \text{Multiplying by the constant } \frac{1}{2\sigma^2} \tag{44}$$

$$= \Sigma_{i=1}^{N}(y_i - f(\mathbf{x_i}))^2 \quad \text{as } \mathbf{w}^T\mathbf{b}(x_i)) = f(x_i) \tag{45}$$

$$\tag{46}$$

Since we took the negative log likelihood, we look for the *argmin* when we optimize for $\mathbf{w}$. Therefor this formulation is exactly the same as the LS Objective.

- c) let $1/K = P(y_{1:N}|x_{1:N})$

$$E(w) = P(\mathbf{w}|x_{1:N}, y_{1:N}) \quad \text{Definition of MAP} \tag{47}$$

$$= \frac{P(y_{1:N}|x_{1:N}, \mathbf{w})p(\mathbf{w})}{P(y_{1:N}|x_{1:N})} \quad \text{Bayes Rule} \tag{48}$$

$$= K \cdot P(y_{1:N}|x_{1:N}, \mathbf{w})p(\mathbf{w}) \quad \text{Since } 1/K = P(y_{1:N}|x_{1:N}) \tag{49}$$

$$= K \cdot P(y_{1:N}|x_{1:N}, \mathbf{w})(\Pi_{i=1}^{K} \frac{1}{\sqrt{2\pi\alpha^{-1}}} exp\{\frac{-1}{2\alpha^{-1}} w_i^2\}) \quad \text{Since } \mathbf{w} \sim N(0, \alpha^{-1}\mathbf{I}) \tag{50}$$

$$= K \cdot P(y_{1:N}|x_{1:N}, \mathbf{w})((\frac{1}{2\pi\alpha^{-1}})^{N/2} exp\{\frac{-1}{2\alpha^{-1}} \Sigma_{i=1}^{K} w_i^2\}) \tag{51}$$

$$-ln(E(w)) = -ln(K \cdot P(y_{1:N}|x_{1:N}, \mathbf{w})((\frac{1}{2\pi\alpha^{-1}})^{N/2} exp\{\frac{-1}{2\alpha^{-1}} \Sigma_{i=1}^{K} w_i^2\})) \quad \text{negative ln both sides} \tag{52}$$

$$= -lnK - lnP(y_{1:N}|x_{1:N}, \mathbf{w}) - ln(\frac{1}{2\pi\alpha^{-1}})^{N/2} - ln(exp\{\frac{-1}{2\alpha^{-1}} \Sigma_{i=1}^{K} w_i^2\}) \tag{53}$$

$$= -lnP(y_{1:N}|x_{1:N}, \mathbf{w}) - ln(exp\{\frac{-1}{2\alpha^{-1}} \Sigma_{i=1}^{K} w_i^2\}) \quad \text{Remove unnecessary constants} \tag{54}$$

$$= -lnP(y_{1:N}|x_{1:N}, \mathbf{w}) + \frac{1}{2\alpha^{-1}} \Sigma_{i=1}^{K} w_i^2 \tag{55}$$

$$= \frac{1}{2\sigma^2} \Sigma_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{b}(\mathbf{x_i}))^2 + \frac{1}{2\alpha^{-1}} \Sigma_{i=1}^{K} w_i^2 \quad -lnP(y_{1:N}|x_{1:N}, \mathbf{w}) \text{ from part b} \tag{56}$$

$$= \Sigma_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{b}(\mathbf{x_i}))^2 + \frac{\sigma^2}{\alpha^{-1}} \Sigma_{i=1}^{K} w_i^2 \quad \text{multiply by constant } 2\sigma^2 \tag{57}$$

$$= \Sigma_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{b}(\mathbf{x_i}))^2 + (\sigma^2\alpha)||\mathbf{w^T w}||_2^2 \tag{58}$$

- d) Since we took the negative log MAP, we look for the *argmin* when we optimize for $\mathbf{w}$.
  Therefor this formulation is exactly the same as the Regularized LS Objective (ridge regression) where $\lambda = \sigma^2\alpha$.

- e)Since for a uniform distribution, $P(\mathbf{w})$ is constant for all $\mathbf{w} \in [l, u]$
  Therefor, P(w) has no effect on the minimum or maximum of the objective function. so the MAP and ML objectives would both be equivalent to the LS objective.