6143 Machine Learning Project

Guandong Kou, Tianhao Zhou

Abstract

The purpose of this project is to train a text classifier which could determine the variety of the wine being reviewed based on the review text. We will be using the wine magazine comment dataset provided by Kaggle at https://www.kaggle.com/zynicide/wine-reviews. This project focuses on the feature engineering, model selection, model training and performance evaluation of multinomial naive bayes, logistic regression, support vector machine and neural networks.

The background of this project can be found on Kaggle (Kaggle wine dataset). The code can be found from the link to the Colab Notebook.

Methodology

The feature engineering of the comment data is done via TF-IDF. Then, four different models are trained and evaluated.

Model selection

- Naive Bayes classifier

1. Introduction

Naive Bayes classifiers are a set of probability classifiers which are based on Bayes' theorem. Those classifier require all the features' value are independent so it may not work so well in some situation that features are highly correlated. However, it is simple to use and only require a small amount of training data, which turns out quite useful in many complex real-world situations.

2. Mathematical model

$$\hat{y} = \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}}\ p(C_k)\prod_{i=1}^{n} p(x_i \mid C_k).$$

In this formula, x represents some n independent features, C_k represents k class labels.

3. Advantages
a. Easy to use and understand.
b. Can be done in linear time, especially time-saving when data set is large.
c. Perform well when training data is limited.
d. Robust to noise and missing training point.
4. Drawbacks
a. Accuracy is limited due to the independent assumption.
b. Bad at language processing and combination problems.

- Logistic Regression

1. Introduction

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. It is a extremely universal algorithm can be used in many fields. First of all, logistic regression is a super star in medical imaging field. A variety of cancer identifications are assisted by logistic regression, for instance, breast cancer, prostate cancer, liver cancer, etc. Combining logistic regression algorithm and expert observation, the accuracy of diagnosing could be more than 99 percent. Another than medical imaging field, logistic algorithm can be also very useful in engineering and marketing. We could use it to predict the probability of certain custom buying a product based on custom’s age, income, sex, education, so on and so forth. predicting the probability of failure of a given process or system is also the speciality of logistic regression. Furthermore, in natural language processing, logistic regression could be used to process sequential data. However, logistic regression is bad at problems like identifying hand-written characters and numbers. In those cases, classification can not be easily decided by the overlap between test and train data.

2. Mathematical model

Logistic Model for Binary Classification

$$z = \sum jw_j x_j + w_0$$

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-z}}$$

Logistic Loss Function for binary classification

$$J(w) = \sum_{i=1}^{n} \ln[1 + e^{z_i}] - y_i z_i$$

3. Advantages

a. Easy to implement and efficient to train.

b. Make no assumptions in feature space.

c. Good accuracy for many simple data sets and it performs well when the data set is linearly separable.

d. Can use coefficients to indicate the importance of features.

e. Can use L1 or L2 regularization to avoid over-fitting.

4. Drawbacks

a. Perform badly when the number of observations less than the number of features.

b. It only construct linear boundaries.

c. Make assumption of linearity between the dependent variable and the independent variables.

d. It is tough to obtain complex relationships using logistic regression.

- SVM

1. Introduction

Support-vector machine(SVM) is another simple and powerful algorithm in machine learning. Unlike the logistic regression, SVM performs pretty well on non-linear classification using kernel trick. More formally, a support-vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin, the lower the generalization error of the classifier.Whereas the original problem may be stated in a finite-dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. Therefore, SVM is helpful in text and hypertext categorization, image classification, recognizing hand-written characters and permutation tests, etc.

2. Mathematical model

Hinge Loss

$$L(w,b) = \max(0,1 - y_i(w^T x_i + b))$$

SVM Optimization

$$J(w,b) = C\sum_{i=1}^{N} \max(0,1 - y_i(w^T x_i + b)) + \frac{1}{2}\|w\|^2$$

In this formula, C represents a coefficient control the final margin.

Margin equals to $\frac{1}{\|w\|}$.

Kernel Trick

$$K(x_i, x) = \phi(x_i)^T \phi(x)$$

$$z = b + \sum_{i=1}^{N} a_i y_i K(x_i, x)$$

$$\hat{y} = sign(z)$$

3. Advantages

a. SVM works relatively well when there is a clear margin of separation between classes.

b. SVM is more efficient in high dimensional spaces.

c. SVM is relative memory efficient.

