Introduction to Machine Learning (incomplete)

Aswin

Overview

I made this document as a way to learn ML for my Maters' thesis . Its not an original work,but a compilation of scribed notes of the ML course by Dr Sanjoy Dasgupta(UCSD), [SD] which I audited.Some parts are drawn directly/inspired from Andrew NG's Stanford CS229 course notes, [ANG].My primary reference was 'The Elements of Statistical Learning' by Trevor Hastie, Robert Tibshirani,Jerome Friedman, [HTF].So there will be considerable overlap with the aforementioned materials.This note might lack mathematical rigor but I have tried to give proofs and supplementary topics wherever necessary.Please write to me if you find any innaccuracies.I hope this this proves at least moderately interesting or useful for you.

Supervised Learning

In a typical scenario, we have an outcome measurement, usually quantitative (such as a stock price) or categorical (such as heart attack/no heart attack), that we wish to predict based on a set of features (such as diet and clinical measurements). We have a training set of data, in which we observe the outcome and feature measurements for a set of objects (such as people). Using this data we build a prediction model, or learner, which will enable us to predict the outcome for new unseen objects. A good learner is one that accurately predicts such an outcome. The examples above describe what is called the supervised learning problem. It is called supervised because of the presence of the outcome variable to guide the learning process.[HTF]

Variable Types and Terminology

The outcome measurement which we wish to predict denoted as ***outputs*** depend on a set of variebles denoted as ***inputs***. Classicaly,the *inputs* are independent varibales whereas *outputs* are dependent variables. The term *features* will be used interchangeably with inputs.

The *outputs* which we wish to predict can be qualitative or quantitative(as in blood sugar level). When the *outputs* are qualitative (as in spams or not spams),it is referred as categorical or discrete variables and are typically represented numerically by codes,as in -spam or not spam can be coded as -1 or 1. Depending upon the kind of output varible, the prediction task can be of two types: *regression* when we predict quantitative outputs and *classification* when we predict qualitative outputs.