CSCC11 Assignment 1.
Ti, Zhang
1004424517

1. Consider Least-Squares (LS) basis function regression for a single variable input/output problem, i.e.,

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

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

(a) Please formulate the LS objective in terms of a vector of known inputs, a corresponding vector of known outputs, and the appropriate matrices. Remember to include a bias term in the model (incorporated into the basis function matrix and the weight vector). **Note**: Please put the bias related terms as the first element/row/column of the vector/matrix.

(b) Then show each step of taking the gradient of the objective function. **Note**: You may use the matrix identities handout on the course website.

(c) Then solve for the optimal weight vector $\mathbf{w}$ (which includes the bias w_0).

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

Vector of known input: $\bar{x} = [x_1 \cdots x_N]^T$

Vector of known output: $\bar{y} = [y_1 \cdots y_N]^T$

Matrix form: $\bar{w} = [w_0 . w_1 \cdots w_K]^T$ Vector of k regression coefficient.

$B = [b_0(x_i) \; b_1(x_i) \cdots b_K(x_i)]$ vector of bias function value with x given.

$$= \begin{bmatrix} 1 & b_1(x_1) & b_2(x_1) & \cdots & b_K(x_1) \\ \vdots & \vdots & & & \vdots \\ 1 & b_1(x_N) & \cdots & \cdots & b_K(x_N) \end{bmatrix}$$

* $b_0(x_i) = 1$

Error function: $E(\bar{w}) = \| \bar{y} - B\bar{w} \|^2$

b) $E(\bar{w}) = \| \bar{y} - B\bar{w} \|^2 = (\bar{y} - B\bar{w})^T(\bar{y} - B\bar{w})$

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

$= \bar{y}^T y - \bar{w}^T B^T \bar{y} - \bar{y}^T B\bar{w} + \bar{w}^T B^T B\bar{w}$

$= \bar{y}^T y - \underline{(\bar{y}^T B\bar{w})^T} - \underline{\bar{y}^T B\bar{w}} + \bar{w}^T B^T B\bar{w}$ * underlined items are constant.

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

$\partial E(\bar{w}) / \partial \bar{w} = 0 - 2\bar{y}^T B + [B^T B + (B^T B)^T] \bar{w}^T = -2\bar{y}^T B + 2 B^T B \bar{w}^T$ $\blacksquare$

* $\partial x^T A x / \partial x = (A + A^T) x^T$

c). Let $\partial E(\bar{w}) / \partial \bar{w} = 0$.

$\Rightarrow \quad -2\bar{y}^T B + 2B^T B \bar{w}^T = 0.$

$B^T B \bar{w}^T = \bar{y}^T B$

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

$\bar{w} B^T B = B^T \bar{y}$

$\bar{w} = (B^T B)^{-1} B^T \bar{y}$ $\blacksquare$

Addition: $\partial^2 E(\bar{w}) / \partial \bar{w}^2 = 2B^T B > 0$. $\therefore$ Above $\bar{w}$ is the optimal value.

CSCC11 Assignment 1.

Ti, Zhang

1004424517

2. In the following problem we consider cases in which the model above is not well constrained by the data alone.

(a) Assume that you have at least one data point, $N \geq 1$. In general, under what conditions will the solution to the normal equations in Q1(c) *not* be unique? Set up and provide a simple regression problem where the optimal weight vector ($\mathbf{w}$) is not unique.

(b) Consider L2 regularized regression (a.k.a. ridge regression), which adds a term to the LS objective that penalizes the squared L2 norm of $\mathbf{w}$, scaled by a positive regularization parameter λ (see Eqn (10) in Chapter 3 of the online notes). Use the gradient of this regularized objective to derive the normal equations for the solution in this case, and explain the way in which this regularization helps overcome the problem above when the data alone do not fully constrain the solution.

(c) Show that the solution for regularized regression in part (b) can alternatively be obtained via (ordinary) least squares regression with augmented basis function matrix $\hat{\mathbf{B}}$ and known outputs $\hat{\mathbf{y}}$ as defined below:

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

where $\mathbf{B}$ is the basis function matrix and $\mathbf{y}$ is the vector of known outputs in Q1. $\mathbf{I}_{K+1}$ represents the identity matrix in $\mathbb{R}^{(K+1)\times(K+1)}$ and $\mathbf{0}_{K+1}$ is a zero vector in $\mathbb{R}^{(K+1)}$. Note that the quantities inside brackets are block matrices. Explain in short how the above formulation constrains $\mathbf{w}$.

a) Condition: $\text{rank}(B) < K+1$. Then $\bar{y} = B\bar{w}$ has infinity solutions.