4.Drawbacks

a. SVM is not efficient when data sets are large.
b. SVM does not perform well when data sets have noticeable noise.
c. In cases where number of features for each data point exceeds the number of training data sample, the SVM will underperform.

- Decision Trees

1. Introduction

Decision tree is one of the predictive modelling approaches which is constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both regression and classification tasks. Tree models where the target variable can take a discrete set of values are called classification trees. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Classification And Regression Tree (CART) is general term for this.

2. Mathematical model

Gini impurity

Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.

$$I_{G(p)} = \sum_{i=1}^{J}(p_i \sum_{k\neq i} p_k)$$

$$\sum_{k\neq i} p_k = 1 - p_i$$

p_i represents the probability of an item with label i being chosen,

p_k represents the probability of a mistake in categorizing that item.

J represents number of classes.

Information gain

$$H(T) = I_E(p_1, p_2, \ldots, p_J) = -\sum_{i-1}^{J} p_i \log_2 p_i$$

$$IG(T, a) = H(T) - H(T \mid a)$$

$$= -\sum_{i-1}^{J} p_i \log_2 p_i - \sum_{i-1}^{J} - p_r(i \mid a) \log_2 p_r(i \mid a)$$

$H(T)$ represents information entropy, where $p_1, p_2 \ldots$ are fractions that add up to 1 and represent the percentage of each class present in the child node that results from a split in the tree.

3. Advantages

a. Can handle both categorical and numerical data
b. Resistant to outliers, hence require little data preprocessing
c. New features can be easily added
d. Can be used to build larger classifiers by using ensemble methods.
e. Unlike Neural Network, decision tree model is a open-box model.

4. Drawbacks

a. Trees can be very non-robust. A small change in the training data can result in a large change in the tree and consequently the final predictions
b. Decision trees are vulnerable to over-fitting because of the prune process.
c. Can create biased learned trees if some classes dominate.

- Neural Network

1. Introduction

Neural Network is a computing system gets inspired by biological neural networks. It is built by basic neurons, which take inputs, do some computations and produce output[Fig1]. Neural Network consist of connected neurons. One neuron’s input is other neuron’s output. The neurons are typically organized into multiple layers[Fig2]. Neurons of one layer connect only to neurons of the immediately preceding and immediately following layers. The layer that receives external data is the input layer. The layer that produces the ultimate result is the output layer. In between them are zero or more hidden layers. Single layer and unlayered networks are also used. Between two layers, multiple connection patterns are possible. They can be fully connected, with every neuron in one layer connecting to every neuron in the next layer. They can be pooling, where a group of neurons in one layer connect to a single neuron in the next layer, thereby reducing the number of neurons in that layer. Neurons with only such connections form a directed acyclic graph and are known as feedforward networks. Alternatively, networks that allow connections between neurons in the same or previous layers are known as recurrent networks. Because of the ability of reproducing and modeling nonlinear process, Neural Network can be found in countless disciplines and situations.

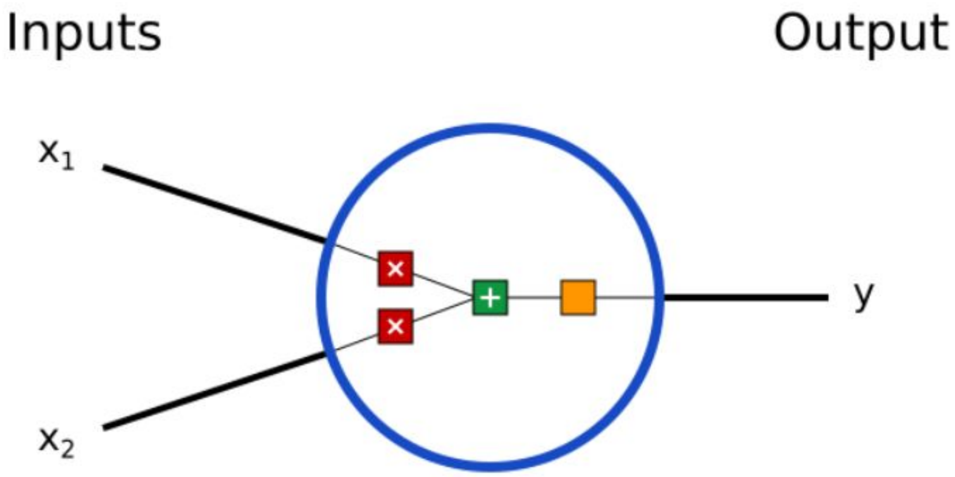


Fig1. A 2-input neuron example.

