Character distribution

Amadou Kountché Djibrilla

December 16, 2015

1 Introduction

The χ^2 test can be use to assess the independence of two random variables or to test the hypothesis that an individual variable is drown from a distribution. In the case of intrusion detection, we are going to use the second case.

2 Definitons

Given v independent variables, each normally distributed with mean u_i and σ_i^2, then :

$$\chi^2 = \sum_{i=1}^{v} (\frac{(x_i - \mu_i)^2}{\sigma_i^2} \tag{1}$$

Ideally, given the random fluctuations of the values of ξ about their mean value μ_i, each term in the sim will be of order of unity, hence if u_i and σ_i are choosen correctly, the χ^2 value will be approximatlly eaqual to v.

If this is the case, it can be concluded that u_i and σ_i describe well the data, the we can not reject the hypothesis.

If χ^2 is greater than v, and we have correctly estimated the value of σ_i, we may possibly conclude that our data are not well described by our hypothesized set of the u_i.

This is the general idea of χ^2 test.

3 The chi^2 distribution

The distribution of the random variable χ^2 is :

$$f(\chi^2) = \frac{1}{2^{v/2}\Gamma(v/2)} e^{-\chi^2/2} (\chi^2)^{(v/2)^{-1}} \tag{2}$$

where:

- v is the degree of freedom
- $\Gamma(p)$ is the gamma function.

The gamma function is defined by :

$$\Gamma(p+1) \equiv \int_0^\infty x^p e^{-x} dx \tag{3}$$

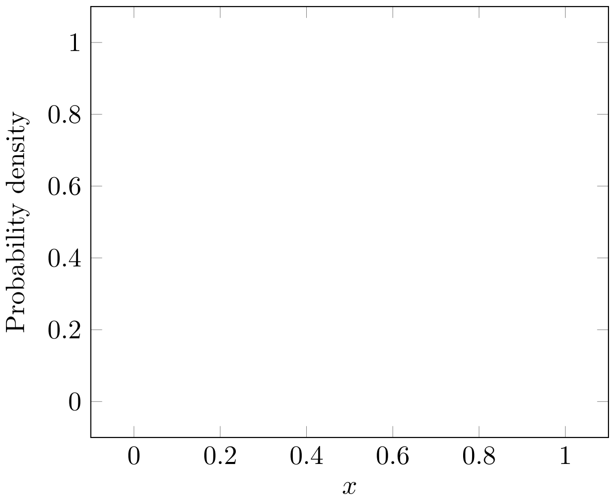

Figure 1: Illustration of the χ^2 distribution

Figure 2: Illustrate here the equation of $\chi^2_{v,\alpha}$.

- It is a generalization of the factorial function to non-integer value of p;
- if p is an, $\Gamma(p+1) = p!$
- in general, $\Gamma(p+1) = p\Gamma(p)$
- $Gamma(1/2) = \sqrt{\pi}$

The χ^2 distribution is skewed for small values and tend toward the normal distribution.

4 Using the χ^2 for statistical test

- Suppose we have N experimental measured quantities x_i,
- we want to known whether ther are well described by some set of hypothesized values μ_i
- Determine the value of χ^2 as described in the equation. In determining the sum, we must use estimates for the σ_i that are independently obtained for each σ_i.

  We can generalise from above discussion, to say that we expect a single measured value of χ^2 will have a aprobaility α of being greater than $\chi^2_{v,\alpha}$ defined by :

$$\int_{\infty} f(\chi^2) d\chi^2 = \alpha \tag{4}$$

The following steps Illustrate how to use the test :

1. We hypothesize that our data are approprially described by our chosen function, or set of μ_i. This is the hypothesis we are going to test.

2. From our data sample, we calculate a sample value of χ^2, along with v, and so determine χ^2/v (the normalized chi-squarre, or chi-square per degree of freedom) for our sample.

3. we choose a value of the significance level α (0.05 is a common value) anf from an appropriate table or graph, determine the corresponding value of $\chi^2_{v,\alpha}/v$. We then compare this value with our sample value of chi^2/v

4. If we find that $\chi^2/v > \chi^2_{v,\alpha}$, we may conclude that either (i) the model represented by the μ_i is a valid one but thant a staistically improbable excursion of χ^2 has occured, or (ii) that our model is so poorly choosen that an unacceptably large value of χ^2 has resulted.

   (i) will happen with a probability α, so if we are satisfied that (i) and (ii) are the only possibilities, (ii) will happen with a probability of $1 - \alpha$.

   Thus, if we find that $\chi^2/v > \chi^2_{v,\alpha}$, we are $100 \times (1 - \alpha)$ per cent confident in *rejecting* our model. Note that this reasoning breaks down if there is a possibily (iii), for example if our data are not normally distributed. The theory of the chi-squarre test relies on the assumption that chi-square is the sum of the squares of random normal deviates, that is, that each x_i is normally distributed about its mean value.

   However for some experiments, there may be occasional non-normal data points that are too far from the mean to be real. It is appropriate to discard data points that are clearly outliers.

5. If we find that χ^2 is too small, that is, if $\chi^2/v < 1 - \chi^2_{v,\alpha}$, we may conclude only that either (i) our model is valid but that a staistically improbable excursion of χ^2 har occured, or (ii) we have, too conservatively, over-estimated the values of σ_i or (iii) someone has given us faudulent data, that is, data 'too good to be true'. A too-small value of χ^2 cannot be indicative of poor model. A poor model can only increase χ^2.