UCLA Samueli Computer Science

CS145 Discussion: Week 10

Naive Bayes, Topic Modeling, Final Review

Junheng Hao

Friday, 12/11/2020

UCLA
Engineer Change.

Announcements

Homework #6	Due on 11:59 PM, **Dec 14** (Monday, final week)
Final Exam (100 minutes via CCLE)	Morning Session: Around 8:00-9:40 AM, **Dec 16** (Wednesday, final Week) Evening Session: Around 6:00-7:40 PM, **Dec 16** (Wednesday, final Week)
Project: Final Report	Due on 11:59 PM, **Dec 18** (Friday, final week)
All regrade request	Due on 11:59 PM, **Dec 21** (Monday, the week after final week)

Note: Practice exam for final exam setup has been added on CCLE, under the tab of “Week 11”.

UCLA
Engineer Change.

Roadmap

- Basics
  - Text Data
  - MLE v.s. MAP
- Naïve Bayes
- pLSA
- Final Exam Q & A

How to represent text data?

- Most common way: Bag-of-Words
  - Ignore the order of words
  - keep the count

c1: *Human* machine *interface* for Lab ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user*-perceived *response time* to error measurement

m1: The generation of random, binary, unordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

For document d, $\boldsymbol{x}_d = (x_{d1}, x_{d2}, \ldots, x_{dN})$, where x_{dn} is the number of words for nth word in the vocabulary

Vector space model

UCLA
Engineer Change.

MLE v.s. MAP

- *Maximum Likelihood* (maximize the likelihood):

$$h_{ML} = \arg\max_{h \in H} P(X \mid h)$$

- *Maximum a posteriori* (maximize the posterior):
  - Useful observation: it does not depend on the denominator P(X)

$$h_{MAP} = \arg\max_{h \in H} P(h \mid X) = \arg\max_{h \in H} P(X \mid h) P(h)$$

UCLA
Engineer Change.

MLE v.s. MAP

- *Maximum Likelihood* (maximize the likelihood):

h is a parameter

$$h_{ML} = \arg\max_{h \in H} P(X \mid h)$$

- *Maximum a posteriori* (maximize the posterior):
  - Useful observation: it does not depend on the denominator P(X)

h is a random variable

$$h_{MAP} = \arg\max_{h \in H} P(h \mid X) = \arg\max_{h \in H} P(X \mid h)P(h)$$

UCLA
Engineer Change.

Naïve Bayes: Modeling

- A document is represented as a bag of words
  - $\boldsymbol{x}_d = (x_{d1}, x_{d2}, \ldots, x_{dN})$, where x_{dn} is the number of words for nth word in the vocabulary
- Model $p(\boldsymbol{x}_d|y)$ for class y
  - Follow multinomial distribution with parameter vector $\boldsymbol{\beta}_y = (\beta_{y1}, \beta_{y2}, \ldots, \beta_{yN})$, i.e.,
    - $p(\boldsymbol{x}_d|y) = \frac{(\sum_n x_{dn})!}{x_{d1}!x_{d2}!\ldots x_{dN}!} \prod_n \beta_{yn}^{x_{dn}}$
- Model $p(y = j)$
  - Follow categorical distribution with parameter vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_m)$, i.e.,
    - $p(y = j) = \pi_j$

UCLA
Engineer Change.

Naïve Bayes: Inference

- Find y that maximizes $p(y|\boldsymbol{x}_d)$, which is equivalently to maximize

$$y^* = \underset{y}{argmax}\ p(\boldsymbol{x}_d, y)$$

$$= argmax_y\ p(\boldsymbol{x}_d|y)p(y)$$

$$= argmax_y \frac{(\sum_n x_{dn})!}{x_{d1}!\, x_{d2}! \ldots x_{dN}!} \prod_n \beta_{yn}^{x_{dn}} \times \pi_y$$

Constant for every class, denoted as c_d

$$= argmax_y \prod_n \beta_{yn}^{x_{dn}} \times \pi_y$$

$$= argmax_y \sum_n x_{dn} log\beta_{yn} + log\pi_y$$

UCLA
Engineer Change.

