CSCC11 - Assignment 1 Written Component

Masheil Mir
Mirmashe
1005245600
Feb 4, 2021

1a)

Consider LS basis function for single variable input/output

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

Let $\{(x_i, y_i)\}_{i=1}^{N}$ be the training data

$$y = f(x) = w_0 + w_1 b_1(x) + w_2 b_2(x) + \ldots + w_K b_K(x)$$

Let $\vec{x} = [x_1, \ldots, x_N]^T$ be the N inputs
Let $\vec{y} = [y_1, \ldots, y_N]^T$ be the N outputs
Let $\vec{w} = [w_0, w_1, \ldots, w_K]^T$ be the weights
↳ w_0 is our bias term

$$B = \begin{bmatrix} 1 & b_1(x_1) & b_2(x_1) & \cdots & b_K(x_1) \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & b_1(x_N) & b_2(x_N) & \cdots & b_K(x_N) \end{bmatrix}_{N \times (K+1)}$$

With least square objective f°:

$$E(\vec{w}) = \sum_{i=1}^{N} (y_i - f(x_i))^2$$
$$= \sum_{i=1}^{N} \left[y_i - \left(w_0 + \sum_{k=1}^{K} w_k b_k(x_i) \right) \right]^2$$

1b)

Matrix form:

$$E(\vec{w}) = \|\vec{y} - B\vec{w}\|_2^2$$
$$= (\vec{y} - B\vec{w})^T (\vec{y} - B\vec{w})$$
$$= (\vec{y}^T - \vec{w}^T B^T)(\vec{y} - B\vec{w}) \quad \text{(1b) \& (1c) from matrix identities}$$
$$= \vec{y}^T\vec{y} - \vec{w}^T B^T \vec{y} - \vec{y}^T B\vec{w} + \vec{w}^T B^T B\vec{w}$$
$$= \vec{y}^T\vec{y} - 2\vec{w}^T B^T \vec{y} + \vec{w}^T B^T B\vec{w} \qquad = 0 - 2B^T\vec{y}$$

$$\nabla_w E(\vec{w}) = \frac{\partial E}{\partial \vec{w}} \left(\vec{y}^T\vec{y} - 2\vec{w}^T B^T\vec{y} + \vec{w}^T B^T B\vec{w} \right)$$
$$= \underbrace{\frac{\partial}{\partial \vec{w}} (\vec{y}^T\vec{y})}_{=0} - 2\frac{\partial}{\partial \vec{w}} (\vec{w}^T B^T \vec{y}) + \frac{\partial}{\partial \vec{w}} (\vec{w}^T B^T B\vec{w})$$

Consider $\frac{\partial}{\partial \vec{w}} (\vec{w}^T B^T \vec{y})$

Let $C = B^T y$

$\frac{\partial}{\partial \vec{w}} (\vec{w}^T C)$ (6e)

$= C = B^T\vec{y}$

Consider $\frac{\partial}{\partial \vec{w}} (\vec{w}^T B^T B\vec{w})$

Let $C = B^T B$

$\frac{\partial}{\partial \vec{w}} (\vec{w}^T C\vec{w})$

$= (C + C^T)\vec{w}$ (5b)

$= (B^T B + (B^T B)^T)\vec{w}$

$= (B^T B + B^T B)\vec{w}$

$= 2B^T B\vec{w}$

1b) Cont'd

$$\nabla_w E(\vec{w}) = \frac{\partial}{\partial \vec{w}}(\vec{y}^T\vec{y}) - 2\frac{\partial}{\partial \vec{w}}(\vec{w}^T B^T \vec{y}) + \frac{\partial}{\partial \vec{w}}(\vec{w}^T B^T B\vec{w})$$
$$= -2B^T\vec{y} + 2B^TB\vec{w}$$
$$= 2(B^TB\vec{w} - B^T\vec{y})$$

1c) Optimal weight vector, denote this as w^*

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

$\Rightarrow 2(B^TB\vec{w} - B^T\vec{y}) = 0$

$\Rightarrow B^TB\vec{w} - B^T\vec{y} = 0$

$\Rightarrow B^TB\vec{w} = B^T\vec{y}$

$\Rightarrow \vec{w} = (B^T\vec{y})(B^TB)^{-1}$

$\Rightarrow \vec{w}^* = (B^T\vec{y})(B^TB)^{-1}$

2a) In general, the following conditions will make the equation from (1c) not unique:

