Precept 2: Graphical models, Naïve Bayes and Text analysis

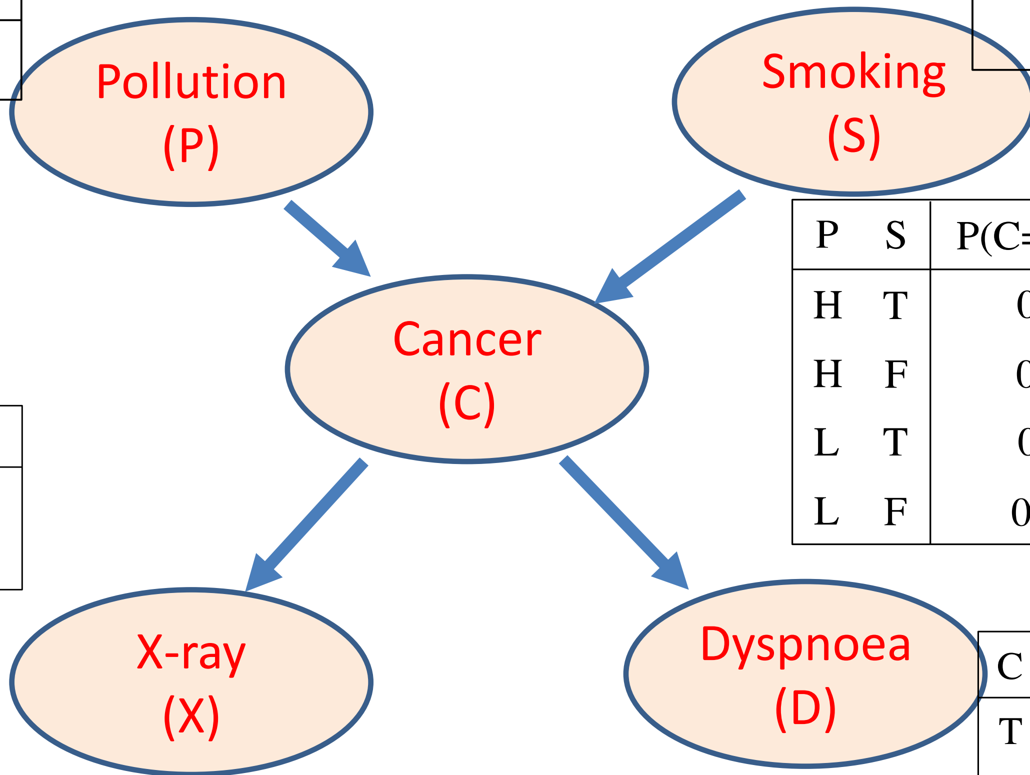
COS424 Spring 2017

Xiaoyan Li

Answers to Quiz 1

- https://docs.google.com/forms/d/1i5NODaS_fm81Xen900_VJU9Qrbew12pa53XJ44IxHsQ/edit?ts=5898acf9
- Q1:A
  - The sufficient statistics of a data are all we need to know about the data to infer the model parameters.
- Q2:C
  - Train error decreases; test error initially decreases and then increases with increasing iterations
- Q3:D Regression
- Q4:A(multinomial) or C(Dirichlet)

Graphical models/Bayes nets—reason with uncertainty



P(P=L)
0.90

P(S=T)
0.30

P	S	P(C=T|P,S)
H	T	0.05
H	F	0.02
L	T	0.03
L	F	0.001

C	P(X=pos|C)
T	0.90
F	0.20

C	P(D=T|C)
T	0.65
F	0.30

Full joint probability:

Pr[X, D, C, P, S] = Pr[X|C] Pr[D|C] Pr[C|P, S] Pr[S] Pr[P]

