



Naïve Bayes Classifier

Learning the Optimal Classifier

Task: Predict whether or not a picnic spot is enjoyable

Training Data: $X = (X_1 \quad X_2 \quad X_3 \quad \dots \quad \dots \quad X_d) \quad Y$

	Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
n rows	Sunny	Warm	Normal	Strong	Warm	Same	Yes
	Sunny	Warm	High	Strong	Warm	Same	Yes
	Rainy	Cold	High	Strong	Warm	Change	No
	Sunny	Warm	High	Strong	Cool	Change	Yes

Lets learn $P(Y|X)$ – how many parameters?

Prior: $P(Y = y)$ for all y

$K-1$ if K labels


Likelihood: $P(X=x|Y = y)$ for all x, y

$(2^d - 1)K$ if d binary features

Learning the Optimal Classifier

Task: Predict whether or not a picnic spot is enjoyable

Training Data: $X = (X_1 \quad X_2 \quad X_3 \quad \dots \quad \dots \quad X_d) \quad Y$

n rows 

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

Lets learn $P(Y|X)$ – how many parameters?

$2^d K - 1$ (K classes, d binary features)

Need $n \gg 2^d K - 1$ number of training data to learn all parameters



Conditional Independence

- X is **conditionally independent** of Y given Z:

probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall x, y, z) P(X = x | Y = y, Z = z) = P(X = x | Z = z)$$

- Equivalent to:

$$P(X, Y | Z) = P(X | Z)P(Y | Z)$$

- e.g., $P(\text{Thunder} | \text{Rain}, \text{Lightning}) = P(\text{Thunder} | \text{Lightning})$

Note: does NOT mean Thunder is independent of Rain

Prediction using Conditional Independence

- Predict Lightning
- From two **conditionally Independent** features
 - Thunder
 - Rain

parameters needed to learn likelihood given L

$$P(T, R | L) \quad (2^2 - 1)2 = 6$$

With conditional independence assumption

$$P(T, R | L) = P(T | L) P(R | L) \quad (2 - 1)2 + (2 - 1)2 = 4$$



Naïve Bayes Assumption

- Naïve Bayes assumption:
 - Features are independent given class:

$$\begin{aligned}P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2|Y) \\ &= P(X_1|Y)P(X_2|Y)\end{aligned}$$

- More generally:

$$P(X_1 \dots X_d|Y) = \prod_{i=1}^d P(X_i|Y)$$

- How many parameters now? **$(2-1)dK$ vs. $(2^d-1)K$**
 - Suppose \mathbf{X} is composed of d binary features



Naïve Bayes Classifier

- Given:
 - Class Prior $P(Y)$
 - d conditionally independent features \mathbf{X} given the class Y
 - For each X_i , we have likelihood $P(X_i | Y)$

- Decision rule:

$$\begin{aligned} f_{NB}(\mathbf{x}) &= \arg \max_y P(x_1, \dots, x_d | y) P(y) \\ &= \arg \max_y \prod_{i=1}^d P(x_i | y) P(y) \end{aligned}$$

- If conditional independence assumption holds, NB is optimal classifier! But worse otherwise.

Naïve Bayes Algo – Discrete features

8

- Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^n$ $X^{(j)} = (X_1^{(j)}, \dots, X_d^{(j)})$

- Maximum Likelihood Estimates

- For Class Prior

$$\hat{P}(y) = \frac{\{\#j : Y^{(j)} = y\}}{n}$$

- For Likelihood

$$\frac{\hat{P}(x_i, y)}{\hat{P}(y)} = \frac{\{\#j : X_i^{(j)} = x_i, Y^{(j)} = y\}/n}{\{\#j : Y^{(j)} = y\}/n}$$

- NB Prediction for test data $X = (x_1, \dots, x_d)$

$$Y = \arg \max_y \hat{P}(y) \prod_{i=1}^d \frac{\hat{P}(x_i, y)}{\hat{P}(y)}$$

Subtlety 1 – Violation of NB Assumption

- Usually, features are not conditionally independent:

$$P(X_1 \dots X_d | Y) \neq \prod_i P(X_i | Y)$$

- Actual probabilities $P(Y|X)$ often biased towards 0 or 1 (Why?)
- Nonetheless, NB is the single most used classifier out there
 - NB often performs well, even when assumption is violated
 - [Domingos & Pazzani '96] discuss some conditions for good performance

Subtlety 2 – Insufficient training data

- What if you never see a training instance where $X_1=a$ when $Y=b$?
 - e.g., $Y=\{\text{SpamEmail}\}$, $X_1=\{\text{'Earn'}\}$
 - $P(X_1=a \mid Y=b) = 0$
- Thus, no matter what the values X_2, \dots, X_d take:
 - $P(Y=b \mid X_1=a, X_2, \dots, X_d) = 0$

$$P(X_1 = a, X_2 \dots X_n | Y) = P(X_1 = a | Y) \prod_{i=2}^d P(X_i | Y)$$

- What now???