1) If the matrix B^TB is singular for the following equation:

$$\nabla_w E(w) = 0 \Rightarrow B^TB\vec{w} = B^T\vec{y}$$

i.e. Showing that the determinant is 0

2) When training data $N < K$; not sufficient # of training data

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

Let $K=3$ and let our training data $\{(x_i, y_i)\} = \{(0,1),(1,0)\}$, $N=2$; Note $N<K$

Let $b_k(x) = x^k$

$f(x) = w_0 + w_1x + w_2x^2 + w_3x^3$

$$\vec{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}; \quad \vec{y} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}; \quad B = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \end{bmatrix}; \quad \vec{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix}$$

$$B\vec{w} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \end{bmatrix}\begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} w_0 + w_1x_1 + w_2x_1^2 + w_3x_1^3 \\ w_0 + w_1x_2 + w_2x_2^2 + w_3x_2^3 \end{bmatrix} = \begin{bmatrix} w_0 \\ w_0 + w_1 + w_2 + w_3 \end{bmatrix}$$

→ by plugging in $x_1 = 0$, $x_2 = 1$ from training data

$$B^TB\vec{w} = \begin{bmatrix}1 & 1\\ 0 & 1\\ 0 & 1\\ 0 & 1\end{bmatrix}\begin{bmatrix}w_0\\ w_0+w_1+w_2+w_3\end{bmatrix} = \begin{bmatrix}2w_0+w_1+w_2+w_3\\ w_0+w_1+w_2+w_3\\ w_0+w_1+w_2+w_3\\ w_0+w_1+w_2+w_3\end{bmatrix} = \begin{bmatrix}2 & 1 & 1 & 1\\ 1 & 1 & 1 & 1\\ 1 & 1 & 1 & 1\\ 1 & 1 & 1 & 1\end{bmatrix}$$

$$B^T\vec{y} = \begin{bmatrix}1 & 1\\ 0 & 1\\ 0 & 1\\ 0 & 1\end{bmatrix}\begin{bmatrix}1\\ 0\end{bmatrix} = \begin{bmatrix}1\\ 0\\ 0\\ 0\end{bmatrix} \quad ; \quad B^TB\vec{w} = B^T\vec{y} \Rightarrow \left[\begin{array}{cccc|c}2 & 1 & 1 & 1 & 1\\ 1 & 1 & 1 & 1 & 0\\ 1 & 1 & 1 & 1 & 0\\ 1 & 1 & 1 & 1 & 0\end{array}\right]$$

From the above it is shown that not having enough training data will make the normal eq= have multiple solutions since our system ends up being inconsistent (weights will have linear dependency).

2b) $E(\vec{w}) = \underbrace{\|\vec{y} - B\vec{w}\|^2}_{\text{data term}} + \underbrace{\lambda\|\vec{w}\|^2}_{\text{Smoothness term}}$

$$= (\vec{y}-B\vec{w})^T(\vec{y}-B\vec{w}) + \lambda\vec{w}^T\vec{w}$$
$$= (\vec{y}-B\vec{w})^T(\vec{y}-B\vec{w}) + \lambda\vec{w}^T\vec{w}$$
$$= (\vec{y}^T-\vec{w}^TB^T)(\vec{y}-B\vec{w}) + \lambda\vec{w}^T\vec{w}$$
$$= \vec{y}^T\vec{y} - \vec{w}^TB^T\vec{y} - \vec{y}^TB\vec{w} + \vec{w}^TB^TB\vec{w} + \lambda\vec{w}^T\vec{w}$$
$$= \vec{y}^T\vec{y} - 2\vec{w}^TB^T\vec{y} + \vec{w}^TB^TB\vec{w} + \lambda\vec{w}^T\vec{w}$$

$$\nabla_w E(\vec{w}) = \frac{\partial E}{\partial \vec{w}}(\vec{y}^T\vec{y} - 2\vec{w}^TB^T\vec{y} + \vec{w}^TB^TB\vec{w} + \lambda\vec{w}^T\vec{w})$$
$$= \frac{\partial}{\partial\vec{w}}(\vec{y}^T\vec{y}) - 2\frac{\partial}{\partial\vec{w}}(\vec{w}^TB^T\vec{y}) + \frac{\partial}{\partial\vec{w}}(\vec{w}^TB^TB\vec{w}) + \lambda\frac{\partial}{\partial\vec{w}}(\vec{w}^T\vec{w})$$
$$= 0 - 2B^T\vec{y} + 2B^TB\vec{w} + 2\lambda\vec{w}$$
$$= 2(B^TB+\lambda I)\vec{w} - 2B^T\vec{y}$$

