Feature selection

LING 572
Fei Xia
Week 4: 1/26/2010

Creating attribute-value table

	f_1	f_2	…	f_K	y
x_1					
x_2					
…					

- Choose features:
  - Define feature templates
  - Instantiate the feature templates
  - Dimensionality reduction: feature selection

- Feature weighting
  - The weight for f_k: the whole column
  - The weight for f_k in d_i: a cell

An example: text classification task

- Define feature templates:
  - One template only: word
- Instantiate the feature templates
  - All the words appeared in the training (and test) data
- Dimensionality reduction: feature selection
  - Remove stop words
- Feature weighting
  - Feature value: term frequency (tf), or tf-idf

Outline

- Dimensionality reduction
- Some scoring functions **
- Chi-square score and Chi-square test
- Hw4

In this lecture, we will use “term” and “feature” interchangeably.

Dimensionality reduction (DR)

Dimensionality reduction (DR)

- What is DR?
  - Given a feature set r, create a new set r', s.t.
    - r' is much smaller than r, and
    - the classification performance does not suffer too much.

- Why DR?
  - ML algorithms do not scale well.
  - DR can reduce overfitting.

Types of DR

- r is the original feature set, r' is the one after DR.

- Local DR vs. Global DR
  - Global DR: r' is the same for every category
  - Local DR: a different r' for each category

- Term extraction vs. term selection

Term selection vs. extraction

- Term selection: r’ is a subset of r
  - Wrapping methods: score terms by training and evaluating classifiers.

    ➔ expensive and classifier-dependent

  - Filtering methods

- Term extraction: terms in r’ are obtained by combinations or transformation of r terms.
  - Term clustering:
  - Latent semantic indexing (LSI)

Term selection by filtering

- Main idea: scoring terms according to predetermined numerical functions that measure the “importance” of the terms.
- It is fast and classifier-independent.
- Scoring functions:
  - Information Gain
  - Mutual information
  - chi square
  - …

Quick summary so far

- DR: to reduce the number of features
  - Local DR vs. global DR
  - Term extraction vs. term selection

- Term extraction
  - Term clustering:
  - Latent semantic indexing (LSI)

- Term selection
    - Wrapping method
    - Filtering method: different functions

Some scoring functions

Basic distributions (treating features as binary)

Probability distributions on the event space of documents:

$P(t_k)$: The % of docs where t_k occurs

$P(\bar{t_k})$, $P(c_i)$, $P(\bar{c_i})$

$$P(t_k, c_i),\ P(t_k, \bar{c_i}),\ P(\bar{t_k}, c_i),\ P(\bar{t_k}, \bar{c_i}).$$

$$P(t_k | c_i),\ P(t_k | \bar{c_i}),\ P(\bar{t_k} | c_i),\ P(\bar{t_k} | \bar{c_i}).$$

Calculating basic distributions

	$\bar{c_i}$	c_i
$\bar{t_k}$	a	b
t_k	c	d

$$P(t_k, c_i) = d/N$$

$$P(t_k) = (c + d)/N, P(c_i) = (b + d)/N$$

$$P(t_k|c_i) = d/(b + d)$$

where $N = a + b + c + d$

Term selection functions

- Intuition: for a category c_i , the most valuable terms are those that are distributed most <u>differently</u> in the sets of possible and negative examples of c_i.

Term selection functions

Document frequency:

the num of docs in which t_k occurs

Pointwise mutual information:

$$MI(t_k, c_i) = log\frac{P(t_k,c_i)}{P(c_i)P(t_k)}$$

Information gain: $IG(t_k, c_i) =$

$$P(t_k, c_i)log\frac{P(t_k,c_i)}{P(c_i)P(t_k)} + P(\bar{t_k}, c_i)log\frac{P(\bar{t_k},c_i)}{P(c_i)P(\bar{t_k})}$$

Information gain

- IG(Y|X): We must transmit Y. How many bits on average would it save us if both ends of the line knew X?

