Linear Regression

training set: $D = \{x_i, y_i\}_{i=1}^N$

Where $\alpha_i = (x_i, x_i, \cdots, x_D)^T$

Want to find: $f(x,w)$

$\Rightarrow y = f(x,w) = w^T x = \sum_{j=0}^{D} w_j x_j$ ← Bias

i.e. $y: (x_0, x_1, x_2 \cdots, x_D) \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_D \end{pmatrix}$

along Set $x_0 = 1$

$= \underbrace{w_0}_{\text{Bias}} + w_1 x_1 + w_2 x_2 + \cdots + w_D x_D$

MSE Mean Squared Error

$L(w) = \frac{1}{N} \sum_{i=1}^{n} (y_i - w^T x_i)^2$

We want to minimize $L(w) \Rightarrow \bar{w} = \arg\min_w L(w)$

$\Rightarrow$ Let its gradient to 0. i.e. $\nabla L(w) = 0$

Using Matrix to express Training Set.

$X = \begin{bmatrix} x_{1,0} & x_{1,1} & x_{1,2} & \cdots & x_{1,D} \\ & & & & \\ x_{N,0} & & & & x_{N,D} \end{bmatrix} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}$

$y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$ $\Rightarrow L(w) = \frac{1}{N} \|y - Xw\|^2$

$= \frac{1}{N} (y - Xw)^T (y - Xw)$

$(AB)^T = B^T A^T$

$= \frac{1}{N} (y^T - (Xw)^T)(y - Xw)$

标量的转置不变 $y^T Xw = (y^T Xw)^T = (Xw)^T y$

$= \frac{1}{N} (y^T y - 2(Xw)^T y + w^T X^T X w)$

$\nabla_w L(w) = \frac{\partial L}{\partial w} = \begin{bmatrix} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_n} \end{bmatrix}$

Least Square Regression

data set $\begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$ $\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ "label"

$\therefore \nabla_w L(w) = \frac{\partial}{\partial w} \sum_{i=1}^{n} (y_i - x_i^T w)^2$

$= 2 \sum_{i=1}^{n} (y_i - \alpha_i^T w)\, \alpha_i = 0$

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

$\alpha_i^T w = x_{i,1} w_1 + x_{i,2} w_2 + \cdots + x_{i,n} w_n$

$\frac{\partial (\alpha_i^T w)}{\partial w} = \begin{bmatrix} \frac{\partial x_i^T w}{\partial w_1} \\ \vdots \\ \frac{\partial x_i^T w}{\partial w_n} \end{bmatrix} = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,n} \end{bmatrix} = \alpha_i$

$\sum_{i=1}^{n} (y_i - \alpha_i^T w)\, x_i = 0$

$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i^T w\, x_i = 0$

$\sum_{i=1}^{n} \alpha_i y_i = \sum_{i=1}^{n} (\alpha_i^T x_i) w$

$X^T y = X^T X w$

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

Polynomial Rregression

Linear regression: $\bar{x} = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix}$

Polynomial regression: $\phi(x) = \begin{bmatrix} x_0 \\ x_1 \\ x_1^2 \\ x_1^3 \end{bmatrix}$ ← Capture more features

feature mapping

$\Rightarrow y = f(x) = w^T \phi(x)$

d: degree

Model Capacity: Overfitting and Underfitting

How d impact L^{train} and L^{test}

Hypothesis space

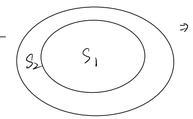
$S_1 = \{ y = w_0 + w_1 x_1 \mid w_0, w_1 \}$.

regression task: pick one solution from S

With d increasing:

$S_2 = \{ y = w_0 + w_1 x_1 + w_2 x_1^2 + w_3 x_1^3 \mid w_0, w_1, w_2, w_3 \}$

$\Rightarrow S_1 \subset S_2$ $\Rightarrow S_2$: higher Capacity



Underfitting: Model too simple!

cannot fit pattern well

$\Rightarrow$ high L^{train}, L^{test}

☆ Overfitting: Model too Complex!

Assume: Training Set: pattern (real) + noise 1

Tese Set: pattern + noise 2.

Model fit pattern + noise 1 well $\Rightarrow L^{train}$ low

But. Since noise 1 ≠ noise 2.