Bayes net—Exact inference

- Two types of interesting calculations

Marginal distribution:

$$\mathbb{P}(A) = \sum_b \mathbb{P}(A, B = b)$$

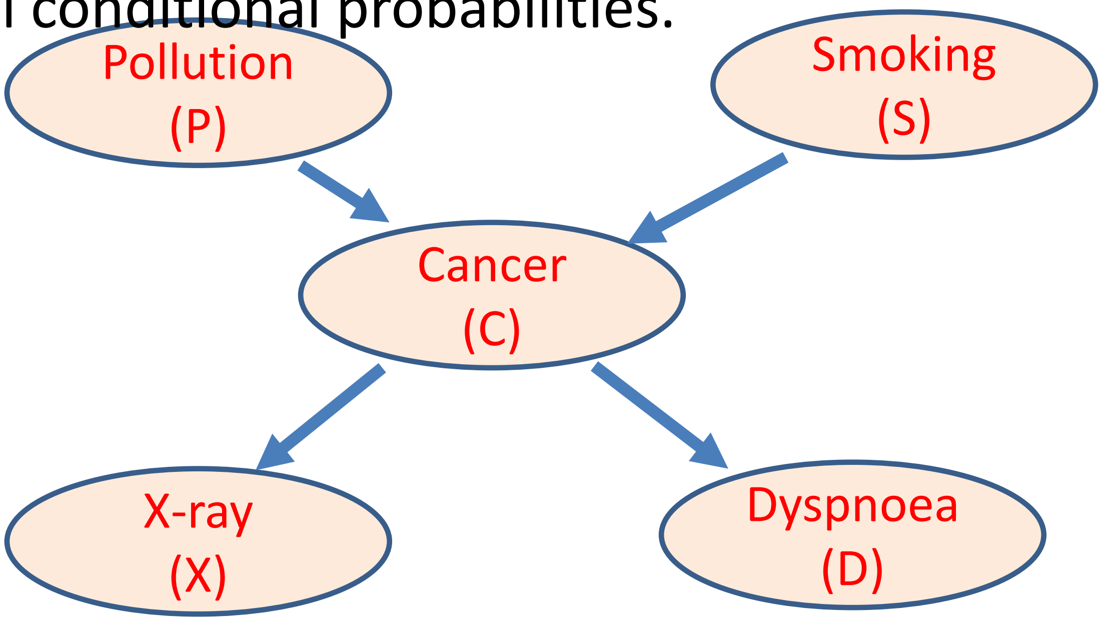
Conditional distribution:

$$\mathbb{P}(A \mid B = b) = \frac{\mathbb{P}(A, B=b)}{\mathbb{P}(B=b)} = \frac{P(A, B = b)}{\sum_a P(A = a, B = b)}$$

For both tasks, we need to marginalize out some variables.

Bayes net—Approximate inference

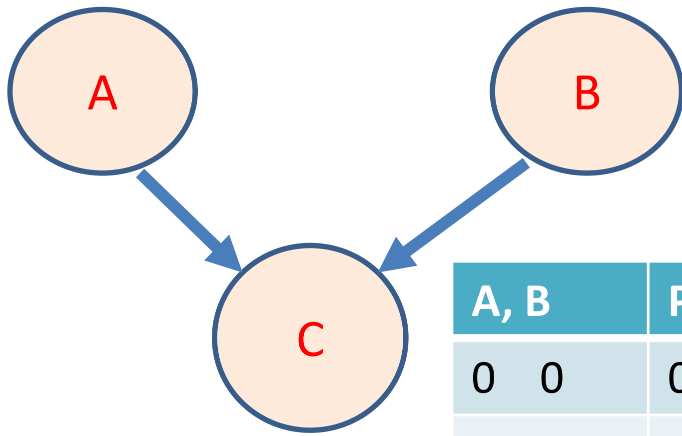
- Random sampling:
  - Randomly assign values to all the random variables in the network(P=H,S=T,C=T,X=pos,D=F)
  - Sample parent nodes before children nodes based on local conditional probabilities.



Pr[X, D, C, P, S] = Pr[X|C] Pr[D|C] Pr[C|P, S] Pr[S] Pr[P]

Explaining away:

P(A=1)
0.01



P(B=1)
0.6

A, B	P(C=1|A,B)
0 0	0.01
0 1	0.9
1 0	0.9
1 1	1

Two random variables become dependent if a common child is observed. Knowing A affects the probability of B when C is observed.