Q1(c) has unique sol^n iff $\text{Null}(B) = \{0\}$. Then $(B^TB)^{-1}$ exist, and we can have unique sol^n $\bar{w} = (B^TB)^{-1}B^T\bar{y}$ hold.

Example for non-unique w: $y(x) = w_0 + w_1 x + w_2 x^2$ with $N > 2$.

b). $E(\bar{w}) = \|\bar{y} - B\bar{w}\|^2 + \lambda\|\bar{w}\|^2$

$= (\bar{y} - B\bar{w})^T(\bar{y} - B\bar{w}) + \lambda\bar{w}^T\bar{w}$

$= \bar{y}^Ty - 2(\bar{y}^TB\bar{w}) + \bar{w}^TB^TB\bar{w} + \bar{w}^T(\lambda I)\bar{w}$

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

$\partial E(\bar{w}) / \partial \bar{w} = 0 - 2\bar{y}^TB + [B^TB + \lambda I + (B^TB + \lambda I)^T]\bar{w}^T$

$= -2\bar{y}^TB + [B^TB + \lambda I + B^TB + \lambda I]\bar{w}^T$

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

Set $\partial E(\bar{w}) / \partial \bar{w} = 0$ $\therefore$ $-2\bar{y}^TB + 2(B^TB + \lambda I)\bar{w}^T = 0$

$(B^TB + \lambda I)\bar{w}^T = \bar{y}^TB$

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

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

This helps as when $\text{rank}(B) < K+1$, i.e. B is un-invertible. $\det(B) = 0$

i.e. $(B^TB)^{-1}$ does not exist.

i.e. $\bar{w} = (B^TB)^{-1}B^T\bar{y}$ cannot proceed.

But we're able to proceed with $\bar{w} = (B^TB + \lambda I)^{-1}B^T\bar{y}$

c) $\hat{B} = \begin{bmatrix} B \\ \sqrt{\lambda} I_{K+1} \end{bmatrix} = \begin{bmatrix} 1 & b_1(x_1) & \cdots & b_K(x_1) \\ \vdots & \vdots & & \vdots \\ 1 & b_1(x_N) & \cdots & b_K(x_N) \\ \sqrt{\lambda} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & \cdots & \cdots & \sqrt{\lambda} \end{bmatrix}$ $\quad \hat{y} = \begin{bmatrix} \bar{y} \\ 0_{K+1} \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

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

Here $\hat{B} = \begin{bmatrix} B \\ \sqrt{\lambda} I \end{bmatrix}$ $\quad \hat{B}^T = [B^T \ \sqrt{\lambda} I]$

$$\hat{B}^T\hat{B} = [B^T \ \sqrt{\lambda} I] \begin{bmatrix} B \\ \sqrt{\lambda} I \end{bmatrix} = B^T B + \lambda I$$

$$(\hat{B}^T\hat{B})^{-1} = (B^T B + \lambda I)^{-1}$$

$$\hat{B}^T \hat{y} = [B^T \ \sqrt{\lambda} I] \begin{bmatrix} \bar{y} \\ 0 \end{bmatrix} = B^T \bar{y}$$

$\therefore \bar{w} = (\hat{B}^T\hat{B})^{-1} \hat{B}^T \hat{y} = (B^T B + \lambda I)^{-1} B^T \bar{y}$ is the same as in part (B).

This way we can see that this is equivalent to the expression in L_2 regularization regression.

3. Now consider a probabilistic formulation of basis function regression. This is useful as a way to incorporate measurement noise. For example, images may contain white noise due to lighting variations, hardware issues, and other reasons. Here, we'll assume the target output y is equal to $f(x)$ plus Gaussian noise. Specifically, we assume y given x follows a Gaussian distribution with mean $f(x)$, and variance σ^2. We write this as $y \sim \mathcal{N}(f(x), \sigma^2)$.

As above, we assume a single variable input/output problem, with training data $\{(x_i, y_i)\}_{i=1}^N$, and that $f(x)$ is a weighted sum of basis functions evaluated at x, with weights $\mathbf{w} = [w_0, \ldots, w_K]^T$.

(a) Formulate the Maximum Likelihood (ML) objective (without solving for the weights).

(b) What can you say about the negative log likelihood as compared to the LS objective above in Q1?

(c) Now, suppose that the model parameters (weights) follow a Gaussian distribution with zero mean with some fixed isotropic covariance $\alpha^{-1}\mathbf{I}$, i.e., $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$. Formulate and take the negative log of the Maximum a Posteriori (MAP) objective. **Note**: You may ignore the evidence term as it does not depend on the parameters of interest.

