Lecture 8: Classification, Part 1
Modeling Social Data, Spring 2017
Columbia University

March 10, 2017

Notes from ad3222

1 Part I - Professor Wiggins

1.1 Classification

Goal: Get as familiar with classification as we are with regression

Learning by example:

- Spam Detection
  - unwanted email
  - How can you detect spam?

Building a theory of 3s

- Trying to predict a categorical value
- no sense of distance
- Another Example
  - Image recognition/OCR
  - 1996 - Optical character recognition
  - Just nee to successfully predict (don't need exact estimates)
  - Like deep learning + atari

1.1.1 What is Classification? - Bananas vs. Oranges

What would Gauss do?
Under Gauss, each banana and orange comes from the normal distribution and we get factors like length and height. We then use prior probabilities on choosing a banana.

What if you don't know all the features or have many covariates?
For example, we can have things like time of purchase or smell, etc... Thus, we want to be able to build a model for high dimensional data.

1.1.2 Game Theory

Assume the worst possible distribution. Note that with more features, the harder this is to do.

Large Deviation Theory
This idea of maximum margin, where margin is the distance separating bananas from oranges.

SVMS are a class of models geared toward finding this margin.

1.1.3 Example Application - New York Times

Goal: Want to investigate if there is a link between traffic fatalities and a particular air bag manufacturer - takata

To do this, we want to find "interesting cases" so we want to construct some labels/features that classifies a data entry of a particular incident as interesting or not.

Example phrases that were interesting were "suddenly deployed" or "unexpectedly". This allowed journalist to find cases and this type of computer assisted reporting led to a massive recall of takata airbags.

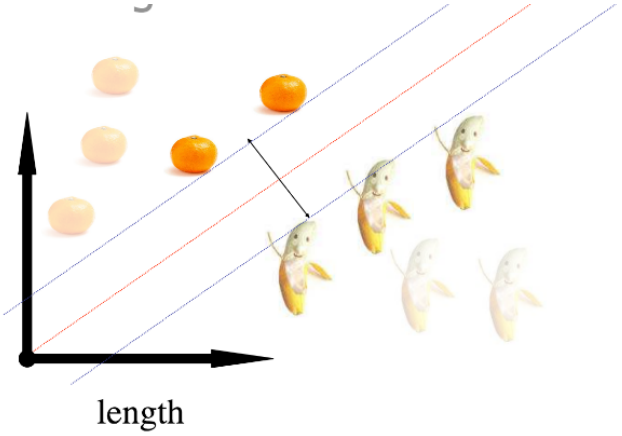

Figure 1: Bananas v Oranges: Margin/Large Deviation Theory

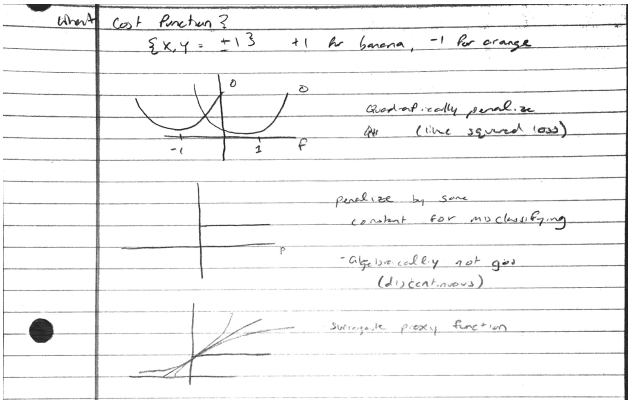

Figure 2: Examples of various cost functions

1.1.4 What is a cost function?

$$x, y = \pm 1 \rightarrow +1 = banana, -1 = orange$$

Misclassification often occurs from having no idea on how to step. Support proxy function gives tweaks on how to step.

1.1.5 MLE for Classification

Binary/Dichotomous/Boolean features + Naive Bayes
Goal: generalize and maintain linearity. Consider the spam problem of recognizing whether a peice of mail is spam.

Example - Bayes Rules Review Assume there you are testing for a disease that has infected 1 percent of the population

99 percent of sick patients test positive and 99 percent of healthy students test negative. Given that a patient tests positive, what is the probability that the patient is sick?