MLE vs. MAP Review

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

What if we toss the coin too few times?

- You say: Probability next toss is a head = 0
- Billionaire says: You're fired! ...with prob 1 😊

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

- Beta prior equivalent to extra coin flips
- As $N \rightarrow \infty$, prior is “forgotten”
- **But, for small sample size, prior is important!**

Naïve Bayes Algo – Discrete features

- Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^n$ $X^{(j)} = (X_1^{(j)}, \dots, X_d^{(j)})$
- Maximum A Posteriori Estimates – add m “virtual” examples

Assume priors

$$Q(Y = b)$$

$$Q(X_i = a, Y = b)$$

MAP Estimate

$$\hat{P}(X_i = a | Y = b) = \frac{\{\#j : X_i^{(j)} = a, Y^{(j)} = b\} + mQ(X_i = a, Y = b)}{\{\#j : Y^{(j)} = b\} + \underbrace{mQ(Y = b)}_{\substack{\text{\# virtual examples} \\ \text{with } Y = b}}}$$

Now, even if you never observe a class/feature posterior probability never zero.



Case Study: Text Classification

- Classify e-mails
 - $Y = \{\text{Spam}, \text{NotSpam}\}$
- Classify news articles
 - $Y = \{\text{what is the topic of the article?}\}$
- Classify webpages
 - $Y = \{\text{Student, professor, project, ...}\}$
- What about the features **X**?
 - The text!

Features X are entire document —¹⁴ X_i for i^{th} word in article



Article from rec.sport.hockey

Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.e
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year's biggest and worst (opinic
Date: 5 Apr 93 09:53:39 GMT

I can only comment on the Kings, but the most obvious candidate for pleasant surprise is Alex Zhitnik. He came highly touted as a defensive defenseman, but he's clearly much more than that. Great skater and hard shot (though wish he were more accurate). In fact, he pretty much allowed the Kings to trade away that huge defensive liability Paul Coffey. Kelly Hrudey is only the biggest disappointment if you thought he was any good to begin with. But, at best, he's only a mediocre goaltender. A better choice would be Tomas Sandstrom, though not through any fault of his own, but because some thugs in Toronto decided



NB for Text Classification

- $P(\mathbf{X}|Y)$ is huge!!!
 - Article at least 1000 words, $\mathbf{X}=\{X_1, \dots, X_{1000}\}$
 - X_i represents i^{th} word in document, i.e., the domain of X_i is entire vocabulary, e.g., Webster Dictionary (or more), 10,000 words, etc.
- NB assumption helps a lot!!!
 - $P(X_i=x_i|Y=y)$ is just the probability of observing word x_i at the i^{th} position in a document on topic y

$$h_{NB}(\mathbf{x}) = \arg \max_y P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$



Bag of words model

- Typical additional assumption – **Position in document doesn't matter**: $P(X_i=x_i | Y=y) = P(X_k=x_i | Y=y)$
 - “Bag of words” model – order of words on the page ignored
 - Sounds really silly, but often works very well!

$$\prod_{i=1}^{LengthDoc} P(x_i|y) = \prod_{w=1}^W P(w|y)^{count_w}$$

When the lecture is over, remember to wake up the person sitting next to you in the lecture room.



Bag of words model

- Typical additional assumption – **Position in document doesn't matter**: $P(X_i=x_i | Y=y) = P(X_k=x_i | Y=y)$
 - “Bag of words” model – order of words on the page ignored
 - Sounds really silly, but often works very well!

$$\prod_{i=1}^{LengthDoc} P(x_i|y) = \prod_{w=1}^W P(w|y)^{count_w}$$

in is lecture lecture next over person remember room
sitting the the the to to up wake when you

Bag of words approach

the world of

TOTAL



all about the company

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.