(d) What can you say about minimizing the negative log posterior compared to the LS objective above?

(e) What happens if we assume that the model parameters follow a Uniform distribution? What can you say about ML and MAP estimates in that case?

CSCC11 Assignment 1.

Ti, Zhang

1004424517

a) $y = f(x) = w_0 + w_1 b_1(x) + \cdots + w_K b_K(x) + \varepsilon. \Rightarrow$

$\Downarrow$

$y \sim N(f(x), \sigma^2)$

$$L(y_i|w) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left\{\frac{-[y_i - f(x)]^2}{2\sigma^2}\right\}$$

$$L(w) = \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left\{\frac{-[y_i - f(x_i)]^2}{2\sigma^2}\right\}\right) \text{ for } \{(x_i, y_i)\}_{i=1}^{N} \quad \blacksquare$$

CSCC11 Assignment 1.
Ti. Zhang
1004424517

b) $$\log[L(w)] = \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left\{\frac{-[y_i - f(x_i)]^2}{2\sigma^2}\right\}\right)$$
$$= \sum_{i=1}^{N} \left\{-\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{[y_i - f(x_i)]^2}{2\sigma^2}\right\}$$
$$= -N \cdot \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{\sum_i^N [y_i - f(x_i)^2]}{2\sigma^2}$$
$$-\log[L(w)] = N \cdot \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) + \frac{\sum_i^N [y_i - f(x_i)^2]}{2\sigma^2}$$
$$= N \cdot \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) + \frac{1}{2\sigma^2}\|y - \bar{B}w\|^2 \quad \blacksquare$$

The calculation is the same as in Q1, as calculating MLE is equivalent to finding minimum least square solution.

c) Given $w \sim N(0, \alpha^{-1}I)$, $w_1, w_2 \cdots w_k$ are iid.
$$f(w) = \prod_{i=1}^{k}\left[\frac{1}{\sqrt{2\pi\alpha^{-1}}} \cdot \exp\left\{\frac{-(w_i - 0)^2}{2\alpha^{-1}}\right\}\right]$$
MAP objective $P(w|Y) = \frac{P(Y|w)P(w)}{\underbrace{P(Y)}_{\text{constant.}}} \propto P(Y|w)P(w)$
$$P(Y|w) = \prod_{i=1}^{N}\left(\frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left\{\frac{-[y_i - f(x_i)]^2}{2\sigma^2}\right\}\right)$$
$$P(w) = \prod_{i=1}^{k}\left[\frac{1}{\sqrt{2\pi\alpha^{-1}}} \cdot \exp\left\{\frac{-(w_i - 0)^2}{2\alpha^{-1}}\right\}\right]$$
$$\therefore P(w|Y) \propto P(Y|w)P(w) = \prod_{i=1}^{N}\left(\frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left\{\frac{-[y_i - f(x_i)]^2}{2\sigma^2}\right\}\right)\prod_{i=1}^{k}\left[\frac{1}{\sqrt{2\pi\alpha^{-1}}} \cdot \exp\left\{\frac{-(w_i - 0)^2}{2\alpha^{-1}}\right\}\right] = g(w)$$
$\therefore$ According to (b), $-\log(g(w)) = N\log(\sqrt{2\pi}\sigma) + \frac{1}{2\sigma^2}\|\bar{y} - B\bar{w}\|^2 + N\log(\sqrt{2\pi\alpha^{-1}}) + \frac{\alpha}{2}\|w\|^2$

d) In (c) I'm calculating the minimum of MAP $[-\log(g(w))]$
This is equivalent to finding the minimum of $\left(\frac{1}{2\sigma^2}\|\bar{y} - B\bar{w}\|^2 + \frac{\alpha}{2}\|w\|^2\right)$,
which is the regularization term, as $\left(N\log(\sqrt{2\pi}\sigma) + N\log(\sqrt{2\pi\alpha^{-1}})\right)$ is a constant.

e) For $w \sim$ Uniform,
MLE: No change as it's depending on y's distn.
MAP: $P(w) = \alpha^{-1}$
$$P(w|Y) \propto \prod_{i=1}^{N}\left(\frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left\{\frac{-[y_i - f(x_i)]^2}{2\sigma^2}\right\}\right) \cdot \alpha^{-1} = g(w)$$
$$-\log(g(w)) = N\log(\sqrt{2\pi}\sigma) + \frac{1}{2\sigma^2}\|\bar{y} - B\bar{w}\|^2 + N\log(\alpha^{-1})$$
Since $N\log(\sqrt{2\pi}\sigma) + N\log(\alpha^{-1})$ is constant, MAP is same as LS without the regularization terms.