Birla Institute of Applied Sciences
बिरला इंस्टिट्यूट ऑफ़ अप्लाइड साइंसेस
Bhimtal, Distt: Nainital, Uttarakhand- 263136

Pattern Recognition

- S. S. Samant

Feature Extraction and Feature Selection

- Importance:
  - Using discriminating features enhances performance
  - Avoids unnecessary computation
  - With less features, a smaller dataset is needed
  - Intended similarity can be captured by distances in a smaller set of features

Types of Feature Selection

- Filter methods: scores each feature
- Wrapper methods: scores subsets of features on a validation set
- Embedded methods: selects features during training itself

Problem: Mutual Information

$$MI = \frac{N_{u_t u_l}}{N} \log_2 \frac{NN_{u_t u_l}}{(N_{u_t \bar{u}_l} + N_{u_t u_l})(N_{u_t u_l} + N_{\bar{u}_t u_l})}$$
$$+ \frac{N_{\bar{u}_t u_l}}{N} \log_2 \frac{NN_{\bar{u}_t u_l}}{((N_{\bar{u}_t u_l} + N_{\bar{u}_t \bar{u}_l})(N_{u_t u_l} + N_{\bar{u}_t u_l})}$$
$$+ \frac{NN_{u_t \bar{u}_l}}{N} \log_2 \frac{NN_{u_t \bar{u}_l}}{(N_{u_t u_l} + N_{u_t \bar{u}_l})(N_{u_t \bar{u}_l} + N_{\bar{u}_t \bar{u}_l})}$$
$$+ \frac{N_{\bar{u}_t \bar{u}_l}}{N} \log_2 \frac{NN_{\bar{u}_t \bar{u}_l}}{(N_{\bar{u}_t u_l} + N_{\bar{u}_t \bar{u}_l})(N_{u_t \bar{u}_l} + N_{\bar{u}_t \bar{u}_l})},$$

Problem: In Reuters corpus, if term is *export* and class is *poultry*. Let's call non-*export* term as *other* term and non-*poultry* class as other class.
The term *export* is present in 49 documents of class poultry and in 27652 documents of other class. There are 141 other terms in documents of poultry class and 774106 other terms in documents of other classes.
Compute MI.

Chi-square Statistic for Feature Selection

- Used to determine if a distribution of observed frequencies differs from the theoretical expected frequencies.
- Used to determine if two variables are independent

Chi-square Statistic for Feature Selection

- Used to determine if a distribution of observed frequencies differs from the theoretical expected frequencies.
- Used to determine if two variables are independent

$$\chi^2 = \sum_i |(N_i - E_i)^2 / E_i|$$

Chi-square Statistic for Feature Selection

- Used to determine if a distribution of observed frequencies differs from the theoretical expected frequencies.
- Used to determine if two variables are independent

$$\chi^2 = \sum_i |(N_i - E_i)^2 / E_i|$$

- NULL Hypothesis: The term and class are independent

Chi-square Statistic for Feature Selection

- Used to determine if a distribution of observed frequencies differs from the theoretical expected frequencies.
- Used to determine if two variables are independent

$$\chi^2 = \sum_i |(N_i - E_i)^2 / E_i|$$

- In feature selection the independence between a term and a class:

$$\chi^2(D,t,l) = \frac{(N_{u_t u_l} - E_{u_t u_l})^2}{E_{u_t u_l}} + \frac{(N_{\bar{u_t} u_l} - E_{\bar{u_t} u_l})^2}{E_{\bar{u_t} u_l}} + \frac{(N_{u_t \bar{u_l}} - E_{u_t \bar{u_l}})^2}{E_{u_t \bar{u_l}}} + \frac{(N_{\bar{u_t} \bar{u_l}} - E_{\bar{u_t} \bar{u_l}})^2}{E_{\bar{u_t} \bar{u_l}}}$$

where
u_t means that the document contains the term t, and
$\bar{u_t}$ means the document does not contain the term t;
u_l means the document is in class l and,
$\bar{u_l}$ means the document is not in class l;
N = observed frequency and,
E = expected frequency.

Chi-square Statistic for Feature Selection

- Used to determine if a distribution of observed frequencies differs from the theoretical expected frequencies.
- Used to determine if two variables are independent

$$\chi^2 = \sum_i |(N_i - E_i)^2 / E_i|$$

- In feature selection the independence between a term and a class:

$$\chi^2(D,t,l) = \frac{(N_{u_t u_l} - E_{u_t u_l})^2}{E_{u_t u_l}} + \frac{(N_{\bar{u}_t u_l} - E_{\bar{u}_t u_l})^2}{E_{\bar{u}_t u_l}} + \frac{(N_{u_t \bar{u}_l} - E_{u_t \bar{u}_l})^2}{E_{u_t \bar{u}_l}} + \frac{(N_{\bar{u}_t \bar{u}_l} - E_{\bar{u}_t \bar{u}_l})^2}{E_{\bar{u}_t \bar{u}_l}}$$

where
u_t means that the document contains the term t, and
$\bar{u}_t$ means the document does not contain the term t;
u_l means the document is in class l and,
$\bar{u}_l$ means the document is not in class l;
N = observed frequency and,
E = expected frequency.

If the value of Chi-square is greater than the value in the Chi-square distribution table, we reject the null hypothesis that the term and class are independent

Link for Chi-square test of independence

https://www.spss-tutorials.com/chi-square-independence-test/

Thank You!