Suppose we have data

| Y | f1 | f2 | f3 |
|---|---|---|---|
| +1 | 1.1 | 10 | 4 |
| +1 | 2 | 20 | 3 |
| +1 | 3.1 | 30 | 2 |
| -1 | 2.1 | 100 | 4.5 |
| -1 | 3 | 200 | 3.5 |
| -1 | 1 | 300 | 1 |

Which of the above three features best discriminates between classes +1 and -1?

If we want to use Pearson correlation we can just evaluate

Pearson(Y, f1), Pearson(Y, f2), and Pearson(Y, f3)

The largest Pearson value will give the most discriminative.

Alternatively we can use the mean and variance of the two classes in each feature to determine discriminative power. For example what is the difference in means between the two classes for a given feature? For f1 it is 0 but for f2 it is 198, which means that feature f2 is a better discriminator than f1.

The chi-square test is similar to Pearson in that it is measuring independence between the variable and label. It is also better applicable to categorical data.

| | f1 |
|---|---|
| +1 | A |
| +1 | A |
| +1 | A |
| -1 | B |
| -1 | B |
| -1 | A |

First we create a contingency table. The contingency is simply counts and comes from the data.

| | A | B |
|---|---|---|
| +1 | 3 (o1,p1) | 0 (o2,p2) |
| -1 | 1 (o3,p3) | 2 (o4,p4) |

n = o1 + o2 + o3 + o4

Each entry in the contingency table is an observed count. We also have a notion of expected counts for each entry of the table. These are expected values under a multinomial distribution.

Since we assume a multinomial distribution each ei = n*pi. But what is pi? Remember that we assume that feature is independent of label and calculate probability under this assumption.

P1 = Prob(A and +1)
    = Prob(A)*Prob(+1)

Prob(A) = (o1+o3)/n
Prob(1) = (o1+o2)/n

E1 = n * (o1+o3)/n *  (o1+o2)/n

In this way we can calculate E2, E3, and E4.

The chi-square value is

Chisquare = ((o1-e1)**2)/e1 + ((o2-e2)**2)/e2 + ((o3-e3)**2)/e3 + ((o4-e4)**2)/e4

We can look up the chi-square p-value using the above formula with degree of freedom = (rows-1)*(cols-1) where rows and cols are number of rows and cols in the contingency table.