P(A=1) = 0.01

P(A=1|B=1)= ?

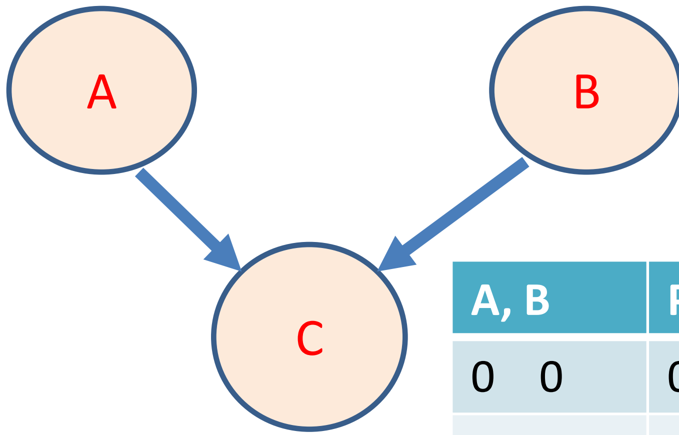
P(A=1|C=1) = ?

P(A=1|C=1,B=0) = ?

P(A=1|C=1,B=1) = ?

Explaining away:

P(A=1)
0.01



P(B=1)
0.6

A, B	P(C=1|A,B)
0 0	0.01
0 1	0.9
1 0	0.9
1 1	1

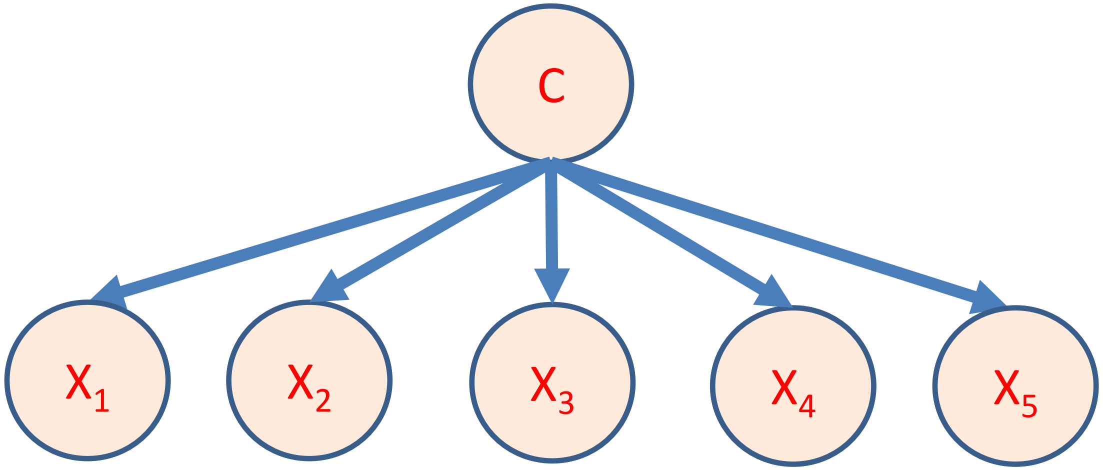
Two random variables become dependent if a common child is observed. Know A affects the probability of B when C is observed.

$P(A=1) = 0.01$, $P(A=1|B=1)=P(A=1)=0.01$

$P(A=1|C=1)= 0.15$

$P(A=1|C=1,B=0) = 0.91$, $P(A=1|C=1,B=1) = 0.11$

Naïve Bayes classifier



Given a sample: $x=(x_1, x_2, x_3, x_4, x_5)$,

What is the label of x? $c^* = \underset{c}{\operatorname{argmax}} P(C = c|x)$

$P(C|x) = \frac{P(C)P(x|C)}{p(x)} \propto P(C)\prod_i P(x_i|C)$ (why?)

Naïve Bayes classifier on sentence/document classification:

No.	Sentences	Class
S1	cat dog cat	1
S2	cat fox cat	1
S3	cat mouse	1
S4	apple banana cat	0
S5	cat apple cat banana cat	?
S6	apple cat elephant	?