Naïve Bayes: Learning

- Given a corpus and labels for each document
  - $D = \{(\boldsymbol{x}_d, y_d)\}$
  - Find the MLE estimators for $\Theta = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_m, \boldsymbol{\pi})$
- The log likelihood function for the training dataset

$$logL = log \prod_d p(\boldsymbol{x}_d, y_d|\Theta) = \sum_d log\, p(\boldsymbol{x}_d, y_d|\Theta)$$

$$= \sum_d log\, p(\boldsymbol{x}_d|y_d)p(y_d) = \sum_d x_{dn} log\beta_{yn} + log\pi_{y_d} + \boxed{logc_d}$$

- The optimization problem

$$\max_{\Theta} \log L$$

$$s.t.$$

$$\pi_j \geq 0 \; and \sum_j \pi_j = 1$$

$$\beta_{jn} \geq 0 \; and \sum_n \beta_{jn} = 1 \; for \; all \; j$$

Does not involve parameters, can be dropped for optimization purpose

UCLA
Engineer Change.

Naïve Bayes: Learning

- $\hat{\beta}_{jn} = \frac{\sum_{d:y_d=j} x_{dn}}{\sum_{d:y_d=j} \sum_{n'} x_{dn'}}$
  - $\sum_{d:y_d=j} x_{dn}$: total count of word n in class j
  - $\sum_{d:y_d=j} \sum_{n'} x_{dn'}$: total count of words in class j
- $\hat{\pi}_j = \frac{\sum_d 1(y_d=j)}{|D|}$
  - $1(y_d = j)$ is the indicator function, which equals to 1 if $y_d = j$ holds
  - |D|: total number of documents

UCLA
Engineer Change.

Naïve Bayes: Smoothing

- What if some word n does not appear in some class j in training dataset?
  - $\hat{\beta}_{jn} = \frac{\sum_{d:y_d=j} x_{dn}}{\sum_{d:y_d=j} \sum_{n'} x_{dn'}} = 0$
  - $\Rightarrow p(\boldsymbol{x}_d | y = j) \propto \prod_n \beta_{yn}^{x_{dn}} = 0$
  - But other words may have a strong indication the document belongs to class j
- Solution: add-1 smoothing or Laplacian smoothing
  - $\hat{\beta}_{jn} = \frac{\sum_{d:y_d=j} x_{dn}+1}{\sum_{d:y_d=j} \sum_{n'} x_{dn'}+N}$
  - N: *total number of words in the vocabulary*
  - Check: $\sum_n \hat{\beta}_{jn} = 1$?

UCLA
Engineer Change.

Naïve Bayes: Example

Index	Word	# in class1	# in class2
1	I	1	9
2	like	2	1
3	data	3	9
4	mining	4	1

beta = ?
pi = ?
classification result of “I like data”?

UCLA
Engineer Change.

Naïve Bayes: Example

Index	Word	# in class1	# in class2
1	I	1	9
2	like	2	1
3	data	3	9
4	mining	4	1

beta11 = 1/10
beta12 = 2/10
beta13 = 3/10
beta14 = 4/10

beta21 = 9/20
beta22 = 1/20
beta23 = 9/20
beta24 = 1/20

pi1 = 1/3
pi2 = 2/3

P(class1 | “I like data”) $\propto$ 1/3 * 1/10 * 2/10 * 3/10 = 6/3000

P(class2 | “I like data”) $\propto$ 2/3 * 9/20 * 1/20 * 9/20 = 81/12000

UCLA
Engineer Change.

Generative vs Discriminative Models

- Training classifiers involve estimating *f: X -> Y*, or P(Y|X)
- Generative classifiers → *"distribution"*
  - Assume some functional form for P(Y), P(X|Y)
  - Estimate parameters of P(X|Y), P(Y) directly from training data
  - Use Bayes rule to calculate P(Y |X)
  - Actually learn the underlying structure of the data
- Discriminative Classifiers → *"boundary"*
  - Assume some functional form for P(Y|X)
  - Estimate parameters of P(Y|X) directly from training data.
  - Learn the mappings directly from the points to the classes

- Generative classifiers
  - Naive Bayes
- Discriminative classifiers
  - Logistic Regression
  - Neural Network
  - SVM

Q: With the aim of classification only, which type of models may less expensive?

