Information Theory

CS5154/6054

Yizong Cheng

10/13/2022

13 *Text classification and Naive Bayes*

13.5 Feature selection

13.5.1 Mutual information

13.5.2 χ^2 Feature selection

Odds Ratio and Retrieval Status Value

(11.18) $$c_t = \log \frac{p_t(1-u_t)}{u_t(1-p_t)}$$

(11.19)

	documents	relevant	nonrelevant	Total
Term present	$x_t = 1$	s	$\mathrm{df}_t - s$	df_t
Term absent	$x_t = 0$	$S - s$	$(N - \mathrm{df}_t) - (S - s)$	$N - \mathrm{df}_t$
	Total	S	$N - S$	N

Using this, $p_t = s/S$ and $u_t = (\mathrm{df}_t - s)/(N - S)$ and

(11.20) $$c_t = K(N, \mathrm{df}_t, S, s) = \log \frac{s/(S-s)}{(\mathrm{df}_t - s)/((N - \mathrm{df}_t) - (S - s))}$$

When a Term Has Nothing to Do with Relevance

Table 2	R=1 (Relevant)	R=0 (on-relevant)	
Term t present	N11	N10	
Term t absent	N01	N00	

(11.19)

	documents	relevant	nonrelevant	Total
Term present	$x_t = 1$	s	$\mathrm{df}_t - s$	df_t
Term absent	$x_t = 0$	$S - s$	$(N - \mathrm{df}_t) - (S - s)$	$N - \mathrm{df}_t$
	Total	S	$N - S$	N

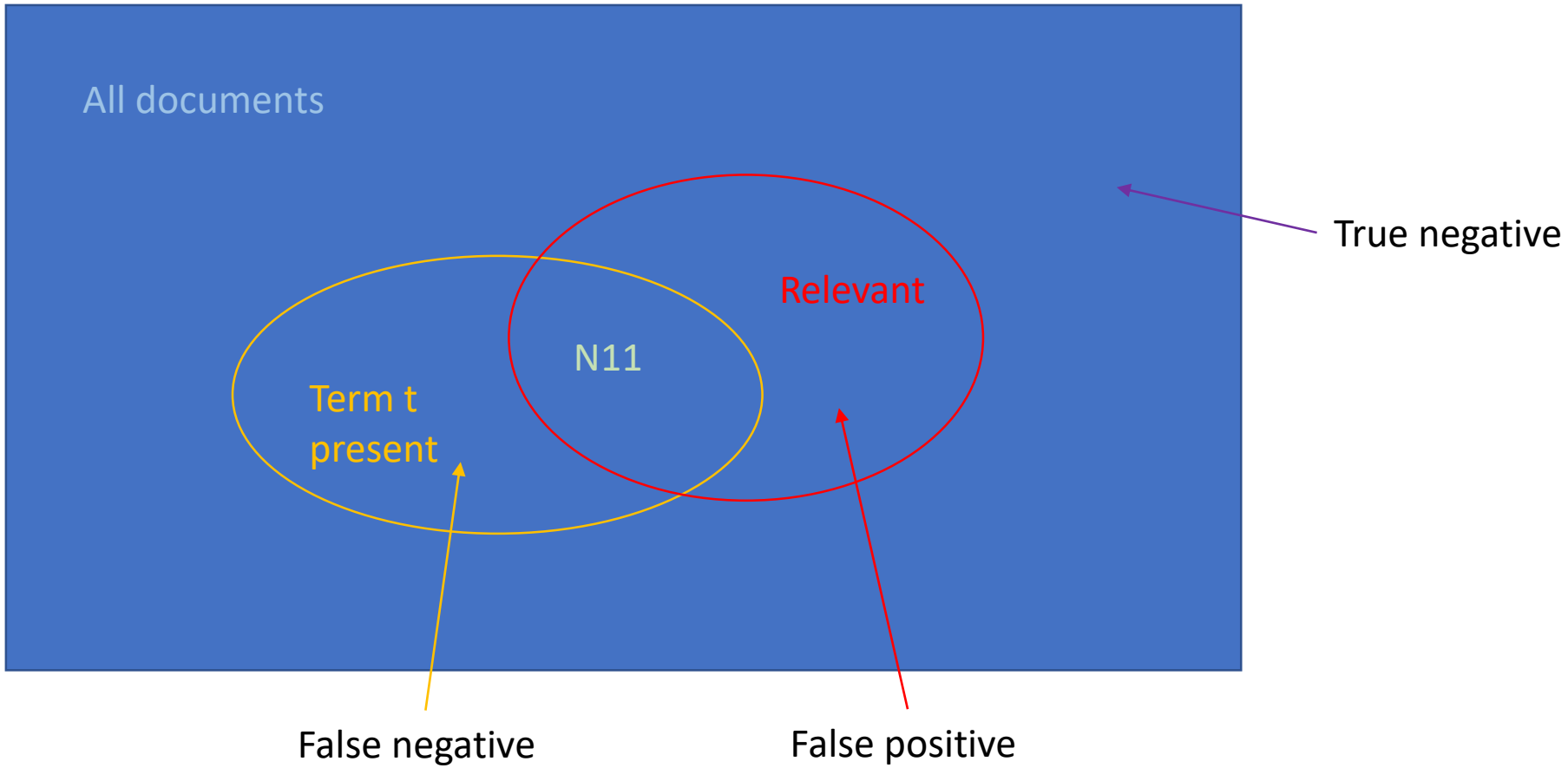
Using this, $p_t = s/S$ and $u_t = (\mathrm{df}_t - s)/(N - S)$ and

