10-601 Fall 2012
Recitation on Naïve Bayes

TA: Selen Uguroglu

September 18, 2012

Outline

- Conditional independence
- Naïve Bayes assumption and its consequences
  - Which (and how many) parameters must be estimated under different generative models (different forms for $P(X|Y)$)
    - and why this matters
- How to train Naïve Bayes classifiers
  - MLE and MAP estimates
  - with discrete and/or continuous inputs X_i

Conditional Independence

- Given random variables X, Y and Z.
- X is conditionally independent of Y given Z, if and only if:

$$P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k), \forall i, j, k$$

$$P(X_1, X_2 | Y) = P(X_1 | X_2, Y) P(X_2 | Y)$$

$$= P(X_1 | Y) P(X_2 | Y)$$

If we have n variables, assuming conditional independence, we can write:

$$P(X_1, X_2, ..., X_n | Y) = \prod_{i=1}^{n} P(X_i | Y)$$

Parameters needed

- How many parameters we need to estimate:

$$P(X_1, X_2, ..., X_n|Y)$$

where X_i and Y are Boolean random variables

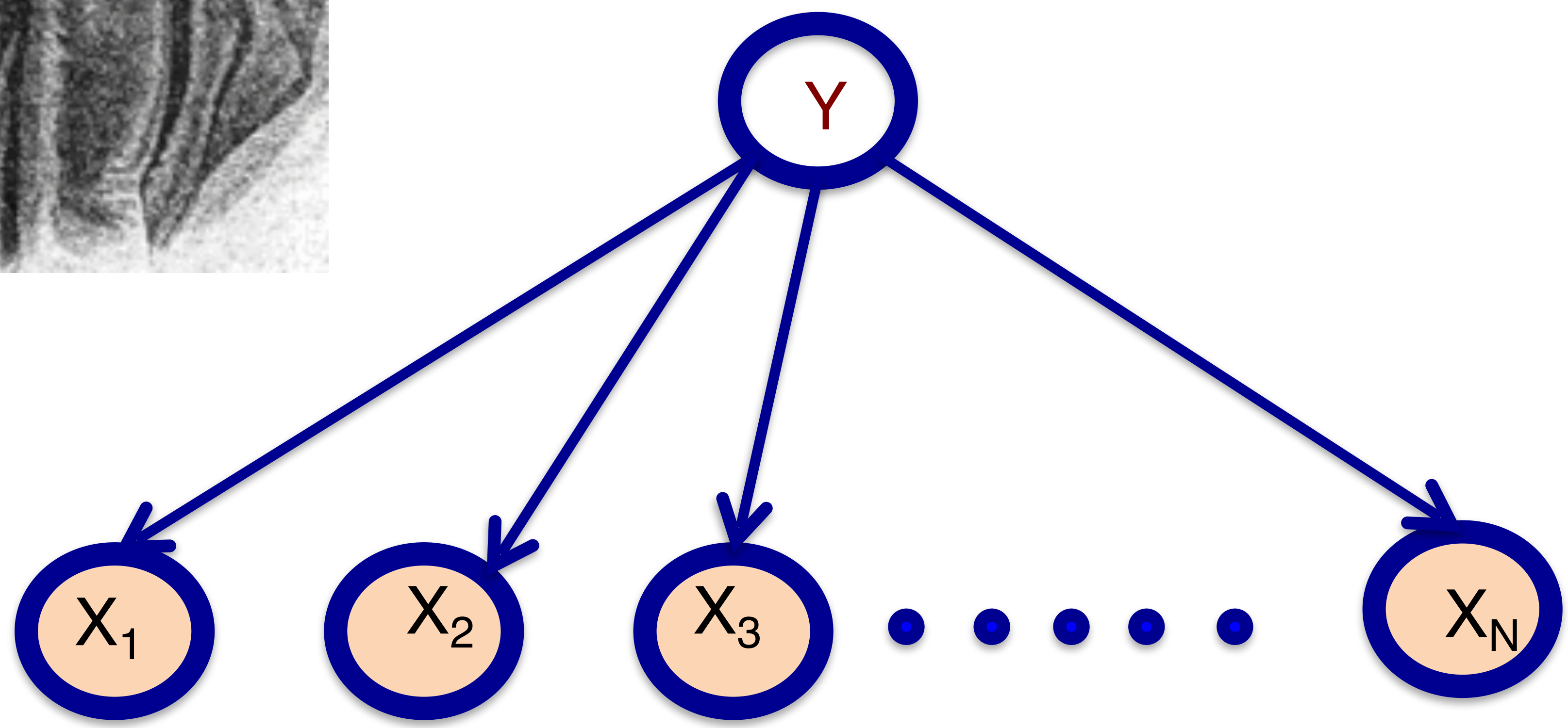
Without conditional independence assumption?