Credit:
https://stats.stackexchange.com/questions/12421/generative-vs-discriminative

Topic Models

Each document can have more than one topic.

A more flexible model than Naïve Bayes

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

UCLA
Engineer Change.

Probabilistic Latent Semantic Analysis (pLSA)

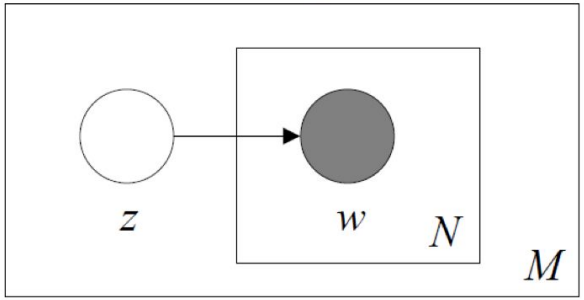
- Word, document, topic
  - w, d, z
- Word count in document
  - $c(w, d)$
- Word distribution for each topic (β_z)
  - β_{zw}: $p(w|z)$
- Topic distribution for each document (θ_d)
  - θ_{dz}: $p(z|d)$ (Yes, soft clustering)

Probabilistic Latent Semantic Analysis (pLSA)

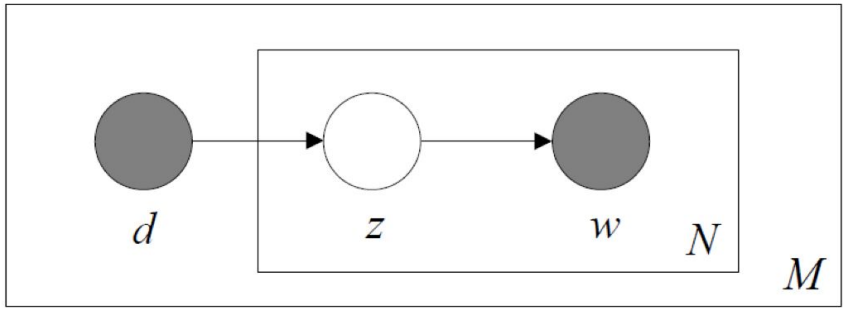
- For each position in d, $n = 1, \ldots, N_d$
  - Generate the topic for the position as
    $z_n \sim Multinoulli(\boldsymbol{\theta}_d), i.e., p(z_n = k) = \theta_{dk}$
    (Note, 1 trial multinomial, i.e., categorical distribution)
  - Generate the word for the position as
    $w_n \sim Multinoulli(\boldsymbol{\beta}_{z_n}), i.e., p(w_n = w) = \beta_{z_n w}$

pLSA: Graphical model

Naïve Bayes
(Multinomial mixture model)



pLSA



UCLA
Engineer Change.

pLSA: Inference

- Probability of a word

$$p(w|d) = \sum_k p(w, z = k|d) = \sum_k p(w|z = k)p(z = k|d) = \sum_k \beta_{kw}\theta_{dk}$$

UCLA
Engineer Change.

pLSA: Learning

$$\max logL = \sum_{dw} c(w,d) log \sum_{z} \theta_{dz}\, \beta_{zw}$$

$$s.t. \sum_{z} \theta_{dz} = 1 \; and \; \sum_{w} \beta_{zw} = 1$$

pLSA: Optimization

- Repeat until converge
  - E-step: for each word in each document, calculate its conditional probability belonging to each topic

$$p(z|w,d) \propto p(w|z,d)p(z|d) = \beta_{zw}\theta_{dz} \; (i.e., p(z|w,d) = \frac{\beta_{zw}\theta_{dz}}{\sum_{z'} \beta_{z'w}\theta_{dz'}})$$