Sentence representation: vector of binary values

No.	Sentences	(apple, banana, cat, dog, fox, mouse)	Class
S1	cat dog cat	(0, 0, 1, 1, 0, 0)	1
S2	cat fox cat	(0, 0, 1, 0, 1, 0)	1
S3	cat mouse	(0, 0, 1, 0, 0, 1)	1
S4	apple banana cat	(1, 1, 1, 0, 0, 0)	0
S5	cat apple cat banana cat		?
S6	apple cat elephant		?

Q1: How to decide the number of features, or what is the length of the vectors?
Q2: How to deal with words in test set but not in the training set?

Sentence representation 1: vector of binary values

No.	Sentences	(apple, banana, cat, dog, fox, mouse)	Class
S1	cat dog cat	(0, 0, 1, 1, 0, 0)	1
S2	cat fox cat	(0, 0, 1, 0, 1, 0)	1
S3	cat mouse	(0, 0, 1, 0, 0, 1)	1
S4	apple banana cat	(1, 1, 1, 0, 0, 0)	0
S5	cat apple cat banana cat	(1, 1, 1, 0, 0, 0)	?
S6	apple cat elephant	(1, 0, 1, 0, 0, 0)	?

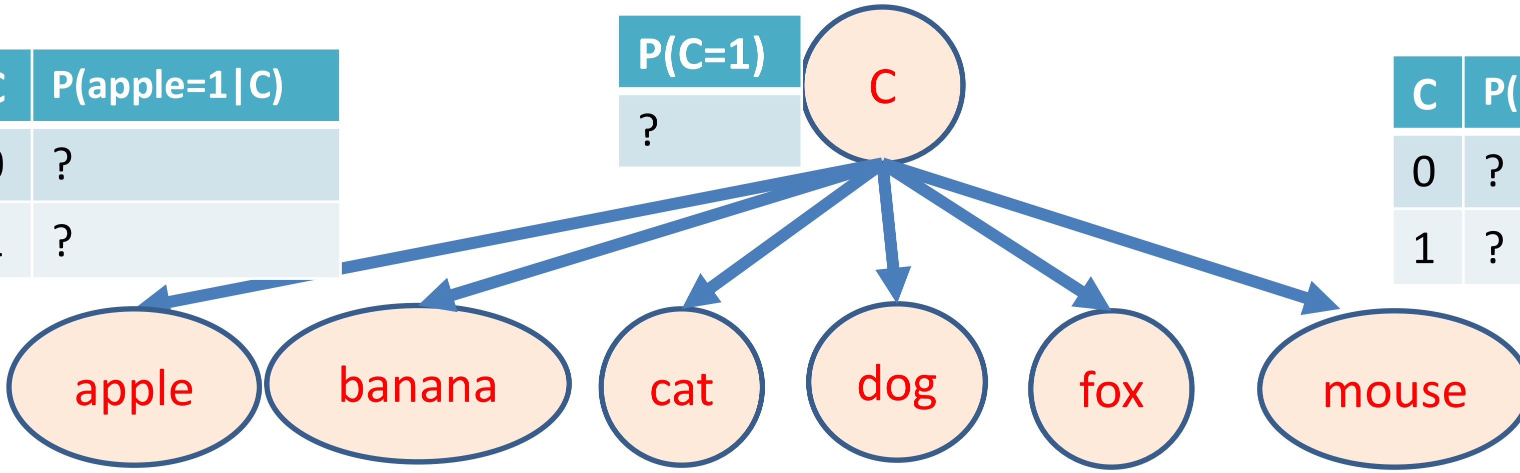
Q1: All words/tokens in the training set. Could apply some feature selection techniques.
Q2: Can simply ignore them("elephant" in S6).

Bernoulli Naïve Bayes Classifier

C	P(apple=1|C)
0	?
1	?

P(C=1)
?

C	P(mouse=1|C)
0	?
1	?



