Chi Square test for feature selection

July 23, 2016 Author: david

Feature selection is an important problem in Machine learning. There are many feature selection methods available such as mutual information, information gain, and chi square test. In this post, I will use simple examples to describe how to conduct feature selection using chi square test. I will show that it is easy to use Spark or MapReduce to conduct chi square test based feature selection on large scale data set.

Problem Statement

Suppose there are N instances, and two classes: positive and negative. Given a feature X, we can use Chi Square Test to evaluate its importance to distinguish the class.

By calculating the Chi square scores for all the features, we can rank the features by the chi square scores, then choose the top ranked features for model training.

This method can be easily applied for text mining, where terms or words or N-grams are features. After applying chi square test, we can select the top ranked terms as the features to build a text mining model.

Understand Chi Square Test

Chi Square Test is used in statistics to test the independence of two events. Given dataset about two events, we can get the observed count O and the expected count E. Chi Square Score measures how much the expected counts E and observed Count O derivate from each other.

In feature selection, the two events are occurrence of the feature and occurrence

of the class.

In other words, we want to test whether the occurrence of a specific feature and the occurrence of a specific class are independent.

If the two events are dependent, we can use the occurrence of the feature to predict the occurrence of the class. We aim to select the features, of which the occurrence is highly dependent on the occurrence of the class.

When the two events are independent, the observed count is close to the expected count, thus a small chi square score. So a high value of χ^2 indicates that the hypothesis of independence is incorrect. In other words, the higher value of the χ^2 score, the more likelihood the feature is correlated with the class, thus it should be selected for model training.

How to use Chi Square test for feature selection

Suppose we have a set of training instances that belonging to positive and negative classes. To calculate the χ^2 score of a feature X, we can build the following table, in which there are four numbers:

A: the number of positive instances that contain feature X

B: the number of negative instances that contain feature X

C: the number of positive instances that do not contain feature X

D: the number of negative instances that do not contain feature X

	Positive class	Negative class	total
feature X occurs	A	B	A+B = M
feature X not occurs	C	D	C+D = N − M
total	A+C = P	B+D = N − P	N

We also define N, M, P, which are describe blow:

N: denote the total number of instances

M = A + B: the number of instances that contain feature X

C + D = N − M: the number of instances that do not contain feature X

A + C = P: the number of positive instances

B + D = N − P: the number of negative instances

Let A, B, C, D denotes the observed value, and E_A, E_B, E_C, E_D denote the expected value,

How to calculate the expected value

Based on the null hypothesis that the two events are independent, we can calculate the expected value E_A using the following formula:

$$\frac{E_A}{A+C} = \frac{A+B}{N}$$

So

$$E_A = (A+C)\frac{A+B}{N}$$

The basic idea is that if the two events are independent, the probability that feature X occurs in the Positive class instances should be equal to the probability that feature X occurs in all the instances of the two classes.

Using the similar idea, we can calculate E_B, E_C, E_D.

Using Chi Square Test for feature selection

Using the formula of Chi Square test:

$$\chi^2 = \frac{1}{d}\sum_{k=1}^{n}\frac{(O_k - E_k)^2}{E_k}$$

We have

$$\chi^2 = \frac{(A - E_A)^2}{E_A} + \frac{(B - E_B)^2}{E_B} + \frac{(C - E_C)^2}{E_C} + \frac{(D - E_D)^2}{E_D}$$

After some simple calculation we have:

$$\chi^2 = \frac{N(AD - BC)^2}{(A + C)(B + D)(A + B)(C + D)}$$

Given

B = M − A

C = P − A

D = N − M − (P − A)

We have:

$$\chi^2 = \frac{N(AN - MP)^2}{PM(N - P)(N - M))}$$

How to implement the Chi Square Test Algorithm

Given a dataset, we can easily obtain:

A: the total number of positive instances that contain feature X,

M: the total number of instances that contain feature X

P: the total number of positive instance,

N: the total number of instance

Apparently, N, P, are constant for all the features,

For each feature, we only need to count A, and M. It is straightforward to implement the algorithm in python or Java for a small dataset.

If we have a large data set with millions of features, we can also easily implement the algorithm using Spark or MapReduce.

Please refer to this article on using chi square test for feature selection on large scale dataset.