Regularisation 101

Regularization 101

Problem

- Dataset with MANY features
- Includes a lot of noise
- Model that takes all features into account
  - Very flexible
  - Very prone to overfitting

- How can you know which features are relevant?
  - Domain knowledge
  - Feature engineering
  - Dimensionality reduction methods
  - Use regularization!

Regularization 101

Introduction

- Type of regression
- Model coefficients penalized
- Avoids overfitting by not-learning a too complex model
- Reduces model variance at the cost of bias increase
- How: by adding a penalty term to cost function

Multivariate Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Ordinary Least Squares

Minimize Cost Function

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \color{red}{PenaltyTerm}$$

Regularization 101

L1 / L2 regularization

Lasso regression (L1 regularization)

- Least Absolute Shrinkage and Selection Operator
- Adds absolute value of magnitude of coefficient
- Penalizes high coefficients, sets to zero

Ridge regression (L2 regularization)

- Adds squared coefficients to cost function
- High coefficients very costly
- Coefficients never zero

Minimize Cost Function

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

Regularization 101

Difference Lasso / Ridge Regression

Lasso Regression

- better, if useless variables existent
- better, if few variables have high coefficients and other variables coefficients close to zero
- removes features from model

→ acts like feature selection

Ridge Regression

- minimizes coefficients
- Reduces coefficients asymptotically close to zero
- Performs better, if there are many variables with similar coefficient values
- Ridge better, if most variables useful

Regularization 101

Penalty Factor Lambda

- Lambda Range: 0 to +∞
- Shrinkage penalty increases with lambda