No.	Sentences (training data)	(apple, banana, cat, dog, fox, mouse)	Class
S1	cat dog cat	(0, 0, 1, 1, 0, 0)	1
S2	cat fox cat	(0, 0, 1, 0, 1, 0)	1
S3	cat mouse	(0, 0, 1, 0, 0, 1)	1
S4	apple banana cat	(1, 1, 1, 0, 0, 0)	0

Q: What is the total number of parameters?

Bernoulli Naïve Bayes Classifier(Training)

No.	Sentences (training set)	(apple, banana, cat, dog, fox, mouse)	Class
S1	cat dog cat	(0, 0, 1, 1, 0, 0)	1
S2	cat fox cat	(0, 0, 1, 0, 1, 0)	1
S3	cat mouse	(0, 0, 1, 0, 0, 1)	1
S4	apple banana cat	(1, 1, 1, 0, 0, 0)	0

Maximum likelihood estimate: P(apple=1|C=1) = 0/3=0, P(apple=1|C=0) = 1/1=1
Add-1-smoothing:

P(apple=1|C=1) = 0+1/3+2=1/5, P(apple=1|C=0) = 1+1/1+2=2/3
P(banana=1|C=1) = 0+1/3+2=1/5, P(banana=1|C=0) = 1+1/1+2=2/3
P(cat=1|C=1) = 3+1/3+2=4/5, P(cat=1|C=0) = 1+1/1+2=2/3
P(dog=1|C=1) = ? , P(dog=1|C=0) = ?
P(fox=1|C=1) = ?, P(fox=1|C=0) = ?
P(mouse=1|C=1) = ? , P(mouse=1|C=0) = ?

Bernoulli Naïve Bayes Classifier(Training)

No.	Sentences (training set)	(apple, banana, cat, dog, fox, mouse)	Class
S1	cat dog cat	(0, 0, 1, 1, 0, 0)	1
S2	cat fox cat	(0, 0, 1, 0, 1, 0)	1
S3	cat mouse	(0, 0, 1, 0, 0, 1)	1
S4	apple banana cat	(1, 1, 1, 0, 0, 0)	0

Class proportion: P(C=1) = 3/4

Maximum likelihood estimate: P(apple=1|C=1) = 0/3=0, P(apple=1|C=0) = 1/1=1

Add-1-smoothing:

P(apple=1|C=1) = 0+1/3+2=1/5, P(apple=1|C=0) = 1+1/1+2=2/3

P(banana=1|C=1) = 0+1/3+2=1/5, P(banana=1|C=0) = 1+1/1+2=2/3

P(cat=1|C=1) = 3+1/3+2=4/5, P(cat=1|C=0) = 1+1/1+2=2/3

P(dog=1|C=1) = 1+1/3+2=2/5, P(dog=1|C=0) = 0+1/1+2=1/3

P(fox=1|C=1) = 1+1/3+2=2/5, P(fox=1|C=0) = 0+1/1+2=1/3

P(mouse=1|C=1) = 1+1/3+2=2/5, P(mouse=1|C=0) = 0+1/1+2=1/3

Bernoulli Naïve Bayes Classifier(predicting)

Class proportion: P(C=1) = 3/4

Maximum likelihood estimate: P(apple=1|C=1) = 0/3=0, P(apple=1|C=0) = 1/1=1

Add-1-smoothing:

P(apple=1|C=1) = 0+1/3+2=1/5, P(apple=1|C=0) = 1+1/1+2=2/3

P(banana=1|C=1) = 0+1/3+2=1/5, P(banana=1|C=0) = 1+1/1+2=2/3

P(cat=1|C=1) = 3+1/3+2=4/5, P(cat=1|C=0) = 1+1/1+2=2/3

P(dog=1|C=1) = 1+1/3+2=2/5, P(dog=1|C=0) = 0+1/1+2=1/3

P(fox=1|C=1) = 1+1/3+2=2/5, P(fox=1|C=0) = 0+1/1+2=1/3

P(mouse=1|C=1) = 1+1/3+2=2/5, P(mouse=1|C=0) = 0+1/1+2=1/3

No.	Sentences(test set)	(apple, banana, cat, dog, fox, mouse)	Class
S5	cat apple cat banana cat	(1, 1, 1, 0, 0, 0)	? 0
S6	apple cat elephant	(1, 0, 1, 0, 0, 0)	?