$2\,(2^n-1)$

With conditional independence assumption?

$2n$

NAÏVE



Suppose we have
n variables

Conditionally
Independent

$$P(X_1, X_2, ..., X_n|Y) = \prod_i P(X_i|Y)$$

Naïve Bayes in a Nutshell

Bayes rule:

$$P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) P(X_1 \ldots X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1 \ldots X_n | Y = y_j)}$$

Assuming conditional independence among X_i's:

$$P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

So, classification rule for $X^{new} = < X_1, \ldots, X_n >$ is:

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

Naïve Bayes Algorithm

for each value y_k :

estimate $\pi_k \equiv P(Y = y_k)$

for each value x_{ij} of each attribute X_i :

estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

TRAIN

Classify X^{new}

$$Y^{new} \leftarrow \text{argmax}_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \text{argmax}_{y_k} \pi_k \prod_i \theta_{ijk}$$

TEST

- In class we constructed a Naïve Bayes classifier to predict if a someone lives in Squirrel Hill based on variables such as:
  - Driving to CMU
  - Shop at Giant Eagle
  - Even # letters in last name
- Note that we used all discrete variables
- We are not limited to discrete X_i!

What if we have continuous X_i ?

Gaussian Naïve Bayes (GNB): assume

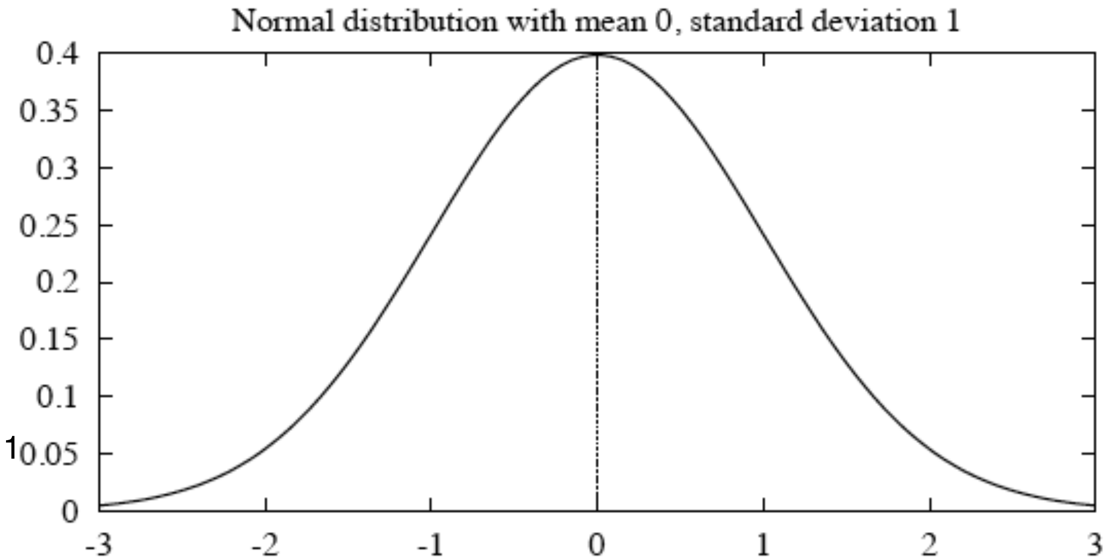
$$p(X_i = x | Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \; e^{-\frac{1}{2}\left(\frac{x-\mu_{ik}}{\sigma_{ik}}\right)^2}$$