  - M-step: given the conditional distribution, find the parameters that can maximize the expected complete log-likelihood

$$\beta_{zw} \propto \sum_d p(z|w,d)c(w,d) \; (i.e., \beta_{zw} = \frac{\sum_d p(z|w,d)c(w,d)}{\sum_{w',d} p(z|w',d)c(w',d)})$$

$$\theta_{dz} \propto \sum_w p(z|w,d)c(w,d) \; (i.e., \theta_{dz} = \frac{\sum_w p(z|w,d)c(w,d)}{N_d})$$

UCLA
Engineer Change.

pLSA: Example

Two documents d1 and d2
Two topics z1 and z2

Initialize beta_wz1 = 1/6
Initialize beta_wiz2
= 1/12 for i <= 4
  1/3 for i > 4

For all z and d,
Initialize theta_zd = 1/2

E step?
M step?

d1

Index	Word (w)	Count	P(z1 | w, d1) ∝	P(z2 | w, d1) ∝
1	frequent	4	?	?
2	pattern	3	?	?
3	data	2	?	?
5	cat	1	?	?

d2

Index	Word (w)	Count	P(z1 | w, d2) ∝	P(z2 | w, d2) ∝
3	data	1	?	?
4	mining	2	?	?
5	cat	3	?	?
6	dog	4	?	?

pLSA: Example

Two documents d1 and d2
Two topics z1 and z2

Initialize beta_wz1 = 1/6
Initialize beta_wiz2
= 1/12 for i <= 4
  1/3 for i > 4

For all z and d,
Initialize theta_zd = 1/2

d1

Index	Word (w)	Count	P(z1 | w, d1) ∝	P(z2 | w, d1) ∝
1	frequent	4	1/12	1/24
2	pattern	3	1/12	1/24
3	data	2	1/12	1/24
5	cat	1	1/12	1/6

d2

Index	Word (w)	Count	P(z1 | w, d2) ∝	P(z2 | w, d2) ∝
3	data	1	1/12	1/24
4	mining	2	1/12	1/24
5	cat	3	1/12	1/6
6	dog	4	1/12	1/6

E step result

pLSA: Example

Two documents d1 and d2
Two topics z1 and z2

Initialize beta_wz1 = 1/6
Initialize beta_wiz2
= 1/12 for i <= 4
  1/3 for i > 4

For all z and d,
Initialize theta_zd = 1/2

d1

Index	Word (w)	Count	P(z1 | w, d1) ∝	P(z2 | w, d1) ∝
1	frequent	4	1/12	1/24
2	pattern	3	1/12	1/24
3	data	2	1/12	1/24
5	cat	1	1/12	1/6

d2

Index	Word (w)	Count	P(z1 | w, d2) ∝	P(z2 | w, d2) ∝
3	data	1	1/12	1/24
4	mining	2	1/12	1/24
5	cat	3	1/12	1/6
6	dog	4	1/12	1/6

M step result:

beta_w1z1 ∝ 4/12
beta_w2z1 ∝ 3/12
beta_w3z1 ∝ 3/12

beta_w1z2 ∝ 4/24
beta_w2z2 ∝ 3/24
beta_w3z2 ∝ 3/24

theta_z1d1 ∝ 10/12
theta_z2d1 ∝

theta_z1d2 ∝ 10/12
theta_z2d2 ∝

UCLA
Engineer Change.

Final Exam Q & A

- Topics after the midterm (80%)
  - Frequent pattern mining
  - Sequential pattern mining
  - Time series data
  - Naïve Bayes
  - Topic modeling
- Topics before the midterm (20%, most likely only T/F)
  - Linear regression
  - Logistic regression
  - SVM
  - Decision Trees
  - Neural Network
  - KNN
  - Clustering

UCLA
Engineer Change.

Continuing Grad-level Courses

- CS247: Advanced Data Mining *(coming Spring 2021)*
- CS245: Graph Neural Network *(coming Winter 2021)*

UCLA Samueli Computer Science

Thank you!

Q & A

Reminder: Students have until **8:00 AM Wednesday, December 16** to log into MyUCLA to complete evaluations for COM SCI 145 section 1C.

Whiteboard

This is an intentionally blank page.