P(C=1|(1,1,1,0,0,0)) ∝ P(C)P(apple=1|C=1)P(banana=1|C=1)P(cat=1|C=1)
P(dog=0|C=1)P(fox=0|C=1)P(mouse=0|C=1)
=3/4*1/5*1/5*4/5*(1-2/5)*(1-2/5)*(1-2/5)=0.005

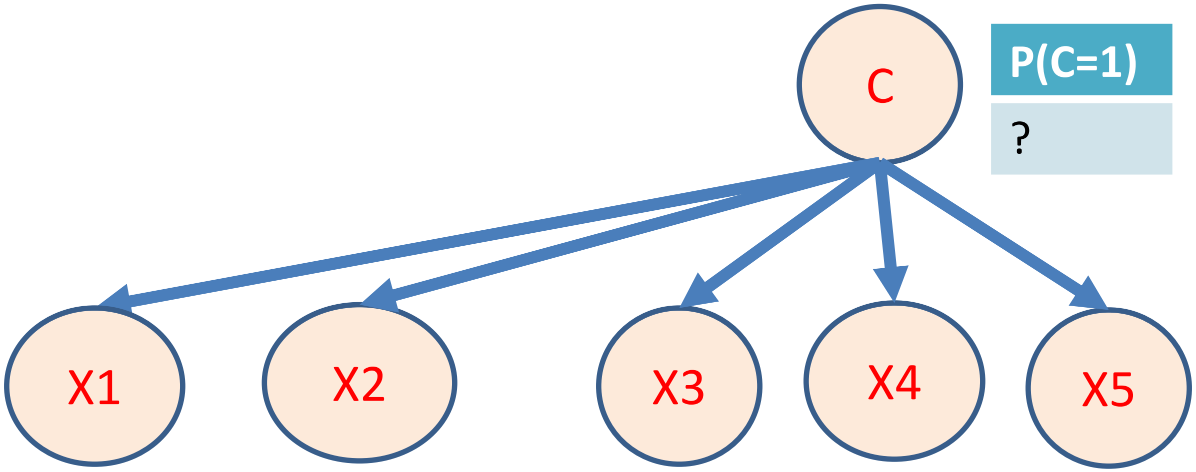
P(C=0|(1,1,1,0,0,0)) ∝ 1/4*2/3*2/3*3/3*(1-1/3)*(1-1/3)*(1-1/3)=0.022

C* = ?

Sentence representation 2: vector of word counts

No.	Sentences	(apple, banana, cat, dog, fox, mouse)	Class
S1	cat dog cat	(0, 0, 2, 1, 0, 0)	1
S2	cat fox cat	(0, 0, 2, 0, 1, 0)	1
S3	cat mouse	(0, 0, 1, 0, 0, 1)	1
S4	apple banana cat	(1, 1, 1, 0, 0, 0)	0
S5	cat apple cat banana cat	(1, 1, 3, 0, 0, 0)	?
S6	apple cat elephant	(1, 0, 1, 0, 0, 0)	?