Bayes Theroem

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$$p(sick|+) = \frac{p(+|sick)p(sick)}{p(+)}$$

$$p(+) = p(+|sick)p(sick) + p(+|healthy)p(healthy)$$

$$p(sick|+) = \frac{p(+|sick)p(sick)}{p(+)} = \frac{99}{198} = 50\%$$

How does this relate to spam?

Build a one-word spam classifier:

$$p(spam|word) = \frac{p(word|spam)p(spam)}{p(word)}$$

where we just estimate the probabilities as ratios of counts.

Naive Bayes *Naive* comes from the fact that we are assuming every word appears independent of other words.

Bernoulli likelihood

$$p(x|c) = \prod_j \theta_{jc}^{x_j}(1 - \theta_{jc}^{1-x_{jc}})$$

Take log-likelihood and derivative to get MLE:

$$log(p(c|x)) = \sum_j x_j \log \frac{\theta_{jc}}{1-\theta_{jc}} + \sum_j \log 1 + \theta_{jc}$$

$$\frac{d}{d\theta} log(p(c|x)) = \frac{\sum_j x_j}{\theta} + \frac{\sum_j 1 - x_j}{1-\theta}$$

$$\theta_c = \frac{\sum_j x_j}{N}$$

which is just like regression

1.2 Boosting

Assume we have a set of training example. We iteratively add weak rules to our model (example: length ¡ 2, etc...). Each feature we add is meant to minimize a loss function which is determined by the weight of our current x_i. Weight increases if our new rules gets it right and decreases if the new rule gets it wrong.

Questions on boosting:

What happens if we get a feature with a high weight?

Boosting is extremely resistant to overfitting as we minimize test error.

Can a rule become useless? We can use decision trees or boosted trees. There are different types of boosting methods, each useful for different situations.

1.3 Summary

This lecture gives an introduction to classification showing various methods like naive bayes and boosting. We derived how naive bayes uses MLE and saw how boosting uses this iterative approach of minimizing a loss function and adding weights. We also got a brief introduction to SVMs.

2 Part II - Code

2.1 Revisiting Part I

How do we deal with discrete and continuous data?
With regression, we were able to have and equations like

$$p(y|x) = wx$$

This wont work here since $p(y|x)$ can't take any value between 0 and 1.

$$p(y = 1|x) = e^{w_1 x}, p(y = 0|x) = e^{w_2 x}$$

These need to sum to 1 so:

$$p(y = 1|x) = \frac{e^{w_1 x}}{e^{w_1 x} + e^{w_2 x}}$$

We can divide through by $e^{w_2 x}$ such that

$$p(y = 1|x) = \frac{e^{w_3 x}}{e^{w_3 x} + 1}$$

We want to get best w after given some data.
Naive Bayes: All weight are independent so you can estimate them separately
Boosting Doesn't assume this so we need to use whole vector

Naive Bayes
Weights are just MLEs so:

$$\log \frac{p}{1-p} = wx \rightarrow p(y|x) = \frac{e^{w_3 x}}{e^{w_3 x} + 1}$$

$$w_1 x_1 + w_2 x_2 + ...$$

Naive Bayes says we can estimate this by just counting. For example, if we are looking at the keyword "money" and whether its spam:

$$\hat{\theta} = \frac{N_{money,spam}}{N_{money}}$$

$$w = \log \frac{\theta_{money}}{1 - \theta_{money}}$$

2.2 Logistic Regression and Boosting

- Have a cost function for wieght
- Do log-likelihood, derivative $\rightarrow$ No closed form sol.
- Use principle of iteratively adding weights minimizing cost function (boosting)

2.3 Code

2.3.1 Shell Script

Grab enron dataset which has a bunch of spam and ham email. Use the spam formula from before for Naive Bayes:

$$p(spam|word) = \frac{p(word|spam)p(spam)}{p(word)}$$

Running on "enron" will give 0 which is not good as you are betting enron will never be spam. Thus we can add pseudocounts:

$$p_{money} = \frac{p(money|spam) + \alpha}{p_{spam} + \beta}$$

We can use cross validation to get best a and b.

2.3.2 R Code

ELSR dataset for spam. Here we can run naive bayes on training data and create a apriori table. We can then predict on test data and get predicted probabilities for every email. Then we can plot distribution or even bin things and plot a calibration plot.

