Bayesian Learning

An Overview Of Naïve Bayes

- **Bayes rule :**

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

- **Assuming conditional independence among X_i's:**

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

- **So, classification rule for** $X^{new} = \langle X_1, \ldots, X_n \rangle$ **is :**

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

Naïve Bayes Algorithm – discrete X_i

- **Train Naïve Bayes (examples)**

  for each* value y_k

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

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

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

- **Classify (X^{new})**

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

$$Y^{new} \leftarrow \arg\max_{y_k} \pi_k \prod_i \theta_{ijk}$$

*probabilities must sum to 1, so need estimate only n-1 of these ...

Estimating Parameters : *Y, Xi* discrete-valued

- Maximum likelihood estimates (MLE's)

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\# D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} \mid Y = y_k) = \frac{\# D\{X_i = x_{ij} \wedge Y = y_k\}}{\# D\{Y = y_k\}}$$

Number of items in dataset D for which *Y* = *y*$_k$

Naïve Bayes : Sublety #1

If unlucky, our MLE estimte for $P(X_i \mid Y)$ might be zero (e.g., X_i = Birthday_Is_January_30_1990)

- **Why worry about just one parameter out of many?**

- **What can be done to avoid this ?**

Estimating Parameters

- Maximum Likelihood Estimate (MLE) : choose θ that maximize probability of observed data D

$$\hat{\theta} = \arg\max_{\theta} P(D \mid \theta)$$

- Maximum a Posteriori (MAP)P estimatee: choose that is most probable given prior probability and the data

$$\hat{\theta} = \arg\max_{\theta} P(\theta \mid D)$$

$$= \arg\max_{\theta} = \frac{P(D \mid \theta) P(\theta)}{P(D)}$$

Estimating Parameters : Y, X_i discrete-valued

- Maximum likelihood estimates :

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\# D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} \mid Y = y_k) = \frac{\# D\{X_i = x_{ij} \wedge Y = y_k\}}{\# D\{Y = y_k\}}$$

- MAP estimates (Beta, Dirichlet priors)

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\# D\{Y = y_k\} + \alpha_k}{|D| + \sum_m \alpha_m}$$

Only differnce :
“imaginary” examples

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} \mid Y = y_k) = \frac{\# D\{X_i = x_{ij} \wedge Y = y_k\} + \alpha'_k}{\# D\{Y = y_k\} + \sum_m \alpha'_m}$$

Another way to view Naïve Bayes (Boolean Y) :

Decision rule : is this quantity greater or less than 1 ?

$$\frac{P(Y=1 \mid X_1 \ldots X_n)}{P(Y=0 \mid X_1 \ldots X_n)} = \frac{P(Y=1)\prod_i P(X_i \mid Y=1)}{P(Y=0)\prod_i P(X_i \mid Y=0)}$$

Naïve Bayes : classifying text documents

- Classify which email are spam ?
- Classify which emails promise an attchment ?

> 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

How shall we represent text documents for Naïve Bayes ?

Learning to classify documents : P $(Y|X)$

- **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 1 : X_i is boolean, 1 if word **i** is in document, else 0

- e.g., $X_{\textbf{pleased}} = 1$

Issues ?

Learning to classify documents : P $(Y|X)$

- **Y discrete valued.**
  - e.g., Spam or not
- **X = <X_1, X_2, …, 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)

$$P(X_i \mid Y) = P(X_j \mid Y)(\forall_i, j)$$

Learning to classify documents : P $(Y|X)$ the "Bag of Words" model

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

- **X_i is a random variables. Each represents the word at its position i in the document**
- Generating a document according to this distribution = rolling a 50,000 sided die, once for each word position in the document

- The observed counts for each word follow a ??? distribution

Multinomial Distribution

- P(θ) and P(θ|D) have the same form

Eg. 2 Dice roll problem (6 outcomes instead of 2)
Likelihood is ~ Multinomial ($\theta = \{\theta_1, \theta_2, ..., \theta_k\}$)

$$P(D \mid \theta) = \theta_1^{\alpha 1} \theta_2^{\alpha 2} \ldots \theta_k^{\alpha k}$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\prod_{i=1}^{k} \theta_i^{\beta_i - 1}}{B(\beta_1, \ldots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \ldots, \beta_k)$$

Then posterior id Dirichlet distribution

$$P(\theta \mid D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$$

For Multinomial, conjugate prior is Dirichlet distribution.

