Linear Regression 3

Agenda:

1) Assumptions of linear Regression.
2) Closed form $\hat{w}^*$ / Normal Equation.
3) Polynomial Regression
4) Train test split
5) Bias/Var Vs. Underfitting

Assumptions:

Linear Regression

Pick a Loss → Minimize Loss (Optimize)

Assumptions

1) Assumption of Linearity

$y = f(x)$

2) No Multicollinearity (Features)

Multicollinearity

$f_1 \quad f_2 \quad f_3$

$f_1 = w_2 f_2 + b$

$f_2 = w_3 f_3 + b$

$f_1 = w_3 f_3 + w_2 w_3 + b$

$\rightarrow$ Linearly correlated features.

$\hat{y} = w_1 f_1 + w_2 f_2 + w_3 f_3 + b$

Error

Why MC is a Problem?

$\hat{y} = w_1 f_1 + w_2 f_2 + w_3 f_3 + b$

$\hat{y} = w_1 f_1 + w_2 f_2 + w_3 f_3 + b$ (assume $w_1 = 0$)

$\hat{y} = w_2 f_2 + w_3 f_3 + b$

$\hat{y} = w_2 f_2 + w_3 f_2 + b$

$\hat{y} = w_2 f_2 + w_3 f_2 + b$

$\hat{y} = (w_2 + w_3) f_2 + b$

$\hat{y} = w_2 f_2 + b$

$\rightarrow w_1 = 0, w_2 = 0.5, w_3 = 0.5$

NOTE: w_1 could be any value for these features.

$\rightarrow$ Predictions will be the same.

$|w|$ $\rightarrow$ w_1 can not be made for future instances.

$w_1 = 0$

If MC $\rightarrow$ weights cannot be interpreted.

Cannot do feature imp.

3) VIF (Variance Inflation Factor)

$f_1 \quad f_2 \quad f_3 \quad f_4$

$\rightarrow$ 4 linear Regression.

for all other features $\rightarrow R^2$

$(-\infty \rightarrow 0-1)$

for f_1 all other features $\rightarrow R^2$

f_1, f_2, f_3, f_4

$\rightarrow R^2$

$w_1 = w_2 f_2 + w_3 f_3 + w_4 f_4 + \dots + b$

$f_1 \rightarrow w_1$

$f_2 \rightarrow w_2$

$f_3 \rightarrow w_3$

$f_4 \rightarrow w_4$

$R^2 \rightarrow$ score

Get $R^2 \rightarrow 1$ $\rightarrow$ we other features with linear corel. can have R^2

Hence MC exists in f_1

$VIF_1 = \frac{1}{1-R^2}$

if $R^2 = 1$ $\rightarrow$ $VIF_1 \rightarrow \infty$

if $R^2 \rightarrow 0$ $\rightarrow$ $VIF_1 \rightarrow 1$

if $R^2 \rightarrow -\infty$ $\rightarrow$ $VIF_1 \rightarrow 0$

$(0, \infty)$

$[1, \infty)$

$VIF_1 \uparrow \Rightarrow$ MC exists.

Rule of thumb

$VIF_1 > 10$: high MC

$VIF_1 > 5$: high MC

log(1.5)

Is MC multicollinearity?

$f_1 \quad f_2 \quad f_3 \quad f_4$

$\rightarrow$ linear fit $\rightarrow$ R^2 $\rightarrow$ MC

4) errors/residuals are normally distributed.

$y = \frac{1}{1+e^{-(w^T x + b)}} + \epsilon$

$\epsilon \sim N(0, \sigma^2)$

$y - \frac{1}{1+e^{-(w^T x + b)}} = \epsilon$

standardization

$y - \hat{y} = \epsilon$

Residuals

$\epsilon = y - \hat{y}$

$\epsilon \sim N(0, \sigma^2)$

Plot ϵ vs $\hat{y}$

ϵ vs $\hat{y}$

$\rightarrow$ Residuals

$\rightarrow$ Non-Linearity

$\rightarrow$ outliers

Impact of Outliers.