Model cannot fit pattern + noise 2 $\Rightarrow L^{test}$ high

Generalization Error = Test error

$L^{(test)}(w) = \frac{1}{N^{(test)}} \| y^{(test)} - w^T x^{(test)} \|^2$

Obj: 1. Make L^{train} ↓

2. Make Gap: L^{train} & L^{test} ↓

$\Rightarrow$ Choose a model with appropriate Capacity

Validation

Data { Training { Training Set: Determine weight; Validation Set: Determine d → "Hyper Parameter" { underfitting: d ↑; Overfitting: d ↓ } } Test

"Hyper Parameter"

↓

Validation

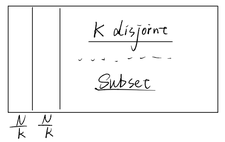
= "Pick the Hypermeter With the Lowest Validation Error"

Held-out Validation

Cons: 1. Not using all data.

2. High Variance (unreliable)

Cross Validation



1. Using one of these subsets as validation set
2. Combining others → Training Set
3. run above k times.

Average the performance in validation set.

i.e. $L^{validation} = \frac{1}{K} \sum_{i=1}^{K} L_{(i)}^{validation}$

Regularization

L_2 regularization

$L(w, w_0) = \underbrace{\frac{1}{N} \sum_{i=1}^{N} (y_i - (w_0 + w^T \phi(x_i)))^2}_{\text{MSE Loss}} + \underbrace{\lambda \|w\|_2^2}_{\text{regularization term}}$

Balance Parameter: How to Choose? Validation

$\lambda \to \infty$: loss pattern → "Underfitting"

appropriate λ ✓

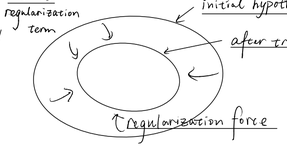
$\lambda \to 0$: "Overfitting"

Min L → { ① Minimizing Loss; ② reduce weight ☆ }

"Suppress the weights of unimportant feature"

↓

Weight Decay



Remark: w_0 is not regularized as it doesn't influence Capacity

ridge regression / penalized least squares Solution:

$\hat{w}_{ridge} = (\underbrace{\lambda I_K}_{\text{ridge expression}} + X^T X)^{-1} X^T y$

Compared with: $\hat{w} = (X^T X)^{-1} X^T y$

LASSO: least absolute shrinkage and selection operato

$L(w, w_0) = \frac{1}{N} \sum_{i=1}^{N} (y_i - (w_0 + w^T \phi(x_i)))^2 + \lambda \|w\|_1$

$\Rightarrow$ force certain coefficient to Zero

Gradient always be ± 1

"Push" the solution to $w_i = 0$

Performanee Metric

① MSE RMSE

② R^2 Score

$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - w^T x_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} = 1 - \frac{MSE}{Variance}$

$R^2 \in [-1, 1]$ $R^2 \to 1$ or $R^2 \to -1$ $\Rightarrow$ Better Performance

Probabilistic Formulation of Linear Rregression

More reasonable linear regression model:

$y = w^T x + \varepsilon$ $(\varepsilon \sim N(0, \sigma^2))$ ε: noise

$\theta = (\bar{w}, \sigma)$

$P(y \mid x, \theta) = N(y \mid \underbrace{\bar{w}^T \bar{x}}_{\text{mean}}, \underbrace{\sigma^2}_{\text{Var}})$

Point Estimation: $\hat{y} = w^T x$

Determine $\theta(w, \sigma)$ by ↘ cross entropy

$\Rightarrow \hat{\theta} = \arg\min_{\theta} -\frac{1}{N} \sum_{i=1}^{N} \log P(y_i \mid x_i, \theta)$

$-\frac{1}{N} \sum_{i=1}^{N} \log P(y_i \mid x_i, \theta)$

$= -\frac{1}{N} \sum_{i=1}^{N} \log (2\pi\sigma^2)^{-\frac{1}{2}} + \boxed{\frac{1}{N2\sigma^2} \sum_{i=1}^{N} (y_i - w^T x_i)^2}$

assume "σ" is fixed (e.g. $\sigma = 1$)

$\hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} (y_i - w^T x_i)^2$

Same as minimize MSE

$\Rightarrow$ MSE is a special case of Cross entropy