Optimization Methods ===================

Machine Learning - Finding an Optimal Model ——————————————-

1. model selection 2. model parameter learning - feature engineering - hyper parameter selection 3. model evaluation

- available data is used to optimize function(m) - closed form solution - problem can be solved mathematically - problem too large? Search for solution

Search Methods ———

* calculus-based * direct-method * indirect-method - guided random search techniques - simulated annealing - evolutionary algorithms - enumerative techniques - dynamic programming

Zero Order / Direct Search Methods

- even sampling / grid search - random sampling / random search

Machine Learning

- simple linear regression (one variable)

x y predict error x y w+wx y-(w+wx) x y w+wx y-(w+wx) ... ... xn yn w+wxn yn-(w+wxn)

- RSS - residual sum of square / sum of square errors

Gradient Descent

differentiatiable function g, fixed step length a, and initial point w k = 1 repeat until stopping condition is met: w$^k = w^(k-1) - ag(w^(k-1))k < -k+1$

gradient: g(w$^(k-1))$

Newtons Method

twice differentiable function g, and initial point w$^0 k = 1 repeatuntilstoppingconditionismetsolvet$
$1))w^k =^2 g(w^(k-1))w^(k-1) - g(w^(k-1)) for w^k k < -k+1$

Tangent Curvature

At high curvature true point, approximation of the curve is better than approximation by Tangent. Optimization (Search) Methods - Mathematical/analytical solution -> applicable when the problem can be defined by mathematical equation and equation should be differentiable.

- Iterative methods -> Numrical optimization

SAmple Error/Loss Functions Goal -> to find or search for the model or solution with the minimum loss or error.

Derivatives of a multivariate function

First derivative -> Jacobian Jacobian is a vector-valued function that takes in matrix of all it's first-order partial derivatives. If the matrix is square, the determinant is referred as Jacobian determininants.

J$_i j = delta f_i / delta x_j$

Second derivative -> Hessian

Zero Order or Direct Search - Popular as it works well in the practice - 1st and second order methods are not appliacble to all nonlinear optimization problems

- Direct search methods have succeeded when more elaborate approaches failed

- THe main issue with this technique is finding the optimal point in the large search space thus both time and accuracy will be concerned.

================================

Batch Gradient Descent ——————-

* batch gradient descent performs parameters update in batch iteration (epoch) + fixed learning rate + straight trajectory - slow if large dataset

Mini-batch Gradient Descent ——————–

* mini-batch uses random mini batches to perform parameters update for n epochs + faster than batch + can avoid redundant samples - may not converge and probably require learning decay

Stochastic Gradient Descent ——————–

* stochastic gradient descent performs parameters update on each sample * like mini-batch but with one example used for each learning step + faster than others - more noise - large variance since only one example used for each learning step

Regularization ==============

- reduce variance-bias trade off - avoid the model overfit - total error = RSS + ||w|| - high - low variance

================================ Bias is the error in fitting a model

Bias high -> underfit Bias low -> just nice Bias very low -> overfit

================================ Variance is the measure of the variation among the fits of different samples training sets on the same population

================================== Bias Variance

AS the complexity of the model increase, the bias decrease. However, teh variance will increase as the model complexity increase to a threshold.

A good fit model has minimum error on the training set, however it may not generalise well to the population. Hence, the validation set might give a higher error.

Therefore, there is a trade off between bias and variance.

——————————————— Identify High Bias and High Variance

HIgh bias: - HIgh training error - Validation erro is similar to training error

HIgh Variance: - Low training error - HIgh validation error

———————————— Ways to Reduce Bias

- Increase training data - Increase model complexity - Increase number of features (increase the information for the model to learn) -> normally used in linear model - Use model selection to increase model selection - Normally use in non-linear model

==

Types —— Different between l1 and l2 is the penalty term

- l1 norm - ||w|| + ||w|| - l2 norm - ||w^2|| $+ ||w^2|| -> ridgeregression -> squaremagnitude$

Ridge Regression ————— - use l2 regularization - reduce model complexity by shrinking model parameters is the tuning parameter that decides how much we want to penalise the fexibility of our model.

Lasso Regression ————— - use l1 regularization - least absolute shrinkage selector operator - perform feature selection (eliminate some features) - reduce other parameters - useful when large number of features involved

Elastic net Regression ——————— - use both l1 and l2 regularization - error = RSS + ||w||¹ + ||w||² - alpha = + - $l1_ratio = /(+) - ifl1_ratio = 1, = 0, lassoregression - ifl1_ratio = 0, = 0, ridgeregression - otherwise, combinationofridgeandlass$

Parameters which define the model architercture are referred as hyperparamters, thus, the process of searching the ideal model architecture is referred as hyperparameter tuning.

Applications ————

1. true performance gain 2. hyper parameters selection