Note: Naive Bayes is over confident since independence of features and multiplying things we are confident about will lead to overconfidence.

Logistic regression can correct for overconfidence by considering correlation between things.

Notes from ka2601

1 Introduction

1.1 What is Classification?

Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

1.1.1 Mathematical Definition

Input: As with regression, in a classification problem we start with measurements x_1, x_2, . . , x_n in an input space X.
Output: The discrete output space Y is composed of K possible classes:
1. $Y = \{$-1, +1$\}$ or $\{0, 1\}$ is called binary classification.
2. $Y = \{1, . . . , K\}$ is called multiclass classification.
Instead of a real-valued response, classification assigns x to a category. For pair (x, y), y is the class of x.

1.1.2 Defining a Classifier

Classification uses a function f (called a classifier) to map input x to class y.

$$y = f(x)$$

2 Naive Bayes Classifier

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes theorem with strong (naive) independence assumptions between the features.

2.1 Assumption

All naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

2.2 Bayes Theorem

A theorem describing how the conditional probability of each of a set of possible causes for a given observed outcome can be computed from knowledge of the probability of each cause and the conditional probability of the outcome of each cause.

Formula:

$$P(\theta|D) = P(\theta)\frac{P(D|\theta)}{P(D)}, \tag{1}$$

2.2.1 Disease Example

Consider a hypothetical population of 10,000 people.

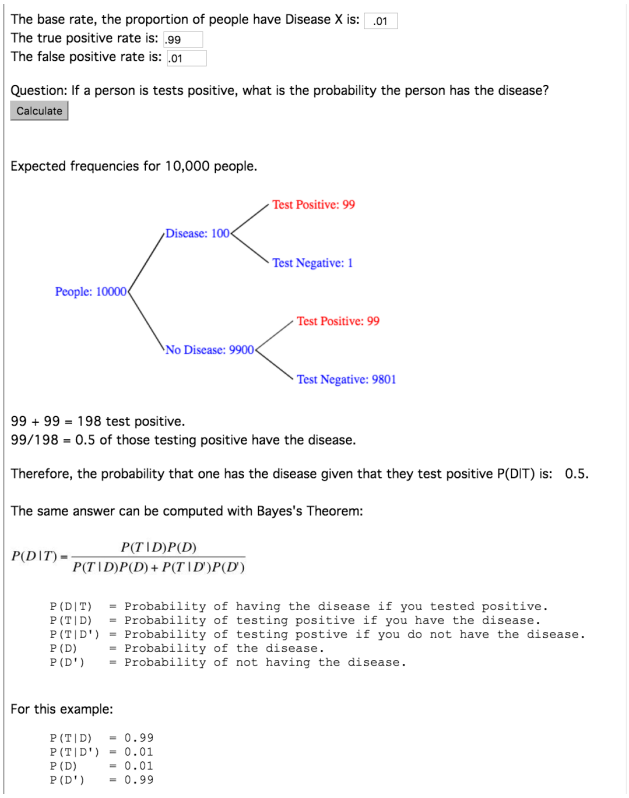

Figure 3: Bayes Theorem Example (Source: http://onlinestatbook.com/2/probability/bayes_demo.html)

2.3 Maximum Likelihood Estimate

The probability of observing the data set in a class C, given parameters θ: (iid assumption)

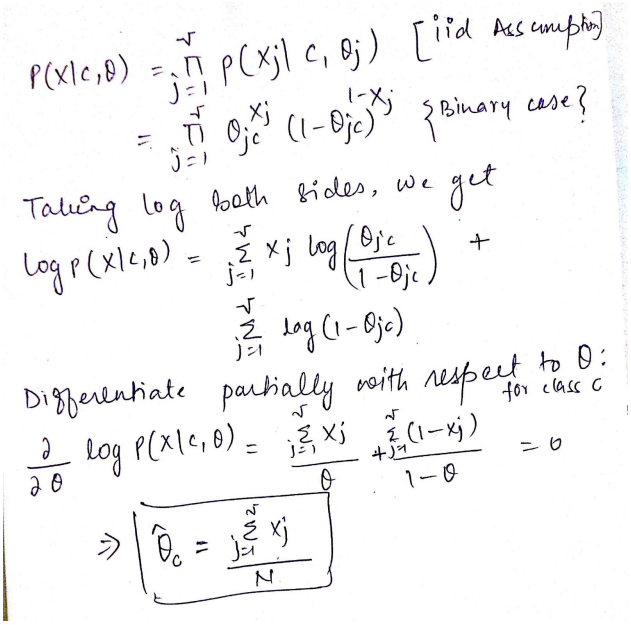

Figure 4: MLE for Naive Bayes

2.4 Advantages & Disadvantages of Naive Bayes

2.4.1 Advantages

- Easy to implement
- Requires a small amount of training data to estimate the parameters
- Good results obtained in most of the cases

