SafwanElmadani / ece523_doc21 Public

<> Code Issues Pull requests Actions Projects Wiki Security

master

ece523_doc21 / ece523 / hw / hw2 / hw2ml.pdf

SafwanElmadani update History

1 contributor

4.78 MB

Hw2

Thursday, February 11, 2021 7:22 PM

① Linear classifier with a Margin :

a data set has 2 data points

$x_1 \in C_1 \; (y_1 = +1)$

$x_2 \in C_2 \; (y_2 = -1)$

Set up the minimization problem w/

constrains on $w^T x_1 + b$

$w^T x_2 + b$

To find the hyperplane, we need to solve

$$\underset{w \in \mathbb{R}^p}{\arg\min} \|w\|_2^2 = \underset{w \in \mathbb{R}^p}{\arg\min} \; w^T w$$

subject to :

$$w^T x_1 + b = 1$$

$$w^T x_2 + b = -1$$

Using Lagrange multiplier λ_1 and λ_2, we can write the following:

$$L = \underset{w \in \mathbb{R}^p}{\arg\min} \left\{ \|w\|_2^2 + \lambda_1 (w^T x_1 + b - 1) + \lambda_2 (w^T x_2 + b + 1) \right\}$$

Taking the derivative w.r.t. w and b and make them equal to 0.

$$\frac{\partial L}{\partial w} = 0 \Rightarrow 2w + \lambda_1 x_1 + \lambda_2 x_2 = 0$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \lambda_1 + \lambda_2 = 0$$

$$\lambda_1 = -\lambda_2$$

$$w = -\frac{1}{2}(\lambda_1 x_1 + \lambda_2 x_2) \Rightarrow \text{substituting } \lambda_1 = -\lambda_2$$

$$w = -\frac{1}{2}(-\lambda_2 x_1 + \lambda_2 x_2)$$

$$\boxed{w = \frac{\lambda_2}{2}(x_1 - x_2)} \qquad 2w = \lambda_2 (x_1 - x_2)$$

$$w^T x_1 + b = -w^T x_2 - b$$

$$2b = -w^T(x_1 + x_2)$$

$$\boxed{b = -\frac{w^T}{2}(x_1 + x_2)}$$

$$2 = 1 + 1 \Rightarrow 2 = (w^T x_1 + b) - (w^T x_2 + b)$$

$$2 = w^T x_1 - w^T x_2$$

$$2 = w^T (x_1 - x_2)$$ We now substitute for w.

$$2 = \frac{\lambda_2}{2}(x_1^T - x_2^T)(x_1 - x_2)$$

$$2 = \frac{\lambda_2}{2}(x_1^T x_1 + x_1^T x_2 - x_2^T x_1 + x_2^T x_2)$$

$X_{n \times d}$

$$\boxed{\lambda_2 = 4\,(x_1^T x_1 + x_1^T x_2 - x_2^T x_1 + x_2^T x_2)^{-1}}$$

$$\lambda_1 = -\lambda_2$$

② Linear Regression with Regularization:

The loss function $L(w) = \sum_{i=1}^{n} (y_i - w^T x_i)^2$

$$= (y - Xw)^T (y - Xw)$$

sum of squared errors from lecture notes

Adding the penalty:

$$L(w) = (y - Xw)^T (y - Xw) + \lambda w^T w$$

$$= (y^T - w^T X^T)(y - Xw) + \lambda w^T w$$

$$= y^T y - y^T X w - w^T X^T y + w^T X^T X w + \lambda w^T w$$

$$= y^T y - 2 y^T X w + w^T X^T X w + \lambda w^T w$$

$$\frac{\partial L}{\partial w} = 0$$

$$\frac{\partial L}{\partial w} = 2X^T X w - 2X^T y + 2\lambda w = 0$$

$$\frac{\partial L}{\partial w} = 2X^TXw - 2X^Ty + 2\lambda w = 0$$

$$X^Ty = X^TX w + \lambda w$$

$$X^Ty = (X^TX + \lambda I) w$$

$$w = (X^TX + \lambda I)^{-1} X^Ty$$ parameter of linear regression with penalty.

It penalizes w for taking large values. It makes W small to prevent the coeffcients from overfitting.

(4) Conceptual.

$P(w|x)$ $\qquad$ $\dfrac{P(x|w)\,P(w)}{P(X)}$

- Easier to model because it doesn't require modeling the joint distribution $P(w,x)$.
- Estimate the posterior directly.
- Can't detect outlier in the data.

- $P(x|w)$ become hard to model if the dimention x is large.
- uses the available data to estimate the prior $P(w)$, likelihood $P(x|w)$, and evidence $P(x)$.
- Known the evidence term $P(x)$ is useful because it normalizes the term and changes the posterior into probabilty $[0,1]$. The likelihood term can be bigger than 1.