Week 1 Report

Haoxuan Yin
Department of Computer Science and Technology
Tsinghua University
yhx18@mails.tsinghua.edu.cn

Abstract

In order to decide if an ML method is good or not,we define training errors and come up with several ways to measure it.Then we introduce two naive models,one is decision tree and the other is neural network.

1 Overview

Machine Learning(ML) is considered a subset of Artificial Intelligence(AI).It relys on "training data" to "learn" how to solve a problem itself,rather than being given explicit instructions.The given dataset is commonly divided into two parts,one is training set which is used for training and the other is testing set which is used to test if the model is good enough.Generally the mission is to learn to predict the output based on a vector of given attributes.Depending on the output,the mission can be classification when the output is discrete,or regression when the output is continuous.

2 Model evaluation

2.1 Training errors

We define error rate as $E = a/m$,where m is the number of test cases and a is number of those which go wrong.Accuracy,on the contrary,is defined as $1 - E$.There are two kinds of errors,one is overfitting,in which case the model mistakes the features only the training set have as general ones,and the other is underfitting,in which case the model doesn't fully learn the features.

2.2 Evaluation mathods

In order to measure if a model is good enough,we need to divide the dataset into training set and testing set.A naive solution,called hold-out,is to randomly choose $2/3$ to $4/5$ of the dataset as training set.

Improvement can be made by a method called cross validation.It divides the dataset into k subsets,and in each of the k rounds,$k - 1$ subsets are used as training set and the other is used as testing set.In particular,if k equals the size of the dataset,it's called LOO(short for Leave One Out).

The method bootstrapping is used when the dataset is small.Every time it randomly picks a sample into the training set without erasing it from the dataset.The action is repeated m times,where m is the size of the dataset.It is guaranteed that there are still around a third of the dataset not picked,which can be used as the testing set.

2.3 Where do errors come from

We define mean squared error as

$$E(f;D) = \int_{x\ D} (f(x) - y)^2 p(x) \mathrm{d}x$$

After calculation we can find that

$$E(f;D) = bias^2(x) + var(x) + \varepsilon^2$$

$bias(x)$ is the difference between this model and the best possible model,which is what we want to minimize.$var(x)$ indicates how the result changes according to the change in the given dataset,that is the stability of the model.ε^2 is the difference between the given label and the actual label i.e. the difficulty of the task itself.

3 Decision tree

Take decision tree of a binary classification as an example.At every node,we choose an attribute that divides the samples best and divide the samples into two branches according to that.

There are a few problems though.If all the attribiutes are used up i.e. the set can't be divided any further,but the classification hasn't been completed i.e. there are still multiple labels in the set,then we have to choose the most frequent label.

An attribute used to divide the set is considered the best,if the two new sets we have are the purest.Information entropy is used to measure the purity of a set.It's defined as

$$Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k$$

The lower the entropy,the higher the purity.

Sometimes the tree can be too large to be effective,so we need pruning.Before dividing we can do prepruning and after building the tree we can do postpruning.

4 Neural network

One naive model is linear model,that is

$$f(x) = \sum_i w_i x_i + b$$

In order to imitate real neurons,sigmoid function is used.It's defined as

$$Sigmoid(x) = \frac{1}{1 + e^{-x}}$$

The combination of linear funtions and Sigmoid function gives us a neural network.