2.4.2 Disadvantages

- Assumptions: class conditional independance, therefore loss of accuracy
- Practically, dependencies exist among variables
- Zero conditional probability problem

2.5 Zero conditional probability problem explained

- If a given class and feature value never occur together in the training set then the frequency based probability estimate will be zero.
- This is problematic since it will wipe out all information in the other probabilities when they are multiplied.
- It is therefore often desirable to incorporate a small sample correction in all probability estimates such that no probability is ever set to be exactly zero.
- Laplace smoothing could be the one solution to eliminate this problem.

3 Logistic Regression

Logistic Regression removes the over-fitting by Naive Bayes for zero prior case. It doesn't assume the feature vectors to be uncorrelated.

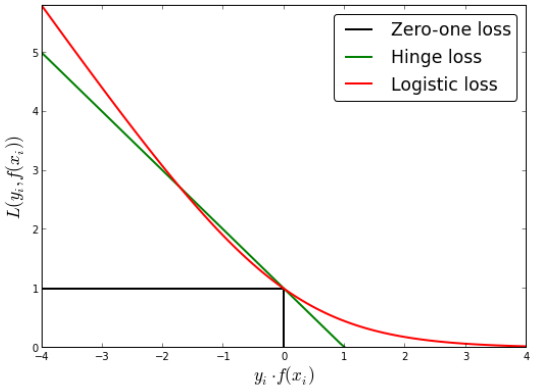

Figure 5: Loss functions

Binary Logistic Regression is a special type of regression where binary response variable is related to a set of explanatory variables, which can be discrete and/or continuous. The important point here to note is that in linear regression, the expected values of the response variable are modeled based on combination of values taken by the predictors. In logistic regression Probability or Odds of the response taking a particular value is modeled based on combination of values taken by the predictors. Like regression (and unlike log-linear models that we will see later), we make an explicit distinction between a response variable and one or more predictor (explanatory) variables.

3.0.1 Log Odds Ratio

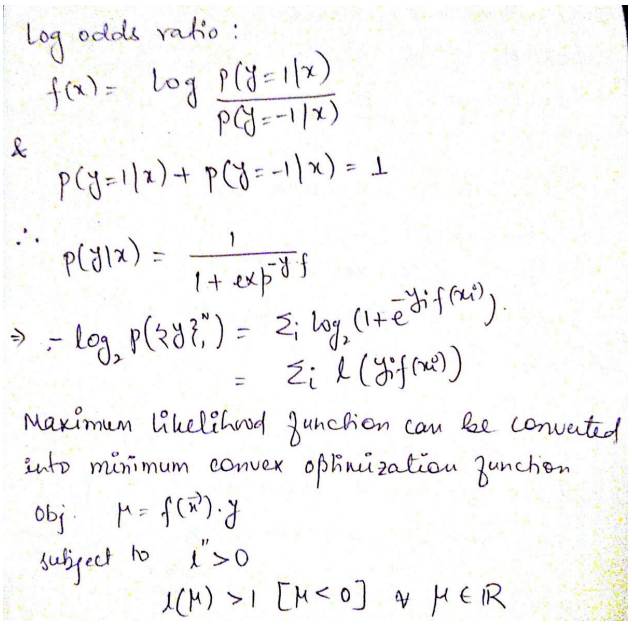

Figure 6: Log Odds Ratio: Logit Function

Despite the probabilistic framework of logistic regression, all that logistic regression assumes is that there is one smooth linear decision boundary. It finds that linear decision boundary by making assumptions that the P(Y—X) of some form, like the inverse logit function applied to a weighted sum of our features. Then it finds the weights by a maximum likelihood approach. The decision boundary it creates is a linear decision boundary that can be of any direction.