Set $\nabla_w E(\vec{w}) = 0 \Rightarrow 2(B^TB+\lambda I)\vec{w} - 2B^T\vec{y} = 0$

$$\Rightarrow \quad 2(B^TB+\lambda I)\vec{w} = 2B^T\vec{y}$$
$$\Rightarrow \quad (B^TB+\lambda I)\vec{w} = B^T\vec{y}$$
$$\Rightarrow \quad \vec{w}^* = (B^TB+\lambda I)^{-1}(B^T\vec{y})$$

Explanation:
This regularization helps ensure that $\vec{w}^*$ has a unique value. The λI term handles the issue of a singular matrix. Note that $\lambda > 0$ is given. The problem arises when we take B^TB being nearly singular. i.e. B^TB is positive semidefinite, it has eigenvals ≥ 0. $\therefore$ $B^TB+\lambda I$ ensures all eigenvalues strictly > 0, implying B^TB is invertible.
Thus this also means $\vec{w}^*$ has only one unique solution.

2c)

$$\hat{B} = \begin{pmatrix} \vec{B} \\ \sqrt{\lambda} I_{K+1} \end{pmatrix} \in \mathbb{R}^{(N+K+1)\times(K+1)} \quad ; \quad \hat{y} = \begin{pmatrix} \vec{y} \\ \vec{0}_{K+1} \end{pmatrix} \in \mathbb{R}^{(N+K+1)}$$

$$\hat{B} = \begin{bmatrix} 1 & b_1(x_1) & b_2(x_1) & \cdots & b_K(x_1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & b_1(x_N) & b_2(x_N) & \cdots & b_K(x_N) \\ \sqrt{\lambda} & & & & \\ & \sqrt{\lambda} & & 0 & \\ & & \ddots & & \\ & 0 & & & \sqrt{\lambda} \end{bmatrix} \quad ; \quad \hat{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad ; \quad \vec{w} = \begin{bmatrix} w_0 \\ \vdots \\ w_K \end{bmatrix}$$

WTS: $E(\vec{w}) = \|\hat{y} - \hat{B}\vec{w}\|_2^2$ is equivalent to $E(\vec{w}) = \|\vec{y} - B\vec{w}\|_2^2 + \lambda\|\vec{w}\|_2^2$

Consider $E(w) = \|\hat{y} - \hat{B}\vec{w}\|_2^2$

$$= \left\| \begin{bmatrix} y_1 \\ \vdots \\ y_N \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} 1 & b_1(x_1) & b_2(x_1) & \cdots & b_K(x_1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & b_1(x_N) & b_2(x_N) & \cdots & b_K(x_N) \\ \sqrt{\lambda} & & & & \\ & \sqrt{\lambda} & & 0 & \\ & & \ddots & & \\ & 0 & & & \sqrt{\lambda} \end{bmatrix} \begin{bmatrix} w_0 \\ \vdots \\ w_K \end{bmatrix} \right\|_2^2$$

$$= \sum_{i=1}^{N} y_i - w_0 - \left(\sum_{k=1}^{K} w_k b_k(x_i)\right)^2 + \sum_{k=0}^{K} (0 - \sqrt{\lambda} w_k)^2$$

$$= \|\vec{y} - B\vec{w}\|_2^2 + \sum_{k=0}^{K} (0 - \sqrt{\lambda} w_k)^2 \quad \text{by def° of LS objective}$$

$$= \|\vec{y} - B\vec{w}\|_2^2 + \lambda \sum_{k=0}^{K} (w_k)^2$$

$$= \|\vec{y} - B\vec{w}\|_2^2 + \lambda \|w_k\|_2^2 \text{ , as wanted}$$

$\therefore$ $E(\vec{w}) = \|\hat{y} - \hat{B}\vec{w}\|_2^2$ is equivalent to $E(\vec{w}) = \|\vec{y} - B\vec{w}\|_2^2 + \lambda\|\vec{w}\|_2^2$, as shown above, we can obtain the solution for regularized regression can be obtained by either method.