Sometimes assume variance

- is independent of Y (i.e., σ_i),
- or independent of X_i (i.e., σ_k)
- or both (i.e., σ)

Gaussian Distribution
(also called “Normal”)

p(x) is a *probability density function*, whose integral (not sum) is 1



$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The probability that X will fall into the interval (a, b) is given by

$$\int_a^b p(x)dx$$

- Expected, or mean value of X, $E[X]$, is

$$E[X] = \mu$$

- Variance of X is

$$Var(X) = \sigma^2$$

- Standard deviation of X, σ_X, is

$$\sigma_X = \sigma$$

Gaussian Naïve Bayes Algorithm

for each value y_k :

estimate $\pi_k \equiv P(Y = y_k)$

For each attribute X_i :

estimate $P(X_i | Y = y_k)$

class conditional mean and variance μ_{ik}, σ_{ik}

TRAIN

Classify X^{new}

$$Y^{new} \leftarrow \text{argmax}_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \text{argmax}_{y_k} \pi_k \prod_i N(X_i^{new}; \mu_{ik}, \sigma_{ik})$$

TEST

Naïve Bayes for Popularity Prediction

IsPopular	Daily Tweets	Has Facebook?	Years left to graduation	ML Grades
no	***85***	***0***	***2***	***85***
no	80	1	2	90
yes	***83***	***0***	***3***	***86***
yes	70	0	1	96
yes	***68***	***0***	***1***	***80***
no	65	1	1	70
yes	***64***	***1***	***3***	***65***
no	72	0	2	95
yes	***69***	***0***	***2***	***70***
yes	75	0	1	80
yes	***75***	***1***	***2***	***70***
yes	72	1	3	90
yes	***81***	***0***	***3***	***75***
no	71	1	1	91

Continuous

Naïve Bayes Classification Rule

$$\underbrace{Y^{new} \leftarrow}_{\text{New example}} \mathrm{argmax}_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

Classify:

X^{new} = < lefttoGraduation = 3, DailyTweets = 60,
HasFacebook = 0, MLGrade = 62>

For continuous variables assume Gaussian Distribution

- P(Daily Tweets I IsPopular) = N(μ,σ^2)

Mean depends on clas
variable and X_i

$$P(\underbrace{X_i = x_{ij}}_{i^{th}\text{ variable taking the } x_{ij}}|\underbrace{Y = y_k}_{\text{Class variable Y taking the } y_k}) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} exp\left(\frac{-(x_{ij} - \mu_{ik})^2}{2\sigma_{ik}^2}\right)$$

Variance depends on
class variable and X_i

Training

Statistics 101

$$\hat{\mu} = \frac{\sum_{l=1}^{D} X_l}{D}$$

$$\hat{\sigma}^2 = \frac{\sum_{l=1}^{D} (X_l - \mu)}{D - 1}$$

μ_dailytweets_yes=? σ_dailytweets_yes=?

μ_dailytweets_no=? σ_dailytweets_no=?

μ_Mlgrades_yes=? σ_MLgrades_yes=?

μ_Mlgrades_no=? σ_Mlgrades_no=?

Training

Statistics 101

$$\hat{\mu} = \frac{\sum_{l=1}^{D} X_l}{D}$$

$$\hat{\sigma}^2 = \frac{\sum_{l=1}^{D} (X_l - \mu)}{D - 1}$$

μ_dailytweets_yes=73 σ_dailytweets_yes=6.2

μ_dailytweets_no=74.6 σ_dailytweets_no=8

μ_Mlgrades_yes=79.1 σ_MLgrades_yes=10.2

μ_Mlgrades_no=86.2 σ_Mlgrades_no=9.7

X^{new} = < lefttoGraduation = 3, DailyTweets = 60, HasFacebook = 0, MLGrade = 62>

