Chi-Squared Test for Machine Learning

Supplementary Lecture

Outline

- What is Chi-square statistics
- Applications in Machine learning
- Example

Chi-square Statistics

- A chi-square statistic is one way to show a relationship between two categorical variables
- The chi-squared statistic is a single number that tells you how much difference exists between your observed counts and the counts you would expect if there were no relationship at all in the population.
- The chi-square distribution can be represented as

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Chi-Square Statistics

- The chi-square distribution can be represented as

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- If there is no difference in observed and expected frequencies, then the chi-square value would be zero. If there is a difference, then the value of chi-square would be more than zero.
- The subscript c represents degree of freedom- the maximum number of independent values

Chi square statistics

- Chi-square test is used to test the ***goodness of fit***
  - determines if a sample data matches a population
- A chi-square **test for independence compares** two variables in a contingency table to see if they are related
  - A very small chi square test statistic means that your observed data fits your expected data extremely well. In other words, there is a relationship.
  - A very large chi square test statistic means that the data does not fit very well. In other words, there isn’t a relationship.

Chi-Square Statistics

- Whether input features are relevant to the outcome to be predicted
- Whether the class (target variable) is dependent or independent on the input features (variable)
    - If independent, we can say the feature is irrelevant in modeling the target class and can be removed **(Feature selection)**

Contingency Table

- Pairs of categorical values can be summarized using **contingency table**
- It shows the distribution of one variable in rows and another in columns
- Example:

Exited\ Gender	Yes	No	Total
Male	38	178	216
Female	44	140	184
Total	82	318	400

Chi square Test

Consider a data-set where we have to determine why customers are leaving the bank,

Let’s perform a Chi-Square test for two variables.

Gender of a customer with values as Male/Female as the predictor and Exited describes whether a customer is leaving the bank with values Yes/No as the response. In this test we will check is there any relationship between Gender and Exited.

Chi square Test

Dataset →

	Gender	……….	Exited
1	Male	……….	Yes
2	Male	……….	Yes
3	Female	……….	No
4	Male	……….	Yes
5	Male	……….	No
6	Male	……….	No
.	.	……….	.
.	.	……….	.
400	Female	……….	No

Chi square Test

Steps to perform the Chi-Square Test:

1. Define Hypothesis.
2. Build a Contingency table.
3. Find the expected values.
4. Calculate the Chi-Square statistic.
5. Accept or Reject the Null Hypothesis

Chi square Test

Define Hypothesis

- Null Hypothesis (H0): Two variables are independent.
- Alternate Hypothesis (H1): Two variables are not independent.

Chi square Test

Contingency Table

- Compute the contingency table

Exited\ Gender	Yes	No	Total
Male	38	178	216
Female	44	140	184
Total	82	318	400

	Gender	....	Exited
1	Male	....	Yes
2	Male	....	Yes
3	Female	....	No
4	Male	....	Yes
5	Male	....	No
6	Male	....	No
.	.	....	.
.	.	....	.
400	Female	....	No

- DoF = (2-1) =1

Chi square Test

Find the Expected Value

- $E_{ij} = (T_i * T_j) / N$
  - T_i = Total in i^{th} row
  - T_j = Total in j^{th} column

- $E_{11} = (216 * 82)/400 = 44.24$
- $E_{12} = (216 * 318)/400 = 171.72$
- $E_{21} = (184 * 82)/400 = 37.72$
- $E_{22} = (184 * 318)/400 = 146.28$

Exited\ Gender	Yes	No	Total
Male	38	178	216
Female	44	140	184
Total	82	318	400

Chi square Test

Exited\Gender	Yes	No	Total
Male	38	178	216
Female	44	140	184
Total	82	318	400

Compute chi-square statistics

	O_i	E_i	$(O_i-E_i)^2/E_i$
1,1	38	44.24	0.880
1,2	178	171.72	0.23
2,1	44	37.72	1.045
2,2	140	146.28	0.27
$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$			2.425

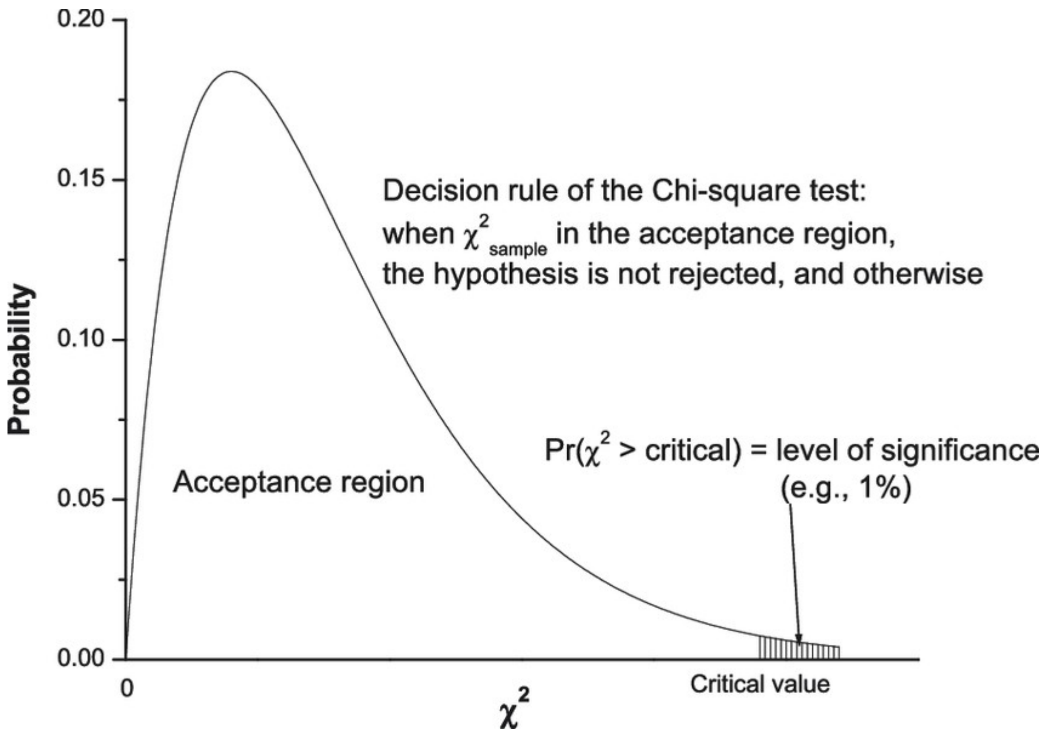
	Yes	NO
Male	44.24	171.72
Female	37.72	146.28