Multinomial Naïve Bayes Classifier

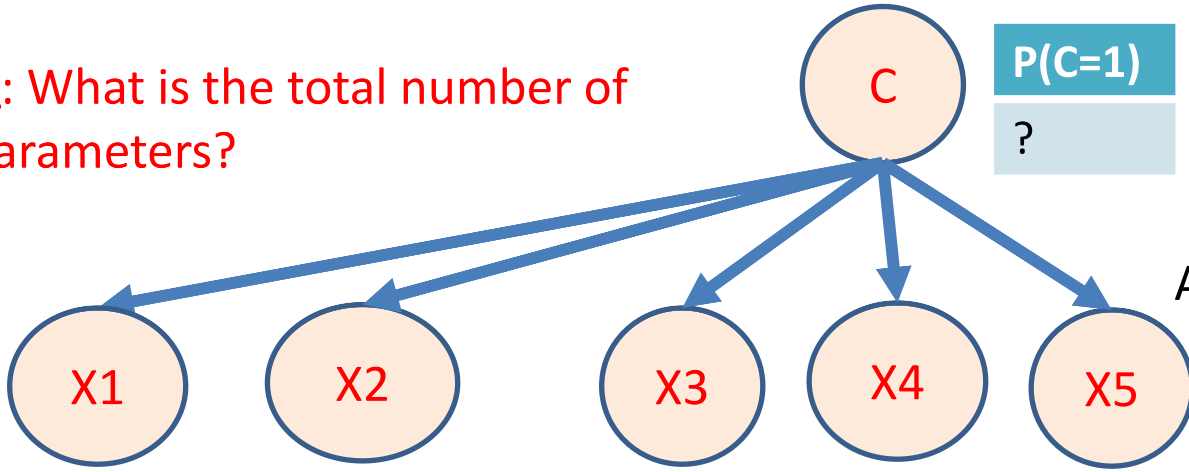


Q: What are X's here?

No.	Sentences (training data)	(apple, banana, cat, dog, fox, mouse)	Class
S1	cat dog cat	(0, 0, 2, 1, 0, 0)	1
S2	cat fox cat	(0, 0, 2, 0, 1, 0)	1
S3	cat mouse	(0, 0, 1, 0, 0, 1)	1
S4	apple banana cat	(1, 1, 1, 0, 0, 0)	0

Multinomial Naïve Bayes Classifier

Q: What is the total number of parameters?



P(C=1)
?

Q: What are X's here?

A: positions in a sentence.

C	P(Xi=apple|C)	P(Xi=banana|C)	P(Xi=cat|C)	P(Xi=dog|C)	P(Xi=fox|C)	P(Xi=mouse|C)
0						
1						

No.	Sentences (training data)	(apple, banana, cat, dog, fox, mouse)	Class
S1	cat dog cat	(0, 0, 2, 1, 0, 0)	1
S2	cat fox cat	(0, 0, 2, 0, 1, 0)	1
S3	cat mouse	(0, 0, 1, 0, 0, 1)	1
S4	apple banana cat	(1, 1, 1, 0, 0, 0)	0

Multinomial Naïve Bayes Classifier(Training)

No.	Sentences (training data)	(apple, banana, cat, dog, fox, mouse)	Class
S1	cat dog cat	(0, 0, 2, 1, 0, 0)	1
S2	cat fox cat	(0, 0, 2, 0, 1, 0)	1
S3	cat mouse	(0, 0, 1, 0, 0, 1)	1
S4	apple banana cat	(1, 1, 1, 0, 0, 0)	0

Class proportion: P(C=1) = 3/4

Maximum likelihood estimate + Add-1-smoothing:

P(apple|C=1) = 0+1/8+6=1/14, P(apple|C=0) = 1+1/3+6=2/9

P(banana|C=1) = 0+1/8+6=1/14, P(banana|C=0) = 1+1/3+6=2/9

P(cat|C=1) = 5+1/8+6=3/7, P(cat|C=0) = 1+1/3+6=2/9

P(dog|C=1) = 1+1/8+6=1/7, P(dog|C=0) = 0+1/3+6=1/9

P(fox|C=1) = 1+1/8+6=1/7, P(fox|C=0) = 0+1/3+6=1/9

P(mouse|C=1) = 1+1/8+6=1/7, P(mouse|C=0) = 0+1/3+6=1/9

Q: What is the total number of parameters? 13 or 11

Multinomial Naïve Bayes Classifier(Predicting)

No.	Sentences(test set)	(apple, banana, cat, dog, fox, mouse)	Class
S5	cat apple cat banana cat	(1, 1, 3, 0, 0, 0)	? 1
S6	apple cat	(1, 0, 1, 0, 0, 0)	?

