Introduction to ML for AI Notes

Chang Liu

May 20, 2015

1 Introduction

This document is used for recording my learning for this article.

2 Content

2.1 Formation of learning

Target: for the samples from unknown process $P(Z)$, we have got some samples $D = \{z_1, z_2, z_3, ...z_n\}$, now that we want to find this process $P(Z)$.

Method: here we could first assume an loss function L, that takes parameters of **decision function** and **these samples** z_i, the decision function is what we assumed before like linear function or quadratic function, samples are the sets D. By minimizing the loss function $L(f, D)$, we can get the unknown variables $A, B...$ in the loss function and get its exact representation.

Result: the loss function L is just our assumed process $P(Z)$, we can assume many loss function and find the best one among them.

2.2 Supervised learning

For supervised learning, in the above section, we can know that there are two situations for that, that is:

1) Regression.
2) Classification.

2.2.1 Regression

In this category, let's assume that Y is the real-valued scalar or vector, then the output of function f(here we should recall that f is the decision function) should also be in the same set of Y, which means its value should be in the set Y.

Then loss function should be :

$$L(f,(X,Y)) = ||f(X) - Y||$$

2.2.2 Classification

Classification is different, as the output Y is finite integer(or symbol) corresponding to the index, and the **loss function** should be the **negative** *log-likelihood* value,

$$f_i(X) = P(Y = i|X)$$

And the loss function should be:

$$L(f,(X,Y)) = -\log f_Y(X)$$

NOTE: the decision function f must have the constrain that:

$$\sum_i f_i(X) = 1, f_i(X) \geq 0$$

2.2.3 Summary

From the above two kinds of **supervised learning**, I know that for regression since we can get continual value, so the loss function is to calculate the **absolute distance** between the decision function and real value.

But for classification problem, since it's the classify label or index, we cannot function, so here we use the likelihood, and to use the *sigmoid* function like the $\log X$ function. And all the samples forms the whole training set, so the constrain is also given as above.

2.3 Unsupervised learning

In unsupervised learning, the learning function f is used to characterize the unknown distribution function $P(Z)$, this function f is may have the following categories:

1) some is just the direct represent of the $P(Z)$, which is called density estimation.

2) some is to characterize where the density concentrates.

Clustering: clustering algorithm is used to divide the space into regions, that have two types:

1) hard partition. (k-means)

2) soft partition. (Gaussian mixed model)

3) form a new representation for Z. (Principle Component Analysis)

2.4 Local generalization

Defintion: when x_i is close to x_j, then the $f(x_i)$ should also be close to $f(x_j)$.

1) When the target unknown function is more than the input parameters X, then it's hard to generalize since there are so many variables, **curse of dimension**.

2) for example, there are n variables with 10 different value, then it has 10^n configurations.