Multinomial Bag of Words

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

Naïve Bayes Algorithm – discrete X_i

- **Train Naïve Bayes (examples)**

  for each* value y_k

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

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

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

  prob that word x_{ij} appears in position i, given $Y = y_k$

- **Classify (X^{new})**

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

$$Y^{new} \leftarrow \arg\max_{y_k} \pi_k \prod_i \theta_{ijk}$$

*Additional assumption : word probability are position independent

$$\theta_{ijk} = \theta_{mjk} \quad for \quad i \neq m$$

MAP estimates for bag of words

- MAP estimate for multinomial

$$\theta_i = \frac{\alpha_i + \beta_i - 1}{\sum_{m=1}^{k} \alpha_m + \sum_{m=1}^{k} \beta_m - 1}$$

$$P(X_i = \text{aardvark}) = \frac{\#\text{observed} \quad \text{'aardvark'} + \#\text{hallucinated} \quad \text{'aardvark'} - 1}{\#\text{observed} \quad \text{words} + \#\text{hallucinated} \quad \text{words} - k}$$

- What β's should we choose ?

Twenty Newsgroups

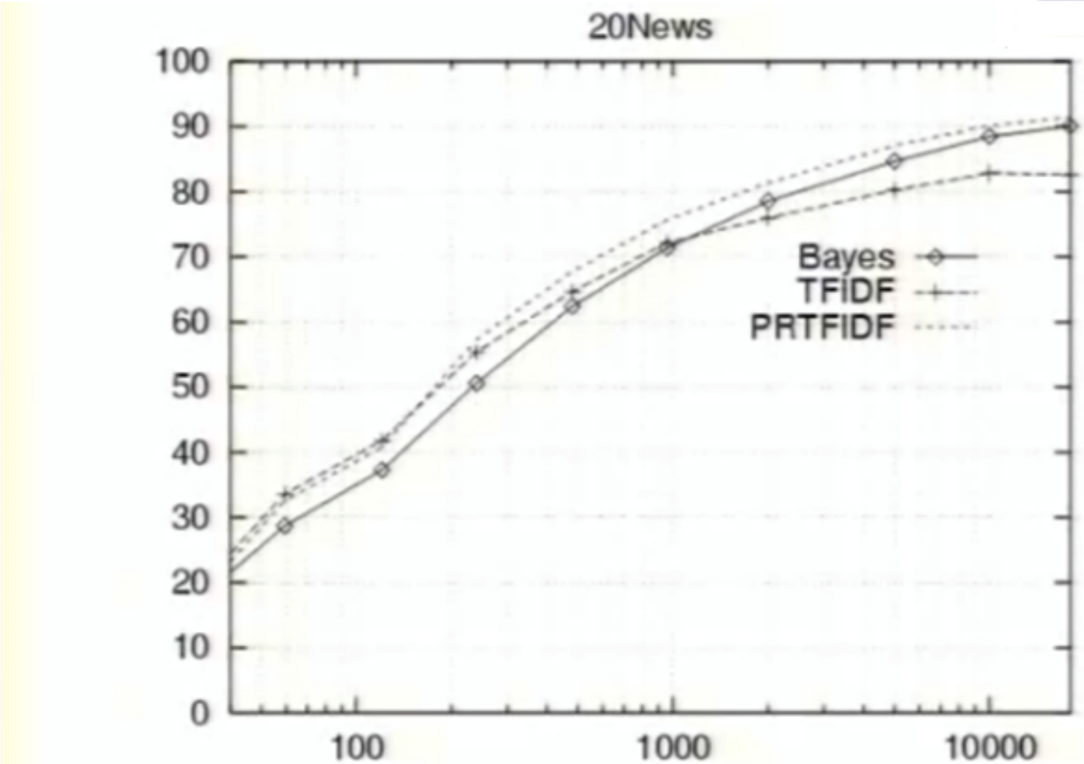
Given 1000 training documents from each group
Learn to classify new documents according to which newsgroup it came from