- Definition:

  $IG(Y, X) = H(Y) - H(Y|X)$

Information gain**

$$\boxed{\sum_i IG(t_k, c_i)}$$

$$= \sum_{c \in C} \sum_{t \in \{t_k, \bar{t_k}\}} P(t,c) log \frac{P(t,c)}{P(c)P(t)}$$

$$= \sum_{c \in C} \sum_t P(t,c) log P(c|t)$$

$$- \sum_c \sum_t P(t,c) log P(c)$$

$$= -H(C|T) - \sum_c ((log P(c)) \sum_t P(t,c))$$

$$= -H(C|T) + H(C) = \boxed{IG(C|T)}$$

More term selection functions**

GSS coefficient:

$$GSS(t_k, c_i) = P(t_k, c_i)P(\bar{t_k}, \bar{c_i}) - P(t_k, \bar{c_i})P(\bar{t_k}, c_i)$$

NGL coefficient: N is the total number of docs

$$NGL(t_k, c_i) = \frac{\sqrt{N}\ GSS(t_k, c_i)}{\sqrt{P(t_k)P(\bar{t_k})P(c_i)P(\bar{c_i})}}$$

Chi-square: (one of the definitions)

$$\chi^2(t_k, c_i) = NGL(t_k, c_i)^2 = \frac{(ad-bc)^2 N}{(a+b)(a+c)(b+d)(c+d)}$$

More term selection functions**

Relevancy score:

$$RS(t_k, c_i) = log\frac{P(t_k|c_i)+d}{P(\bar{t_k}|\bar{c_i})+d}$$

Odds Ratio:

$$OR(t_k, c_i) = \frac{P(t_k|c_i)P(\bar{t_k}|\bar{c_i})}{P(\bar{t_k}|c_i)P(t_k|\bar{c_i})}$$

Global DR

- For local DR, calculate $f(t_k, c_i)$.

- For global DR, calculate one of the following:

Sum: $f_{sum}(t_k) = \sum_{i=1}^{|C|} f(t_k, c_i)$

Average: $f_{avg}(t_k) = \sum_{i=1}^{|C|} f(t_k, c_i) P(c_i)$

Max: $f_{max}(t_k) = \max_{i=1}^{|C|} f(t_k, c_i)$

$|C|$ is the number of classes

Which function works the best?

- It depends on
  - Classifiers
  - Data
  - …

- According to (Yang and Pedersen 1997):

$$\{OR, NGL, GSS\} > \{\chi^2_{max}, IG_{sum}\}$$
$$> \{\#_{avg}\} >> \{MI\}$$

Feature weighting

Alternative feature values

- Binary features: 0 or 1.
- Term frequency (TF): the number of times that t_k appears in d_i.
- Inversed document frequency (IDF): log |D| /$d_{k,}$ where d_k is the number of documents that contain t_k.
- TFIDF = TF * IDF
- Normalized TFIDF: $$w_{ik} = \frac{tfidf(d_i, t_k)}{Z}$$

Feature weights

- Feature weight $\in$ {0,1}: same as DR

- Feature weight $\in$ R: iterative approach:
  - Ex: MaxEnt

➔ Feature selection is a special case of feature weighting.

Summary so far

- Curse of dimensionality ➔ dimensionality reduction (DR)

- DR:
  - Term extraction
  - Term selection
    - Wrapping method
    - <u>Filtering method</u>: different functions

Summary (cont)

- Functions:
  - Document frequency
  - Mutual information
  - Information gain
  - Gain ratio
  - Chi square
  - …

Chi square

Chi square

- An example: is gender a good feature for predicting footwear preference?
  - A: gender
  - B: footwear preference

- Bivariate tabular analysis:
  - Is there a relationship between two random variables A and B in the data?
  - How strong is the relationship?
  - What is the direction of the relationship?