(11.20) $$c_t = K(N, \mathrm{df}_t, S, s) = \log \frac{s/(S-s)}{(\mathrm{df}_t - s)/((N - \mathrm{df}_t) - (S - s))}$$

C_t = log (N11/N01)/(N10/N00)

C_t = 0 iff N11 N00 = N01 N10

Two Ways to Partition (Cluster) Documents



Expected Counts for Independent Events

Actual	R=1 (Relevant)	R=0 (on-relevant)	
Term t present	N11	N10	N1.
Term t absent	N01	N00	N0.
	N.1	N.0	N

S of N documents are relevant and df_t of N documents contains term t. The expected number of document both relevant and containing t is $S\ df_t /N = (N11 + N01)(N11 + N10)/N$. If this expected number is also the actual N11, then events t present and relevance are independent.

Quiz problem: under what condition do we have $(N11 + N01)(N11 + N10)/N = (N11 + N01)(N11 + N10)/(N11 + N01 + N10 + N00) = N11$?

Expected Counts from Observed Counts

Example 13.3: Consider the class *poultry* and the term export in Reuters-RCV1. The counts of the number of documents with the four possible combinations of indicator values are as follows:

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$
$e_t = e_{export} = 1$	$N_{11} = 49$	$N_{10} = 27{,}652$
$e_t = e_{export} = 0$	$N_{01} = 141$	$N_{00} = 774{,}106$

Example 13.4: We first compute E_{11} for the data in Example 13.3:

$$\begin{aligned} E_{11} &= N \times P(t) \times P(c) = N \times \frac{N_{11} + N_{10}}{N} \times \frac{N_{11} + N_{01}}{N} \\ &= N \times \frac{49 + 141}{N} \times \frac{49 + 27652}{N} \approx 6.6 \end{aligned}$$

where N is the total number of documents as before.

We compute the other $E_{e_t e_c}$ in the same way:

	$e_{poultry} = 1$		$e_{poultry} = 0$	
$e_{export} = 1$	$N_{11} = 49$	$E_{11} \approx 6.6$	$N_{10} = 27{,}652$	$E_{10} \approx 27{,}694.4$
$e_{export} = 0$	$N_{01} = 141$	$E_{01} \approx 183.4$	$N_{00} = 774{,}106$	$E_{00} \approx 774{,}063.6$

X^2 : Difference between Expected and Observed

$$X^2(\mathbb{D}, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} \tag{13.18}$$

STATISTICAL SIGNIFICANCE

X^2 is a measure of how much expected counts E and observed counts N deviate from each other. A high value of X^2 indicates that the hypothesis of independence, which implies that expected and observed counts are similar, is incorrect. In our example, $X^2 \approx 284 > 10.83$. Based on Table 13.6, we can reject the hypothesis that *poultry* and export are independent with only a 0.001 chance of being wrong.[8] Equivalently, we say that the outcome $X^2 \approx 284 > 10.83$ is *statistically significant* at the 0.001 level. If the two events are dependent, then the occurrence of the term makes the occurrence of the class more likely (or less likely), so it should be helpful as a feature. This is the rationale of χ^2 feature selection.

An arithmetically simpler way of computing X^2 is the following:

$$X^2(\mathbb{D}, t, c) = \frac{(N_{11} + N_{10} + N_{01} + N_{00}) \times (N_{11}N_{00} - N_{10}N_{01})^2}{(N_{11} + N_{01}) \times (N_{11} + N_{10}) \times (N_{10} + N_{00}) \times (N_{01} + N_{00})} \tag{13.19}$$

This is equivalent to Equation (13.18) (Exercise 13.14).

Statistical Significance (p-Value)

► **Table 13.6** Critical values of the χ^2 distribution with one degree of freedom. For example, if the two events are independent, then $P(X^2 > 6.63) < 0.01$. So for $X^2 > 6.63$ the assumption of independence can be rejected with 99% confidence.

p	χ^2 critical value
0.1	2.71
0.05	3.84
0.01	6.63
0.005	7.88
0.001	10.83

sklearn.feature_selection.chi2

sklearn.feature_selection.chi2(*X*, *y*) [source]

Compute chi-squared stats between each non-negative feature and class.

This score can be used to select the n_features features with the highest values for the test chi-squared statistic from X, which must contain only non-negative features such as booleans or frequencies (e.g., term counts in document classification), relative to the classes.

Recall that the chi-square test measures dependence between stochastic variables, so using this function "weeds out" the features that are the most likely to be independent of class and therefore irrelevant for classification.

Read more in the User Guide.

Parameters:

X : *{array-like, sparse matrix} of shape (n_samples, n_features)*
Sample vectors.

y : *array-like of shape (n_samples,)*
Target vector (class labels).