P(IsPopular = yes)=

P(IsPopular = no)=

P(HasFacebook=0|IsPopular = yes)=

P(HasFacebook=0|IsPopular = yes)=

P(HasFacebook=1IsPopular = no)=

P(HasFacebook=1IsPopular = no)=

P(YearsLeft=1|IsPopular = yes)=

P(YearsLeft=1IsPopular = no)=

P(YearsLeft=2|IsPopular = yes)=

P(YearsLeft=2IsPopular = no)=

P(YearsLeft=3|IsPopular = yes)=

P(YearsLeft=3IsPopular = no)=

P(DailyTweets=60|IsPopular = yes)=

P(DailyTweets=60|IsPopular = no)=

P(MLGrade=62|IsPopular = yes)=

P(MLGrade=60|IsPopular = no)=

X^{new} = < lefttoGraduation = 3, DailyTweets = 60, HasFacebook = 0, MLGrade = 62>

P(IsPopular = yes)= 0.643

P(IsPopular = no)=0.357

P(HasFacebook=0|IsPopular = yes)=0.667

P(HasFacebook=0|IsPopular = yes)=

P(YearsLeft=3|IsPopular = yes)=0.286

P(YearsLeft=3IsPopular = no)=0

P(DailyTweets=60|IsPopular = yes)=0.0071

P(DailyTweets=60|IsPopular = no)=0.0094

P(MLGrade=62|IsPopular = yes)=0.0096

P(MLGrade=60|IsPopular = no)=0.0018

A lot of problems…

- In the training data, there are no yearsLeftGraduation=3|IsPopular=No
- What if the test example has yearsLeftGraduation=4
- It is not in the training data!
- We need to smooth or regularize the estimates to avoid overfitting training dataset

- MLE estimates

$$\theta_{MLE} = \text{argmax}_\theta P(X|\theta)$$

- MAP estimates

$$\theta_{MAP} = \text{argmax}_\theta P(\theta|X)$$

$$\theta_{MAP} = \text{argmax}_\theta P(X|\theta)P(\theta)$$

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

Prior to avoid 0 probabilities

Text Classification

- Y discrete valued.
  - e.g., Spam or not
- $X = <X_1, X_2, \ldots X_n> =$ document

I am pleased to announce that Bob Frederking of the Language Technologies Institute is our new Associate Dean for Graduate Programs. In this role, he oversees the many issues that arise with our multiple masters and PhD programs. Bob brings to this position considerable experience with the masters and PhD programs in the LTI.

I would like to thank Frank Pfenning, who has served ably in this role for the past two years.

****

Randal E. Bryant
Dean and University Professor

- X_i is a random variable describing…

Answer 2:

- X_i represents the i^{th} *word position* in document
- X_1 = “I”, X_2 = “am”, X_3 = “pleased”
- and, let’s assume the X_i are iid (indep, identically distributed)

HW2 Dataset: Reuters Corpus


```
<REUTERS TOPICS=''YES'' LEWISSPLIT=''TRAIN''
CGISPLIT=''TRAINING-SET'' OLDID=''12981'' NEWID=''798''>
<DATE> 2-MAR-1987 16:51:43.42</DATE>
<TOPICS><D>livestock</D><D>hog</D></TOPICS>
<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>
<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Pork
Congress kicks off tomorrow, March 3, in Indianapolis with 160
of the nations pork producers from 44 member states determining
industry positions on a number of issues, according to the
National Pork Producers Council, NPPC.
Delegates to the three day Congress will be considering 26
resolutions concerning various issues, including the future
direction of farm policy and the tax law as it applies to the
agriculture sector. The delegates will also debate whether to
endorse concepts of a national PRV (pseudorabies virus) control
and eradication program, the NPPC said. A large
trade show, in conjunction with the congress, will feature
the latest in technology in all areas of the industry, the NPPC
added. Reuter
\&\#3;</BODY></TEXT></REUTERS>
```

Questions

Acknowledgements

- Some slides are taken from Tom Mitchell's 10-601 lecture notes