Class proportion: P(C=1) = 3/4

P(apple|C=1) = 0+1/8+6=1/14, P(apple|C=0) = 1+1/3+6=2/9

P(banana|C=1) = 0+1/8+6=1/14, P(banana|C=0) = 1+1/3+6=2/9

P(cat|C=1) = 5+1/8+6=3/7, P(cat|C=0) = 1+1/3+6=2/9

P(dog|C=1) = 1+1/8+6=1/7, P(dog|C=0) = 0+1/3+6=1/9

P(fox|C=1) = 1+1/8+6=1/7, P(fox|C=0) = 0+1/3+6=1/9

P(mouse|C=1) = 1+1/8+6=1/7, P(mouse|C=0) = 0+1/3+6=1/9

$P(C=1|(1,1,1,0,0,0)) \propto P(C)P(apple|C=1)P(banana|C=1)P(cat|C=1)^3$
$=3/4*1/14*1/14*(3/7)^3=0.0003$

$P(C=0|(1,1,1,0,0,0)) \propto P(C)\ P(apple|C=0)P(banana|C=0)P(cat|C=0)^3$
$=1/4*2/9*2/9*(2/9)^3=0.0001$

C* = ?

Naïve Bayes classifiers in Scikit learn

- GaussianNB
- BernoulliNB
  - S is a vector of binary values
  - i.e. (1,1,1,0,0,0)
- MultinomialNB
  - S is a vector of word counts, i.e (0,0,2,1,0,0)
  - In practice, it also works when S is a vector of tfidf scores. i.e. (0.1,.0,1, 0.3, 0.2, 0,1, 0.1)
    - tf: term frequency, idf: inverse document frequency.

Review Questions:

- Is a Naive Bayes classifier a generative model? Why?
- How to generate a sentence in the Multinomial Naïve Bayes model?
- How to generate a sentence in the Bernoulli naïve Bayes model?
- Are the estimates of the class probabilities for predictions very accurate? If not, why do we use them?
- What is the graphic model for the Naive Bayes classifier?
- Can the features have different distributions?
- What libraries/packages in python is available for Naive Bayes classification with features of different distributions?

Grading and expectations of assignment 1

- C to B+:
  - Approached the problem correctly,
  - Did very basic tasks and data analysis
  - A complete report
- A-, A
  - Approached the problem correctly, motivated each method used,
  - Did some of the extensions
  - A well written report
- A+
  - Reserved for exceptional work.

Cross-validation

- Fit a model on training set
  - Training error: the error of the fitted model on the training set
    - e.g. the error of on the training set of the 4 sentences.
  - test/generalization error: the error of the fitted model on the test set/unobserved data
    - e.g. the error on the test set of the 2 sentences.
- Cross validation
  - To quantify the generalization error.

Cross-validation: Quantify generalization error

- K-fold cross validation
  - Partition data randomly into k folds, or equal disjointed subsets.
  - For i =1, 2, …, K
    - Let fold i be the test(held out) fold.
    - Fit the model on the other K-1 folds.
    - Predict on the test fold.
  - Compute generalization error from one prediction for each sample

- Leave-one-out cross validation
  - when k = n, n is the total number of samples

Cross-validation: Quantify generalization error

- How to choose the number of folds (K=?)
  - 5-fold and 10-fold cross validations are more commonly used in practice.
  - Trade-off between computation speed for training and the number of samples in training.
    - If you have a very slow method for training your model, you should make K small.

Cross-validation: Hyperparameter fitting

- Often used to fit hyperparamters.
  - e.g. Fit the number K in the K-nearest neighbors classifier; Fit the number of trees (n-estimators) in random forest classifier
- Hyperparameter fitting is a inner loop of training method.
  - Perform K-fold cross validation for hyperparamter estimation on the current training data.
  - Try different values of n, select the n with the lowest generalization error.
  - *Not appropriate to double dip the data(use for both training and test)

Resources:

- Text classification and Naive Bayes
  - http://nlp.stanford.edu/IR-book/html/htmledition/text-classification-and-naive-bayes-1.html
  - Examples in the slides were modified from above resource
- Some slides are taken from lectures and precepts from COS402