3.1 Advantages & Disadvantages of Logistic Regression

3.1.1 Advantages

- Convenient probability scores for observations.
- Multi-collinearity is not really an issue and can be countered with L2 regularization to an extent.

3.1.2 Disadvantages

- Doesn't perform well when feature space is too large.
- Doesn't handle large number of categorical features/variables well.
- Using MLE for parameter might not give closed form solution, therefore use iterative algorithms like Gradient descent (Boosting).

3.2 Boosting

While boosting is not algorithmically constrained, most boosting algorithms consist of iteratively learning weak classifiers with respect to a distribution and adding them to a final strong classifier. When they are added, they are typically weighted in some way that is usually related to the weak learners' accuracy. After a weak learner is added, the data are reweighted: examples that are misclassified gain weight and examples that are classified correctly lose weight.

Notes from vd2334

1 Classification

Classification problem deals with categorization of given examples into a set of categories.

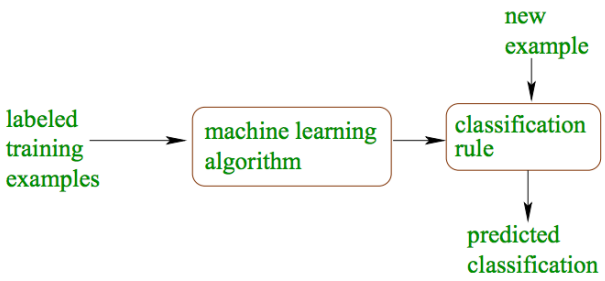

Figure 7: Classification

Input: As with regression, in a classification problem we start with measurements x_1, x_2, . . . , x_n in an input space X.
Output:
1. $Y = \{-1, +1\}$ or $\{0, 1\}$ is called binary classification.
2. $Y = \{1, \ldots, K\}$ is called multiclass classification.
Instead of a real-valued response, classification assigns x to a category. For pair (x, y), y is the class of x.

2 Naive Bayes Classifier

The Naive Bayes classifier is based on the Bayes rule of conditional probability. It makes use of all the attributes contained in the data, and analyses them individually as though they are equally important and independent of each other. In some cases it is also seen that Naive Bayes outperforms many other comparatively complex algorithms.

2.1 Bayes Theorem

A theorem describing how the conditional probability of each of a set of possible causes for a given observed outcome can be computed from knowledge of the probability of each cause and the conditional probability of the outcome of each cause.

Let (Ω,P) be a probability space. (Ω is the sample space; P is the probability distribution.)
For any event A,B in Ω
Formula:

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)}, \tag{2}$$

Let E, H_0, H_1 in Ω.
Conditioned on E, of H_0 and H_1, we find which one is more probable.
Compare $P(H_0) \cdot P(E|H_0)$ to $P(H_1) \cdot P(E|H_1)$

2.1.1 Example

Suppose result of test for genetic disease is correct with probability 95%, and suppose the disease is rare: any given person has disease with probability 1%.

Question: If test comes back positive for disease, is it more likely that you have disease or do not?

$$
\begin{aligned}
E &= \text{test comes back positive for disease} \\
H_0 &= \text{do not have disease} \\
H_1 &= \text{have disease} \\
\mathbb{P}(E \mid H_0) &= 0.05 \\
\mathbb{P}(E \mid H_1) &= 0.95 \\
\mathbb{P}(H_0) &= 0.99 \\
\mathbb{P}(H_1) &= 0.01
\end{aligned}
$$

Want to compare $\mathbb{P}(H_0 \mid E)$ to $\mathbb{P}(H_1 \mid E)$, so compare

$$\mathbb{P}(H_0) \cdot \mathbb{P}(E \mid H_0) = 0.99 \cdot 0.05 \quad \text{and} \quad \mathbb{P}(H_1) \cdot \mathbb{P}(E \mid H_1) = 0.01 \cdot 0.95 .$$

Figure 8: Bayes Theorem Example(Source: http://www.cs.columbia.edu/ djhsu/coms4771-f16/lectures/slides-generative.4up.pdf)