comp.graphics	misc.forsale
comp.os.ms-windows.misc	rec.autos
comp.sys.ibm.pc.hardware	rec.motorcycles
comp.sys.mac.hardware	rec.sport.baseball
comp.windows.x	rec.sport.hockey
alt.atheism	sci.space
soc.religion.christian	sci.crypt
talk.religion.misc	sci.electronics
talk.politics.mideast	sci.med
talk.politics.misc	
talk.politics.guns	

Naive Bayes: 89% classification accuracy

Learning Curve for 20 Newsgroups

For code and data, see
www.cs.cmu.edu/~tom/mlbook.html
click on "Software and Data"



Accuracy vs. Training set size (1/3 withheld for test)

What if we have continuous X_i ?

- Eg., image classification : X_i is real-valued i^{th} pixel

- Naïve Bayes requires P(X_i | Y=y_k), but X_i is real (continuous)

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

- Common approach : assume P(X_i | Y=y_k) follows a Normal (Gaussian) distribution

What if we have continuous X_i ?

- Eg., image classification : X_i is i^{th} pixel

- Gaussian Naïve Bayes : assume

$$P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

- Sometimes assume σ_{ik}
  - is independent of Y (i.e., σ_i)
  - or independent of X_i (i.e., σ_k)
  - Or both (i.e., σ)

Gaussian Naïve Bayes Algorithm – continuous X_i (but still discrete Y)

- **Train Naïve Bayes (examples)**

  for each* value y_k
  estimate $\pi_k \equiv P(Y = y_k)$
  for each* value x_{ij} of each attribute X_i
  estimate $\theta_{ijk} \equiv P(X_i = x_{ij} \mid Y = y_k)$

- **Classify (X^{new})**

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

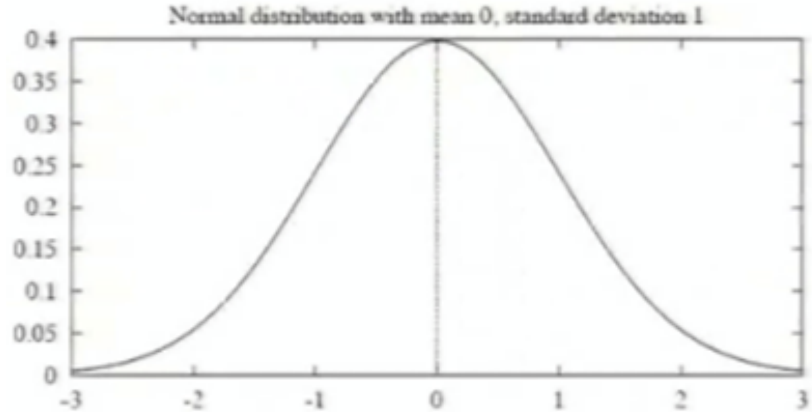
$$Y^{new} \leftarrow \arg\max_{y_k} \quad \pi_k \prod_i Normal(X_i^{new}, \mu_{ik}, \sigma_{ik})$$

*probabilities must sum to 1, so need estimate only n-1 parameters …

Gaussian Distribution

(also called "Normal")

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



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

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

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

- Expected, or menu 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 – Big Picture

Consider boolean Y, continuous X_i, Assume P(Y=1)=0.5

$$Y^{new} \leftarrow \arg\max_{y \in \{0,1\}} P(Y = y) \prod_i P(X_i^{new} \mid Y = y)$$

What is thr minimum possible error ?

Best case :

- conditional independence assumption is satistied
- we know P(Y), P(X|Y) perfectly (e.g., infinite training data)