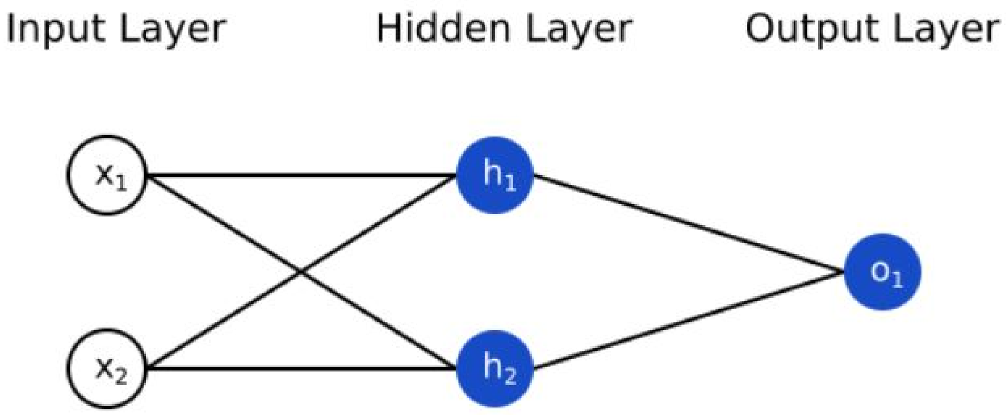


Fig2. A general Neural Network block diagram.

2. Mathematical model

Forward propagation

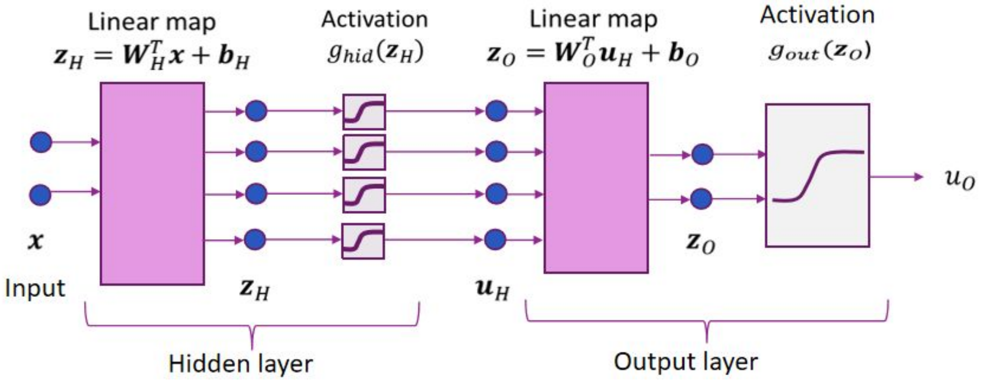
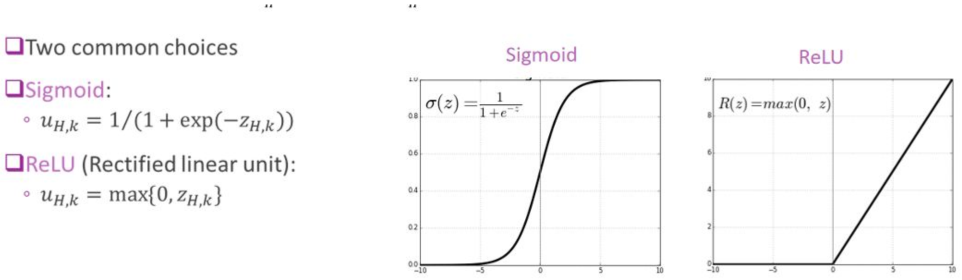

Linear map
$\boldsymbol{z}_H = \boldsymbol{W}_H^T \boldsymbol{x} + \boldsymbol{b}_H$
Activation
$g_{hid}(\boldsymbol{z}_H)$
Linear map
$\boldsymbol{z}_O = \boldsymbol{W}_O^T \boldsymbol{u}_H + \boldsymbol{b}_O$
Activation
$g_{out}(\boldsymbol{z}_O)$
u_O
$\boldsymbol{x}$
Input
$\boldsymbol{z}_H$
$\boldsymbol{u}_H$
$\boldsymbol{z}_O$
Hidden layer
Output layer

Fig3. Neural Network's forward propagation flow chart.

