Week3 Report

Zhengyi Wang
wangzhen17@mails.tsinghua.edu.cn

Abstract

This week, we talked about the fundemental of machine learning. Our talk is basically on *logistic regression* and *boosting*.

1 Logistic Regression

Logistic regression is not a regression question problem. It's actually a classification problem.

1.1 Basic Knowlegde

In the case that $y \in \{0, 1\}$, with parameter $W = (w_1, w_2, ..., w_n)$, we have

$$P(Y = 1|X; W) = \frac{1}{1 + e^{W^T X}}$$

and,

$$P(Y = 0|X; W) = \frac{e^{W^T X}}{1 + e^{W^T X}}.$$

More generally, y can take $\{y_1, y_2, ...y_R\}$, in which

$$P(Y = y_k|X; W) = \frac{e^{w_{k0} + \sum_{i=1}^{n} w_{ki} x_i}}{1 + \sum_{j=1}^{R-1} e^{w_{j0} + \sum_{i=1}^{n} w_{ji} x_i}} \quad (k \neq R)$$

and,

$$P(Y = y_R|X; W) = \frac{1}{1 + \sum_{j=1}^{R-1} e^{w_{j0} + \sum_{i=1}^{n} w_{ji} x_i}}$$

1.2 Optimistic Goal

To find a suitable W such that

$$W = argmax_W \Pi_i P(y_i|x_i; W)$$

which is equivalent to,

$$W = argmax_W \Pi_i \ln P(y_i|x_i; W).$$

Our optimistic goal can be equivalently transformed into the problem of maximize the log-likelihood function

$$l(W) = \sum_l y^l \ln P(y^l = 1|x^l; W) + (1 - y^l) \ln P(y^l = 0|x^l; W).$$

$l(w)$ is a concave function. We can find its maximum by *gradient descent method*.

Reading Group on Machine Learning (May 12th, 2019).

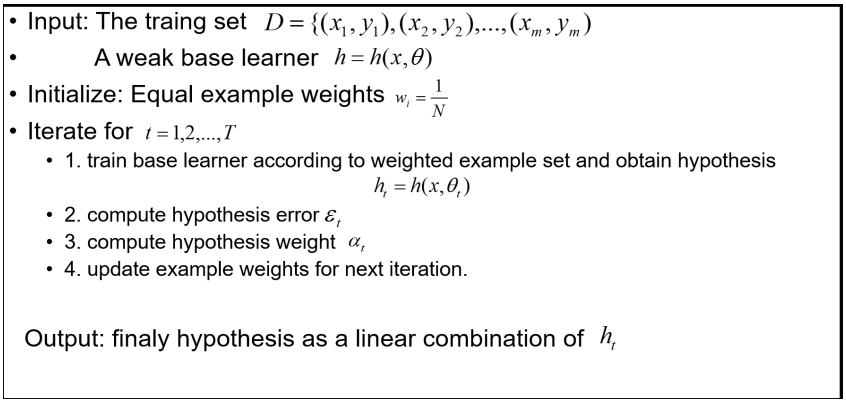

Figure 1: Adaboost Algorithm

2 Boosting

Boosting is one method of *ensemble learning*. The main idea of ensemble learning is to combine a lot of weak learner to get a better learner.
The main idea of boosting is that we adjust the weights of samples while training. The samples that were misclassified by the former model will get more importance in the training afterwards.

2.1 Algorithm

The algorithm can be shown in Figure 1.

In details, the hypothesis error is usually the sum of the weights of misclassified items (after normalization to [0,1]). $\alpha_k = \frac{1}{2} \ln \frac{1-\epsilon_k}{\epsilon_k}$. And $w_i^k = w_i^{k-1} e^{-y_i \alpha_k h(x_i; \theta_k)}$. $h(x_i; \theta_k) (\in \{0, 1\})$ is the prediction made by the current model.

2.2 Personal Veiw and Others

Compared to bagging, adaboost focuses on the misclassified items. Which means that is can not only smaller the variance of each weak learner, but also the bias of each weak learner.
Also, note that $error_{true}(H) \leq error_{train}(H) + O(\sqrt{\frac{Td}{m}})$.