The constrain for $\vec{w}$ can be found using $\nabla_w E(w) = 0$ which was computed in part (2b) where we got the result of $w^* = (B^T B + \lambda I)^{-1}(B^T \vec{y})$.

3a) Let $y = f(x) + \varepsilon$, $y \sim N(f(x), \sigma^2)$

Let $\{(x_i, y_i)\}_{i=1}^N$ be the training data

Let $f(x) = w_0 + w_1 b_1(x) + \ldots + w_K b_K(x)$

Then, since $y \sim N(f(x), \sigma^2)$, we have

$$P(y \mid \vec{w}^T \vec{b}) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{1}{2\sigma^2}(y - \vec{w}^T \vec{b})^2}$$

st $\vec{b} = [1 \quad b_1(x) \quad \ldots \quad b_K(x)]$

Note: $N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

To compute ML objective:

$P(y_{1:N} \mid \vec{w}^T \vec{b}_1, \ldots, \vec{w}^T \vec{b}_N)$ where $\vec{b}_i = [1, b_1(x_i), \ldots, b_K(x_i)]$

$= \prod_{i=1}^N P(y_i \mid \vec{w}^T \vec{b}_i)$ because each y_i is independent

$= \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \vec{w}^T \vec{b}_i)^2}$ Since $y \sim N(f(x), \sigma^2)$

$= \left[\frac{1}{\sqrt{2\pi}\sigma}\right]^N e^{-\frac{1}{2\sigma^2}\sum_{i=1}^N (y_i - \vec{w}^T \vec{b}_i)^2}$

$= \left[\frac{1}{2\pi\sigma^2}\right]^{N/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^N (y_i - \vec{w}^T \vec{b}_i)^2}$

3b) The following defines an energy function to be minimized:

$-\log(P(y_{1:N} \mid \vec{w}^T \vec{b}_1, \ldots, \vec{w}^T \vec{b}_N))$

$= -\log\left(\left[\frac{1}{2\pi\sigma^2}\right]^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^N (y_i - \vec{w}^T \vec{b}_i)^2}\right)$

$= -\log\left(\frac{1}{2\pi\sigma^2}\right)^{N/2} - \log e^{-\frac{1}{2\sigma^2}\sum_{i=1}^N (y_i - \vec{w}^T \vec{b}_i)^2}$

$= -\left(\frac{N}{2}[\underbrace{\log(1)}_{=0} - \log(2\pi\sigma^2)]\right) + \frac{1}{2\sigma^2}\sum_{i=1}^N (y_i - \vec{w}^T \vec{b}_i)^2$

$= \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^N (y_i - \vec{w}^T \vec{b}_i)^2$

$= C + \frac{1}{2\sigma^2}\sum_{i=1}^N (y_i - \vec{w}^T \vec{b}_i)^2$ We can ignore C since it does not affect the derivative & multiply out $\frac{1}{2\sigma^2}$

$E(w) = \sum_{i=1}^N (y_i - \vec{w}^T \vec{b}_i)^2$

$= (\vec{y} - B\vec{w})^T (\vec{y} - B\vec{w})$

$= \|\vec{y} - B\vec{w}\|_2^2$, $B = [\vec{b}_1, \ldots, \vec{b}_N]^T$

Recall LS objective from Q1:

$$E(\vec{w}) = \vec{y}^T\vec{y} - 2\vec{w}^T B^T\vec{y} + \vec{w}^T B^T B\vec{w} \longrightarrow ①$$

Since we took the negative log likelihood, we find argmin to optimize $\vec{w}^*$

Neg. log likelihood: $E(\vec{w}) = \|y - B\vec{w}\|_2^2$

$$= (\vec{y} - B\vec{w})^T(y - B\vec{w})$$
$$= (\vec{y}^T - \vec{w}^T B^T)(y - B\vec{w})$$
$$= \vec{y}^T\vec{y} - 2\vec{w}^T B^T\vec{y} + w^T B^T B\vec{w} \longrightarrow ②$$

Since we can see the equation ① = ②, we can conclude that the negative log likelihood and the LS objective are the same.