Returns:

chi2 : *ndarray of shape (n_features,)*
Chi2 statistics for each feature.

p_values : *ndarray of shape (n_features,)*
P-values for each feature.

From Counts to Probability Distribution

Actual	R=1 (Relevant)	R=0 (on-relevant)	
Term t present	N11/N	N10/N	N1./N
Term t absent	N01/N	N00/N	N0./N
	N.1/N	N.0/N	N/N = 1

$$N = N11 + N01 + N10 + N00$$

Independent	R=1 (Relevant)	R=0 (on-relevant)	
Term t present	$N1.N.1/N^2$	$N1.N.0/N^2$	N1./N
Term t absent	$N0.N.1/N^2$	$N0.N.0/N^2$	N0./N
	N.1/N	N.0/N	N/N = 1

Kullback-Leibler Divergence

- A measure of difference between two probability distributions (over the same set of outcomes).
- $KL(p \parallel q) = \sum_k p_k \log (p_k/q_k)$
- Theorem: $KL(p \parallel q) >= 0$ with equality iff $p = q$.

Mutual Information

- Given two random variables (X = term t present and Y = relevant),
- I(X; Y) = KL(p(X, Y) || p(X)p(Y))
- In our case, both random variables are binary and the estimates of the probabilities are based on counts N11, N01, N10, and N00.
- Iir uses U as the random variable for term t present and C as the random variable for relevance.

$$I(U;C) = \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0.}N_{.0}} \quad (13.17)$$

Example 13.3: Consider the class *poultry* and the term export in Reuters-RCV1. The counts of the number of documents with the four possible combinations of indicator values are as follows:

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$
$e_t = e_{export} = 1$	$N_{11} = 49$	$N_{10} = 27{,}652$
$e_t = e_{export} = 0$	$N_{01} = 141$	$N_{00} = 774{,}106$

After plugging these values into Equation (13.17) we get:

$$\begin{aligned} I(U;C) &= \frac{49}{801{,}948} \log_2 \frac{801{,}948 \cdot 49}{(49+27{,}652)(49+141)} \\ &\quad + \frac{141}{801{,}948} \log_2 \frac{801{,}948 \cdot 141}{(141+774{,}106)(49+141)} \\ &\quad + \frac{27{,}652}{801{,}948} \log_2 \frac{801{,}948 \cdot 27{,}652}{(49+27{,}652)(27{,}652+774{,}106)} \\ &\quad + \frac{774{,}106}{801{,}948} \log_2 \frac{801{,}948 \cdot 774{,}106}{(141+774{,}106)(27{,}652+774{,}106)} \\ &\approx 0.0001105 \end{aligned}$$

UK

london	0.1925
uk	0.0755
british	0.0596
stg	0.0555
britain	0.0469
plc	0.0357
england	0.0238
pence	0.0212
pounds	0.0149
english	0.0126

China

china	0.0997
chinese	0.0523
beijing	0.0444
yuan	0.0344
shanghai	0.0292
hong	0.0198
kong	0.0195
xinhua	0.0155
province	0.0117
taiwan	0.0108

poultry

poultry	0.0013
meat	0.0008
chicken	0.0006
agriculture	0.0005
avian	0.0004
broiler	0.0003
veterinary	0.0003
birds	0.0003
inspection	0.0003
pathogenic	0.0003

coffee

coffee	0.0111
bags	0.0042
growers	0.0025
kg	0.0019
colombia	0.0018
brazil	0.0016
export	0.0014
exporters	0.0013
exports	0.0013
crop	0.0012

elections

election	0.0519
elections	0.0342
polls	0.0339
voters	0.0315
party	0.0303
vote	0.0299
poll	0.0225
candidate	0.0202
campaign	0.0202
democratic	0.0198

sports

soccer	0.0681
cup	0.0515
match	0.0441
matches	0.0408
played	0.0388
league	0.0386
beat	0.0301
game	0.0299
games	0.0284
team	0.0264

► **Figure 13.7** Features with high mutual information scores for six Reuters-RCV1 classes.

sklearn.metrics.mutual_info_score

sklearn.metrics.**mutual_info_score**(*labels_true*, *labels_pred*, *, *contingency=None*) [source]

Mutual Information between two clusterings.

The Mutual Information is a measure of the similarity between two labels of the same data. Where $|U_i|$ is the number of the samples in cluster U_i and $|V_j|$ is the number of the samples in cluster V_j, the Mutual Information between clusterings U and V is given as:

$$MI(U,V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|}\frac{|U_i \cap V_j|}{N}\log\frac{N|U_i \cap V_j|}{|U_i||V_j|}$$

Any Two Sets Have a Mutual Information

- The set of retrieved and the set of relevant documents.
- The set of documents containing term t1 and that containing term t2.
  - A distance measure between terms?
  - I(X; Y) may not be the same as I(Y; X).
- The set of terms in document d1 and that in document d2.
  - Divergence between documents?
- Objective for optimization:
  - Modifying a set of parameters so that the mutual information generated with these parameters against a target distribution is minimized.