220514

① Ridge and Lasso Regression { Overfitting, Bias, Variance, Underfitting }

② ElasticNet Regression

③ Logistic Regression

Underfitted Model

High bias + Low/High Variance

Train ● Test ●

Generalized Model

Low bias + Low variance

Train ● Test ●

Overfitted Model

Low bias + High variance

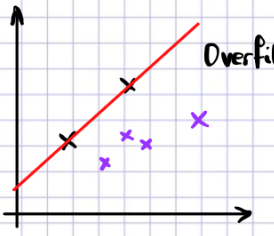
Train ● Test ●

Bias : Training Data

Variance : Test Data

Ridge and Lasso Regression

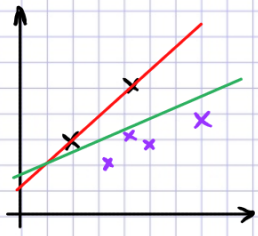
Train Error ↓ | Test Error ↑ { Low bias and high variance }



Cost function

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 = 0$$

Ridge Regression (L2- Regularization)



Cost function (modified)

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda\,(\text{slope})^2$$

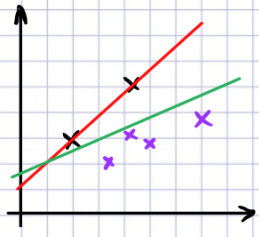
λ: hyperparameter; $(\text{slope})^2$: penalty

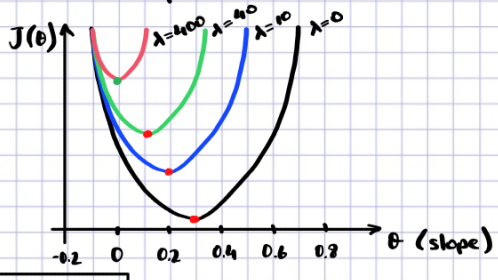
$$= \text{small number} + 1\cdot(0.9^2)$$

(λ = 1, slope = 0.9)

$$\approx 0.81$$

Relationship between λ and $(\text{slope})^2$



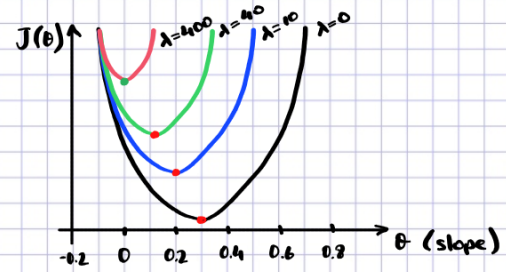


λ ↑	slope ↓	Reduce Overfitting

Global Minima is shifting, while overfitting is being reduced, and if our hypothesis is in form of $h_\theta(x) = \theta_0 + \theta_1 x_1$

by reducing the slope (θ_1) we are neglecting feature x_1.

Lasso Regression (L1-Regularization)



$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \underset{0.13 \to 0}{\theta_3 x_3} + \ldots + \theta_n x_n$$

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda\cdot|\text{slope}|$$

feature not correlating at all ≈ 0 → Feature neglected

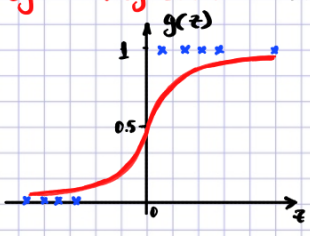
λ ↑	slope ↓	Feature Selection

Elastic Net Regression

Reducing Overfitting — feature selection

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda_1(\text{slope})^2 + \lambda_2|\text{slope}|$$

Logistic Regression [Binary Classification]



$h_\theta(x) = g(\theta_0 + \theta_1 x_1)$

$z := \theta_0 + \theta_1 x_1$

$h_\theta(x) = g(z)$ [activation function]

$g(z) = \frac{1}{1+e^{-z}}$

$$h_\theta(x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1)}}$$

Training set:

$\{(x^1, y^1), (x^2, y^2), \dots (x^n, y^n)\}$

$y = \{0, 1\}$ → 2 outputs → Binary Classification

$h_\theta(z) = \frac{1}{1+e^{-z}}$, $z = \theta_0 + \theta_1 x_1$, $\theta_0 = 0 \Rightarrow$ intercept = 0

Main Aim: Change θ_1, so that our sigmoid function classifies data point

Logistic Regression Cost Function (Log loss)

$$J(\theta) = \begin{cases} -\log(h_\theta(x)) & \text{if } y=1 \\ -\log(1-h_\theta(x)) & \text{if } y=0 \end{cases}$$

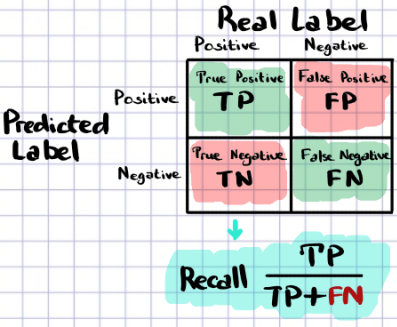
$$J(\theta_0, \theta_1) = -y \cdot \log(h_\theta(x^{(i)})) - (1-y) \cdot \log(1-h_\theta(x^{(i)}))$$

$$J(\theta) = -\frac{1}{2m} \sum_{i=1}^{m} \left[y^{(i)} \cdot \log(h_\theta(x^{(i)})) + (1-y^{(i)}) \cdot \log(1-h_\theta(x^{(i)})) \right]$$

$h_\theta(x^{(i)}) = \frac{1}{1+e^{-(\theta, x_1)}}$ ⇒ hypothesis

repeat until convergence $\{ \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} (J(\theta)) \}$

Performance Metrics (Binary Classification)



$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{Recall} \frac{TP}{TP+FN}$$

We can choose between Precision and Recall after clarifying what our main aim is, whether it's to reduce FN's or FP's.

E.g. Spam classification $\{FP \downarrow\} \Rightarrow$ Precision

Patient's cancer (Y/N) $\{FN \downarrow\} \Rightarrow$ Recall

But say we have a problem where we need minimal values possible of both FP and FN. In this case we can use F-Beta Score using 1 as β-value (F1 score)

$$\text{F-Beta Score} = (1+\beta^2) \frac{\text{Precision} \times \text{Recall}}{\beta^2 \times (\text{Precision} + \text{Recall})}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})}$$ [Harmonic Mean]

We can adjust this metric by using different values as β depending on what is more important for us whether it is to minimize FP's or FN's

Case 1: minimizing FP's is more important : FP >> FN

⇒ reduce β Ex: β=0.5

⇒ F-Beta Score $= (1+0.25) \times \frac{\text{Precision} \times \text{Recall}}{0.25 \times (\text{Precision} + \text{Recall}} = 5 \times \frac{\text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall}}$

Case 2: minimizing FN's is more important : FN >> FP

⇒ reduce β Ex: β=2

⇒ F-Beta Score $= (1+4) \times \frac{\text{Precision} \times \text{Recall}}{4 \times (\text{Precision} + \text{Recall}} = 1.25 \times \frac{\text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall}}$