3c) $\vec{w} \sim N(0, \alpha^{-1}I)$, Gaussian dist.

$y \sim N(f(x), \alpha^{-1}I)$; Let $y = \vec{w}^T\vec{b} + \varepsilon$

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

Need $P(\vec{w} \mid \vec{b}_{1:N}, y_{1:N}) = \dfrac{P(y_{1:N} \mid b_{1:N}, \vec{w})\, P(\vec{w})}{P(y_{1:N} \mid b_{1:N})}$ ← Can ignore / disregard this term bc doesn't depend on $\vec{w}$.

$$= P(y_{1:N} \mid b_{1:N}, \vec{w})\, P(\vec{w})$$

① $P(\vec{w}) = P(w_1, \ldots, w_K)$

$$= \prod_{i=1}^N P(w_i)$$
$$= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\alpha^{-1}}} e^{\left(-\frac{\alpha}{2}\vec{w}_i^2\right)}$$
$$= \left(\frac{1}{\sqrt{2\pi\alpha^{-1}}}\right)^{N/2} e^{\left(-\frac{\alpha}{2}\vec{w}^T\vec{w}\right)}$$

② $\vec{y} = \vec{w}^T\vec{b} + \varepsilon$

$$P(y_1, \ldots, y_N \mid \vec{w}^T\vec{b}_1, \ldots, \vec{w}^T\vec{b}_N) = \left[\frac{1}{2\pi\sigma^2}\right]^{N/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^N (y_i - \vec{w}^T\vec{b}_i)^2}$$

(Computed in part (3a))

③ $P(\vec{w} \mid \vec{b}_{1:N}, y_{1:N}) = P(y_{1:N} \mid b_{1:N}, \vec{w})\, P(\vec{w})$

$$= \left(\left[\frac{1}{2\pi\sigma^2}\right]^{N/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^N (y_i - \vec{w}^T\vec{b}_i)^2}\right)\left(\left(\frac{1}{\sqrt{2\pi\alpha^{-1}}}\right)^{N/2} e^{\left(-\frac{\alpha}{2}\vec{w}^T\vec{w}\right)}\right)$$

④ Negative log of MAP

$$-\log(P(\vec{w}|\vec{b}_{1:N}, y_{1:N})) = -\log\left[\left(\left[\frac{1}{2\pi\sigma^2}\right]^{N/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \vec{w}^T\vec{b}_i)^2}\right)\left(\left(\frac{1}{|2\pi\alpha^{-1}|}\right)^{M/2} e^{\left(-\frac{\alpha}{2}\vec{w}^T\vec{w}\right)}\right)\right]$$

$$= \frac{N}{2}(-\log(1) - (-\log(2\pi\sigma^2)) + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \vec{w}^T\vec{b}_i)^2 + \frac{\alpha}{2}\vec{w}^T\vec{w}$$

$$= \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \vec{w}^T\vec{b}_i)^2 + \frac{\alpha}{2}\vec{w}^T\vec{w}$$

$$= C + \frac{1}{2\sigma^2}\|y - B\vec{w}\|_2^2 + \frac{\alpha}{2}\|\vec{w}\|_2^2$$

We can ignore C, multiply by $2\sigma^2$

$$\Rightarrow E(\vec{w}) = \|y - B\vec{w}\|_2^2 + \sigma^2\alpha\|\vec{w}\|_2^2$$

3d) Looking at the negative log posterior compared to the LS objective above, we can conclude that they are the same, except in the negative log posterior we have the addition of the $\sigma^2\alpha\|\vec{w}\|$ term which is the regularization term with $\lambda = \sigma^2\alpha$. This term aids with the issue of overfitting, and singularity of B^TB.
That being said, we can say that the negative log posterior is almost identical to the regularized LS objective, but it has $\lambda = \sigma^2\alpha$ in this case.

3e) If we assume $\theta \sim U[0,1]$, then $P(\theta)$ is constant $\forall\ \theta \in [0,1]$
$\therefore$ $P(w)$ will not affect the min or max of the objective function and we can conclude that the MAP and ML objectives will be equivalent to the original LS objective.