Selecting the Hidden Activation

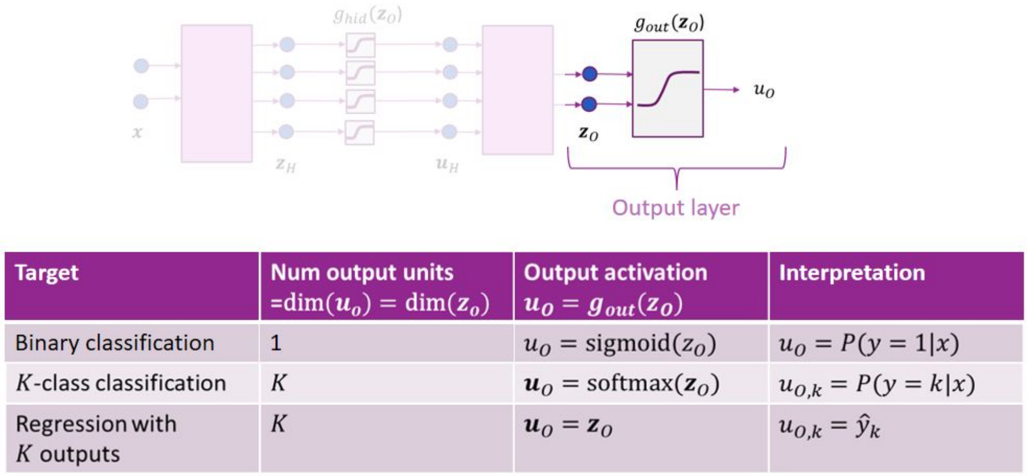

❑Two common choices

❑Sigmoid:
- $u_{H,k} = 1/(1 + \exp(-z_{H,k}))$

❑ReLU (Rectified linear unit):
- $u_{H,k} = \max\{0, z_{H,k}\}$

Fig4. Two common activation functions for hidden layer.

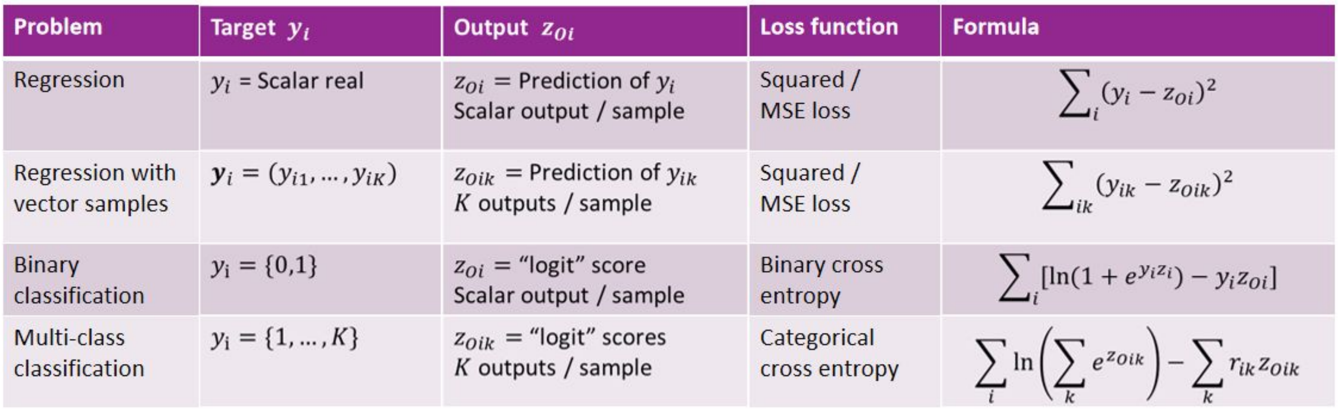
Selecting the Output Activation



Target	Num output units =dim($\boldsymbol{u}_o$) = dim($\boldsymbol{z}_o$)	Output activation $\boldsymbol{u}_O = g_{out}(\boldsymbol{z}_O)$	Interpretation
Binary classification	1	$u_O = \text{sigmoid}(z_O)$	$u_O = P(y = 1 \vert x)$
K-class classification	K	$\boldsymbol{u}_O = \text{softmax}(\boldsymbol{z}_O)$	$u_{O,k} = P(y = k \vert x)$
Regression with K outputs	K	$\boldsymbol{u}_O = \boldsymbol{z}_O$	$u_{O,k} = \hat{y}_k$

Fig5. Common activation functions of hidden layer for binary classification, K-class classification and regression.