2.2 Maximum Likelihood Estimate

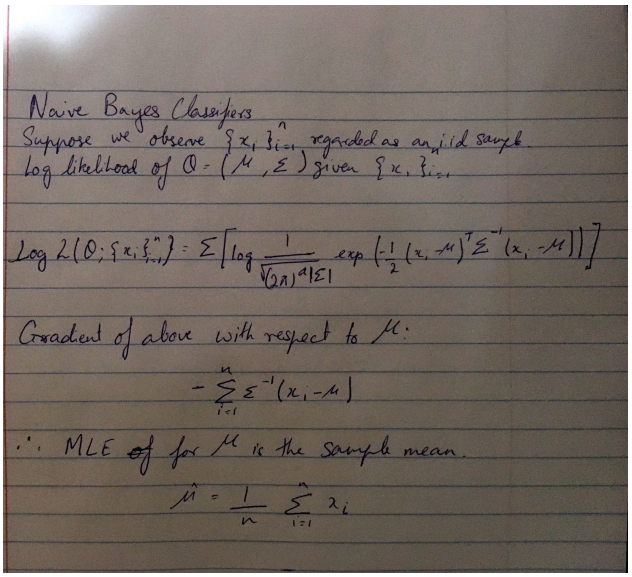

Figure 9: MLE for Naive Bayes

2.2.1 Advantages

- Simple, many variations
- Can leverage domain knowledge of class conditonals
- Can be very efficient when K is large.

2.2.2 Disadvantages

- Classifier relies on formula (via Bayesŕule) that assumes data comes from estimated distribution, which is generally not true.
- Modeling P away from decision boundary between classes is a wasted effort.

3 Logistic Regression

In logistic regression probability of the response taking a particular value is modeled based on combination of values taken by the predictors.

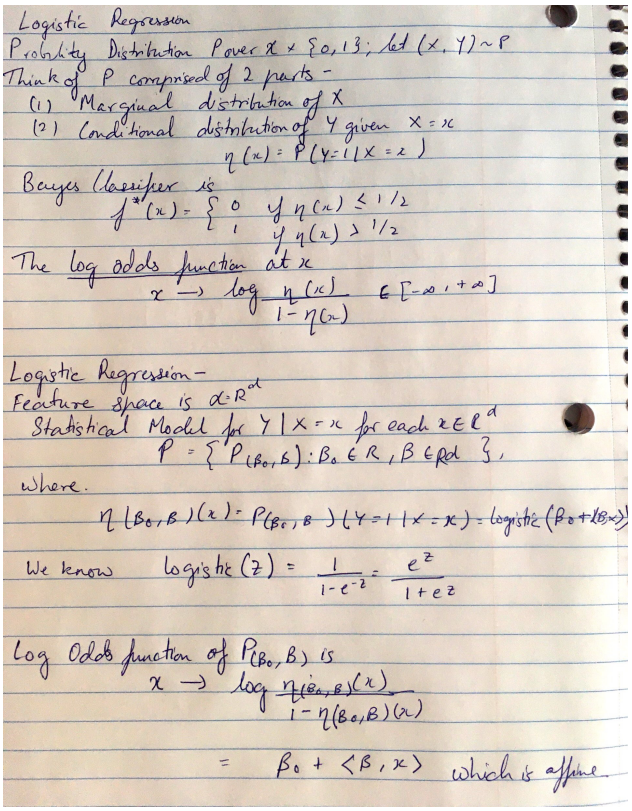

Figure 10: Logistic Regression-Linear Classifier

3.1 Boosting

Boosting is nothing but using a learning algorithm that provides rough rules-of-thumb to construct a very accurate predictor.

Motivation: Easy to construct classification rules that are correct more-often-than-not (e.g., If over 5 percent of the e-mail characters are dollar signs, then its spam.), but seems hard to find a single rule that is almost always correct.

Assumption: Availability of a base or weak learning algorithm which produces a weak classifier. Boosting improves the performance of the weak learning algorithm while treating it as a black box . Weak classifiers are not entirely trivial which means that the error rates are at least a better than a classifier whose every prediction is a random guess. The weak classifiers can be moderately inaccurate, but not as bad as random guessing.

When they are added, they are typically weighted in some way that is usually related to the weak learners' accuracy. After a weak learner is added, the data are assigned weights again. Therefore when an entry is wrongly classified, the classifier which correctly classifies has its weight increased.