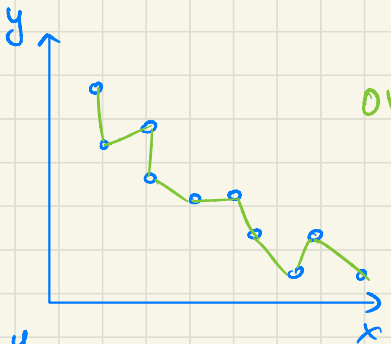
L1 & L2 REGULARISATION



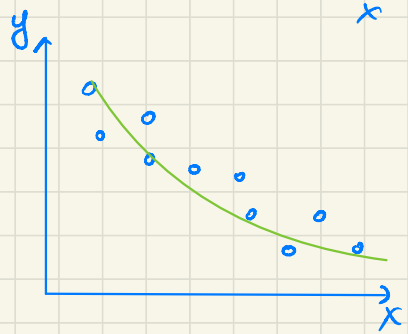
↳ adding information

overfitting: $y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$

↓

try to make $\theta_3 \dots \theta_n$ almost zero so that we rule out those factors ∴ shrink parameters to help generalize better

↓



balanced: $y = \theta_0 + \theta_1 x_1 + \theta_2 x_2$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - h_\theta(x_i))^2$$

controllable ↓

$$\text{now, } MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - h_\theta(x_i))^2 + \lambda \sum_{i=1}^{n} \theta_i^2$$

if θ gets bigger, the term $\lambda \sum_{i=1}^{n} \theta_i^2$ gets bigger

∴ penalizing higher values of θ to ↓ MSE

normally b/w 0 and 1

$L2 = \theta_i^2$

$$\text{in L1, } MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - h_\theta(x_i))^2 + \lambda \sum_{i=1}^{n} |\theta_i|$$

* Regularization techniques are commonly used for linear regression

Ridge Regression:

↳ L2 Regularization

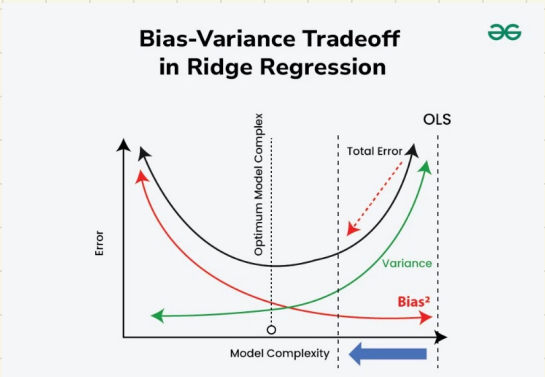
↳ Addresses problem of multicollinearity among predictor variables which occurs when independent variables are highly correlated ∴ leads to unreliable estimates

How? Adds a regularization term to least squares to penalize large coefficients, in the form of penalty ∴ also solves overfitting

Example: Price = 1000 . Size − 500 . Age + Noise

↓ after L2

Price = 800 . Size − 300 . Age + Less Noise



- As λ↑, bias ↓ and var ↑
- goal: find λ that balances bias and variance

How to select a ridge parameter?

↳ Cross-validation: data split into subsets; model trains on some and validates on the others

- K-Fold CV: data split into K subsets, trained on K-1 and validated on one; repeated K times s.t. each fold serves as validation set once
- Leave-One-Out CV

↳ Generalized CV

Pros:

- Bias vs variance
- retains original parameters

Cons:

- Introduces bias (can lead to underestimation of true effects of predictors)
- Choosing ridge parameter can be challenging

eg:

x_i	y_i	Prediction
1	2	2
2	4	4
3	6	6
4	8	8

$\therefore y = \beta_0 + \beta_1 x$

$\therefore y = 0 + 2x$

$\therefore y = 2x$

$\therefore \hat{y} - y_i = 0$

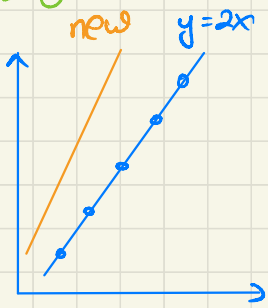
$\therefore (\hat{y} - y_i)^2 = 0$

$\therefore$ cost function $= 0 + \lambda \cdot \text{slope}^2$

$= 0 + 1 \cdot 2^2 = 4$

↑ penalty

$\therefore$ change coefficients



L1 Regularization

- shrinks certain parameters to 0, making them obsolete
- choosing one feature from group of correlated features; in the real world, some features are highly correlated (eg: year was built and # rooms in the house)

$\therefore$ reduces model complexity

- robust to outliers

L2 Regularization

- makes certain parameters small but not 0 so that they still have some influence
- not robust to outliers