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

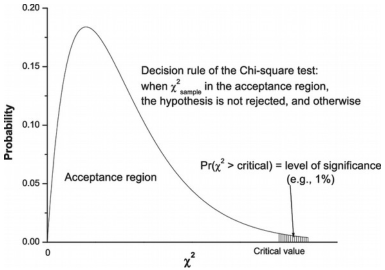
Chi square Test

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Accept or reject the Hypothesis



Chi square Test



$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Accept or reject the Hypothesis

- Compute Chi-square$_{tabulated}$ from Chi-Square table, with level of significance = 0.05 and DoF=1

Table of Chi-square statistics

t-statistics
F-statistics with other P-values: P=0.05 | P=0.01 | P=0.001

df	P = 0.05	P = 0.01	P = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46
7	14.07	18.48	24.32
8	15.51	20.09	26.13
9	16.92	21.67	27.88
10	18.31	23.21	29.59

Chi square Test

Exited\ Gender	Yes	No	Total
Male	38	178	216
Female	44	140	184
Total	82	318	400

Accept or reject the Hypothesis

- Compute Chi-square$_{tabulated}$ from Chi-Square table, with level of significance = 0.05 and DoF=1
- Chi-square$_{tabulated}$ =3. 84
- Chi-square$_{computed}$= 2.425
- Chi-square$_{computed}$< Chi-square$_{tabulated}$

	O_i	E_i	$(O_i-E_i)^2/E_i$
1,1	38	44.24	0.880
1,2	178	171.72	0.23
2,1	44	37.72	1.045
2,2	140	146.28	0.27
		Chi square	2.425

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Chi square Test

Exited\ Gender	Yes	No	Total
Male	38	178	216
Female	44	140	184
Total	82	318	400

	O_i	E_i	$(O_i-E_i)^2/E_i$
1,1	38	44.24	0.880
1,2	178	171.72	0.23
2,1	44	37.72	1.045
2,2	140	146.28	0.27
		Chi square	2.425

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

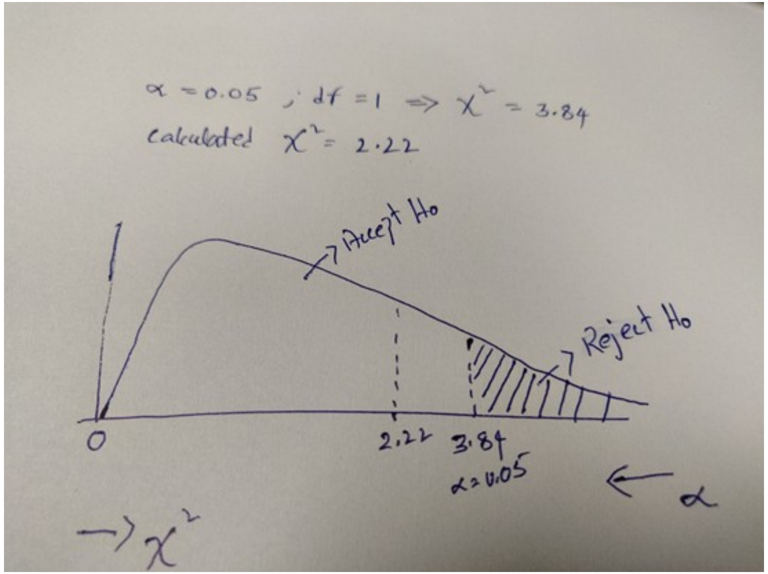
Accept or reject the Hypothesis

- Accepting the null hypothesis since the Chi-square$_{\text{computed}}$ < Chi-square$_{\text{tabulated}}$
- That means, the variables are independent
- Gender variable cannot be selected for training the model.

Chi square Test

Accept or reject the Hypothesis

- Accepting the null hypothesis since the Chi-square$_{computed}$ < Chi-square$_{tabulated}$
- That means, the variables are independent
- Gender variable cannot be selected for training the model.



References

https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223

https://machinelearningmastery.com/chi-squared-test-for-machine-learning/