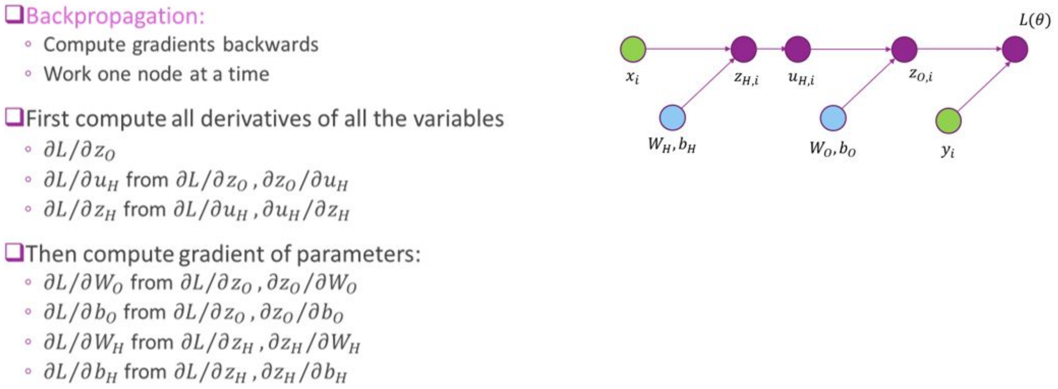
Selecting the Loss Function for different types of problems



Problem	Target y_i	Output z_{Oi}	Loss function	Formula
Regression	y_i = Scalar real	z_{Oi} = Prediction of y_i Scalar output / sample	Squared / MSE loss	$\sum_i (y_i - z_{Oi})^2$
Regression with vector samples	$\boldsymbol{y}_i = (y_{i1}, \dots, y_{iK})$	z_{Oik} = Prediction of y_{ik} K outputs / sample	Squared / MSE loss	$\sum_{ik} (y_{ik} - z_{Oik})^2$
Binary classification	$y_i = \{0,1\}$	z_{Oi} = "logit" score Scalar output / sample	Binary cross entropy	$\sum_i [\ln(1 + e^{y_i z_i}) - y_i z_{Oi}]$
Multi-class classification	$y_i = \{1, \dots, K\}$	z_{Oik} = "logit" scores K outputs / sample	Categorical cross entropy	$\sum_i \ln\left(\sum_k e^{z_{Oik}}\right) - \sum_k r_{ik} z_{Oik}$

Fig6. Loss functions for different tasks .

Back-propagation



❑Backpropagation:
- Compute gradients backwards
- Work one node at a time

❑First compute all derivatives of all the variables
- $\partial L / \partial z_O$
- $\partial L / \partial u_H$ from $\partial L / \partial z_O$, $\partial z_O / \partial u_H$
- $\partial L / \partial z_H$ from $\partial L / \partial u_H$, $\partial u_H / \partial z_H$

❑Then compute gradient of parameters:
- $\partial L / \partial W_O$ from $\partial L / \partial z_O$, $\partial z_O / \partial W_O$
- $\partial L / \partial b_O$ from $\partial L / \partial z_O$, $\partial z_O / \partial b_O$
- $\partial L / \partial W_H$ from $\partial L / \partial z_H$, $\partial z_H / \partial W_H$
- $\partial L / \partial b_H$ from $\partial L / \partial z_H$, $\partial z_H / \partial b_H$

Fig7. Illustration of back-propagation .

3. Advantages

a. Neural Networks are extremely versatile, can be used in both classification and regression problems.
b. Neural Networks are flexible, can be trained with any number of inputs and layers.
c. Unlike SVM or other traditional method, Neural Networks are good at dealing with nonlinear data with large number of inputs; for example, images. It is reliable in an approach of tasks involving many features. It works by splitting the problem of classification into a layered network of simpler elements.
d. The prediction of Neural Networks are efficient after training.

4. Drawbacks

a. Neural networks are black boxes, meaning we cannot know how much each independent variable is influencing the dependent variables.
b. Neural networks require the quantity and quality of training data.

- Ensemble Learning

1. Introduction

Ensemble learning models are not a specific model. They combine the decisions from multiple models to improve the overall performance. Ensemble models have different ensemble technique to solve different problems. Some basic ensemble techniques, such as Max Voting, Averaging and Weighted Average help us understand how ensemble learning works. Although in real problem solving, people usually use algorithm based on advanced ensemble techniques like bagging and boosting. In this article, we will focus on a powerful algorithm called XGB.