$y = \frac{1}{1+e^{-(w^T x + b)}}$

5) Homoscedasticity should not exists.

Heteroscedasticity.

6) No Autocorrelation

Time Series analysis

$y_i = f(x_i) + \epsilon_i$

Avg i^{th} datapoint should be independent/random.

$(x^{(i)}, y^{(i)})$

Closed Form Sol^n / Normal Sol^n

Normal $\hat{w}^*$ $\rightarrow$ Analytically solving

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

$\boxed{\hat{w}^* = (X^T X)^{-1} X^T Y}$

$\frac{\partial (w^T x)}{\partial x} = w$

$\frac{\partial (x^T w)}{\partial x} = x$

X : Input data $(n \times d)$

Y : output $(n \times 1)$

w^* : Best weights $(d \times 1)$

Derivation

$\hat{y} = X w$

$(n \times 1)$

$(n \times d)(d \times 1)$

$\Rightarrow (y - \hat{y})^2$

$\boxed{\text{Error} = (y - Xw)^T (y - Xw)}$

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

$(AB)^{-1} = B^{-1} A^{-1}$

$\Rightarrow (y^T - (Xw)^T)(y - Xw)$

$\Rightarrow (y^T - w^T X^T)(y - Xw)$

$\Rightarrow (y^T - w^T X^T)(y - Xw)$

$\boxed{\text{Error} = y^T y - y^T X w - w^T X^T y + w^T X^T X w}$

$\boxed{\frac{\partial \text{Error}}{\partial w^T} = 0}$

Solve.

You get w^*

$\frac{\partial \text{Error}}{\partial w^T}$ $\rightarrow$ Derivative w.r.t vector

$\rightarrow$ convex

$\Rightarrow -2X^T y + 2X^T X w = 0$

$\Rightarrow 2X^T X w = 2X^T y$

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

Normal $\hat{w}^*$

1) Closed form solution.

2) Complexity is cubic.

3) you need to invert $X^T X$ matrix.

4) Direct method.

5) Training of $(X^T X)^{-1}$ is very costly.

$(n \times d)$

$(d \times d)$

$O(d^3)$

Solve.

1) Iterative approach.

2) Iterative solution.

3) Gradient descent.

4) Stochastic Gradient Descent

5) Mini-Batch G.D

Stochastic Gradient Descent | Mini-Batch G.D

In G.D

$w_t^{(j)} = w_t^{(j)} - \eta \frac{\partial L}{\partial w_t^{(j)}}$

$\boxed{\frac{\partial L}{\partial w_t^{(j)}}}$

$\boxed{\frac{\partial L}{\partial w_t^{(j)}}} = \sum_{i=1}^{n} (y^{(i)} - y^{(i)})^2$

Advantage $n \times 100$

$\rightarrow$ train weights for each data.

Batch Gradient Descent.

Solution?

$\rightarrow$ Randomly select K data points from n.

$\rightarrow$ Calculate gradient using these K data points.

$w_t^{(j)} = w_t^{(j)} - \eta \frac{\partial L}{\partial w_t^{(j)}} \sum_{i=1}^{K} (y^{(i)} - y^{(i)})^2$

$\rightarrow$ if $K = 1$ $\rightarrow$ Stochastic G.D

if $K = n$ $\rightarrow$ Batch G.D

if $K(1, n)$ $\rightarrow$ Mini-Batch G.D

$K = 256$

Gradient per example

$\rightarrow$ convergence

$\rightarrow$ Gradients are noisy.

Polynomial Regression

Linear Reg $\rightarrow$

Linear Reg $\rightarrow$ high bias $O(1)$

$\rightarrow$ Linear fitter.

$\text{Linear reg} = w_1 x + w_0$

$\text{Quad reg} = w_2 x^2 + w_1 x + w_0$

$y^{(i)} = f_w(f_1, f_2)$

$f_1 = x, f_2 = x^2$

Separating is linear feature.

Not MC.

Degree of polynomial

I would say quad or cubic

$\rightarrow$ Bias

$\rightarrow$ Variance

$\rightarrow$ Underfitting

$\rightarrow$ Overfitting