Raw frequencies

	sandal	sneaker	Leather shoe	boots	others
male	6	17	13	9	5
female	13	5	7	16	9

Feature: male/female

Classes: {sandal, sneaker, ….}

Two distributions

Observed distribution (O):

	Sandal	Sneaker	Leather	Boot	Others
Male	6	17	13	9	5
Female	13	5	7	16	9

Expected distribution (E):

	Sandal	Sneaker	Leather	Boot	Others	Total
Male						50
Female						50
Total	19	22	20	25	14	100

Two distributions

Observed distribution (O):

	Sandal	Sneaker	Leather	Boot	Others	Total
Male	6	17	13	9	5	50
Female	13	5	7	16	9	50
Total	19	22	20	25	14	100

Expected distribution (E):

	Sandal	Sneaker	Leather	Boot	Others	Total
Male	9.5	11	10	12.5	7	50
Female	9.5	11	10	12.5	7	50
Total	19	22	20	25	14	100

Chi square

- Expected value =

  row total * column total / table total

- $\chi^2 = \sum_{ij} (O_{ij} - E_{ij})^2 / E_{ij}$

- $\chi^2 = (6-9.5)^2/9.5 + (17-11)^2/11+ \ldots.$

  $= 14.026$

Calculating χ^2

- Fill out a contingency table of the observed values ➔ O
- Compute the row totals and column totals
- Calculate expected value for each cell assuming no association ➔ E
- Compute chi square: $(O-E)^2/E$

When r=2 and c=2

O =

	$\bar{c_i}$	c_i	total
$\bar{t_k}$	a	b	a+b
t_k	c	d	c+d
total	a+c	b+d	N

E =

	$\bar{c_i}$	c_i	total
$\bar{t_k}$	$\frac{(a+c)(a+b)}{N}$	$\frac{(b+d)(a+b)}{N}$	a+b
t_k	$\frac{(a+c)(c+d)}{N}$	$\frac{(b+d)(c+d)}{N}$	c+d
total	a+c	b+d	N

$$\chi^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} = \frac{(ad-bc)^2 N}{(a+b)(a+c)(b+d)(c+d)}$$

χ^2 test

Basic idea

- Null hypothesis (the tested hypothesis): no relation exists between two random variables.

- Calculate the probability of having the observation with that χ^2 value, assuming the hypothesis is true.

- If the probability is too small, reject the hypothesis.

Requirements

- The events are assumed to be independent and have the same distribution.
- The outcomes of each event must be mutually exclusive.
- At least 5 observations per cell.
- Collect raw frequencies, not percentages

Degree of freedom

- Degree of freedom df = (r – 1) (c – 1)

  r: # of rows c: # of columns

- In this Ex: df=(2-1) (5-1)=4

χ^2 distribution table

	0.10	0.05	0.025	0.01	0.001
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
…					

df=4 and 14.026 > 13.277

➔$p<0.01$

➔there is a significant relation

χ^2 to P Calculator

http://faculty.vassar.edu/lowry/tabs.html#csq

Steps of χ^2 test

- Select significance level p_0
- Calculate χ^2
- Compute the degree of freedom
  df = (r-1)(c-1)
- Calculate p given χ^2 value (or get the χ^2_0 for p_0)
- if $p < p_0$ (or if $\chi^2 > \chi^2_0$)
  then reject the null hypothesis.

Summary of χ^2 test

- A very common method for significant test

- Many good tutorials online
  - Ex: http://en.wikipedia.org/wiki/Chi-square_distribution

Hw4

Hw4

- Q1-Q3: kNN
- Q4: chi-square for feature selection
- Q5-Q6: The effect of feature selection on kNN
- Q7: Conclusion

Q1-Q3: kNN

- The choice of k
- The choice of similarity function:
  - Euclidean distance: choose the smallest ones
  - Cosine function: choose the largest ones
- Binary vs. real-valued features

Q4-Q6

- Rank features by chi-square scores
- Remove non-relevant features from the vector files
- Run kNN using the newly processed data
- Compare the results with or without feature selection.