2. Ensemble techniques and Mathematical model

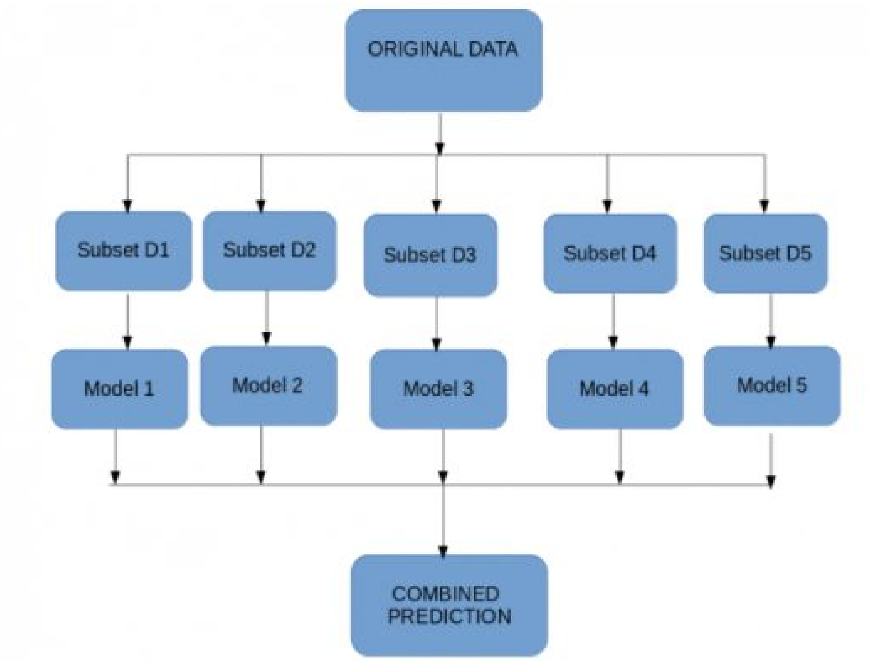
Bagging

The idea behind bagging is combining the results of multiple models to get a generalized result. But if we use same data to fit all the models, there is a high chance that all the results are the same which makes bagging useless. So how do we solve this problem? Bootstrapping is a perfect solution. Bootstrapping is a sampling technique in which we create subsets of observations from the original data set, with replacement. The size of the subsets is the same as the size of the original set.Bagging technique uses these subsets to get a fair idea of the distribution . The size of subsets created for bagging may be less than the original set.

Fig8. Bagging technique combining bootstrapping .

Boosting

Boosting is a sequential process, where each subsequent model attempts to correct the errors of the previous model. The succeeding models are dependent on the previous model. The following steps could illustrate how boosting works:



1. A subset is created from the original data set.
2. Initially, all data points are given equal weights.
3. A base model is created on this subset.
4. This model is used to make predictions on the whole data set
5. Errors are calculated using the actual values and predicted values.
6. The observations which are incorrectly predicted, are given higher weights.
7. Another model is created and predictions are made on the data set.

8. Similarly, multiple models are created, each correcting the errors of the previous model.
9. The final model (strong learner) is the weighted mean of all the models (weak learners).

The final model combines sub-models’ ‘speciality’ to gain great overall performance on the entire data set.

XGBoost

XGBoost (extreme Gradient Boosting) is an advanced implementation of the gradient boosting algorithm. XGBoost has proved to be a highly effective ML algorithm, which has high predictive power and is way faster than the other gradient boosting techniques. It also includes a variety of regularization, which reduces overfitting and improves overall performance. Hence it is also known as ‘regularized boosting‘ technique.

Loss function of XGBoost

$$\widetilde{L}^{(t)} = \sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2}h_i f_t^{\,2}(x_i)] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2$$

$$= \sum_{j=1}^{T}[(\sum_{i\in I_j} g_i)w_j + \frac{1}{2}(\sum_{i\in I_j} h_i + \lambda)w_j^2] + \gamma T$$

3. Advantages

a. General high accuracy, low bias and variance.
b. General high efficiency.
c. Highly flexible and robust.

4. Drawbacks

a. Could have over-fitting problem depending on specific situations .
b. Not perform well processing high dimensional sparse features(GBDT).