► All About The Company

- Global Activities
- Corporate Structure
- TOTAL's Story
- Upstream Strategy
- Downstream Strategy
- Chemicals Strategy
- TOTAL Foundation
- Homepage

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

NB with bag of words for text classification¹⁹

- Learning phase:
 - Class Prior $P(Y)$
 - $P(X_i | Y)$
- Test phase:
 - For each document
 - Use naïve Bayes decision rule

$$h_{NB}(\mathbf{x}) = \arg \max_y P(y) \prod_{i=1}^{LengthDoc} P(x_i | y)$$



Twenty news groups results

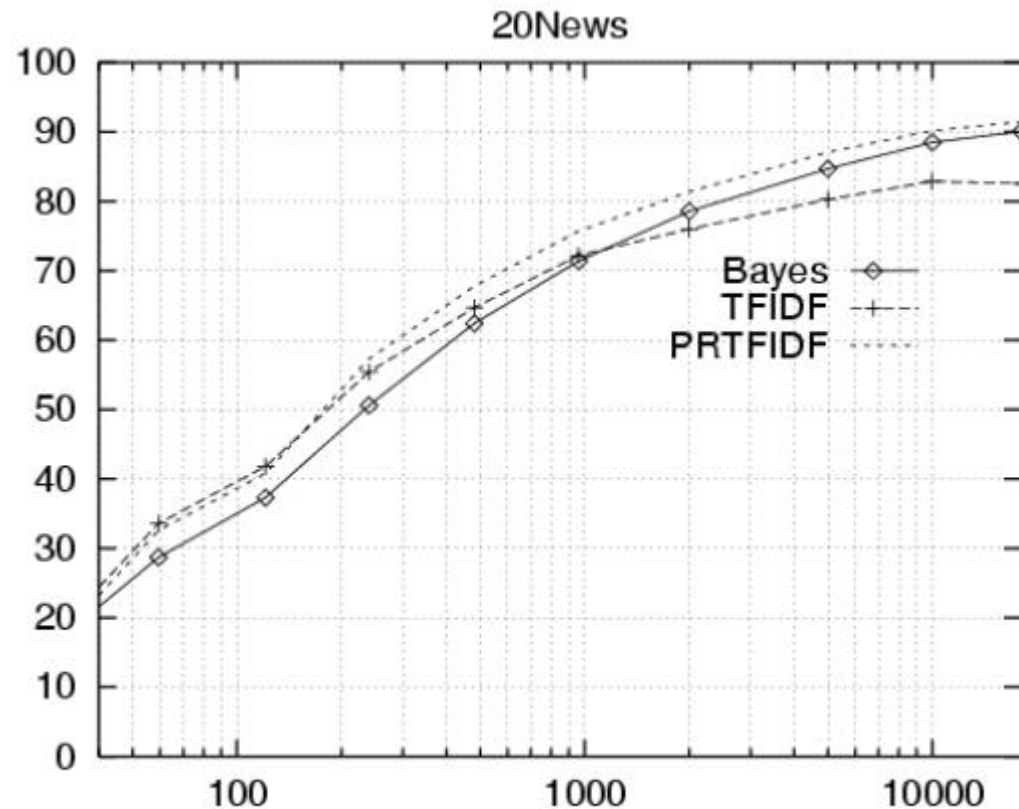
Given 1000 training documents from each group
Learn to classify new documents according to
which newsgroup it came from

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 twenty news groups



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

What if features are continuous?

Eg., character recognition: X_i is intensity at i^{th} pixel



Gaussian Naïve Bayes (GNB):

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

Different mean and variance for each class k and each pixel i .

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)

Estimating parameters:

Y discrete, X_i continuous

Maximum likelihood estimates:

$$\hat{\mu}_{MLE} = \frac{1}{N} \sum_{j=1}^N x_j$$

$$\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k)$$

i^{th} pixel in
 j^{th} training image

k^{th} class

j^{th} training image

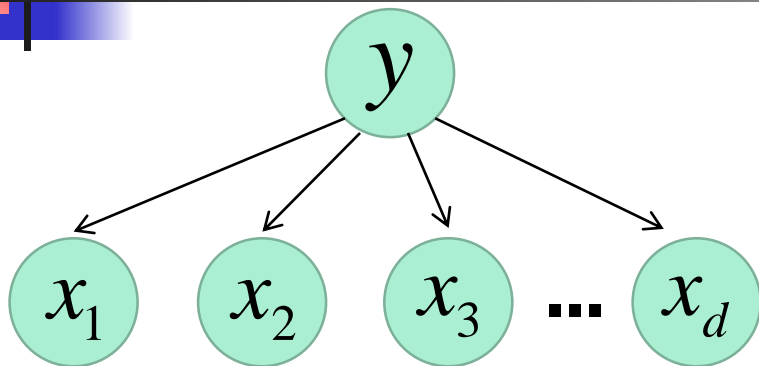
$$\hat{\sigma}_{unbiased}^2 = \frac{1}{N-1} \sum_{j=1}^N (x_j - \hat{\mu})^2$$

$$\hat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k) - 1} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)$$

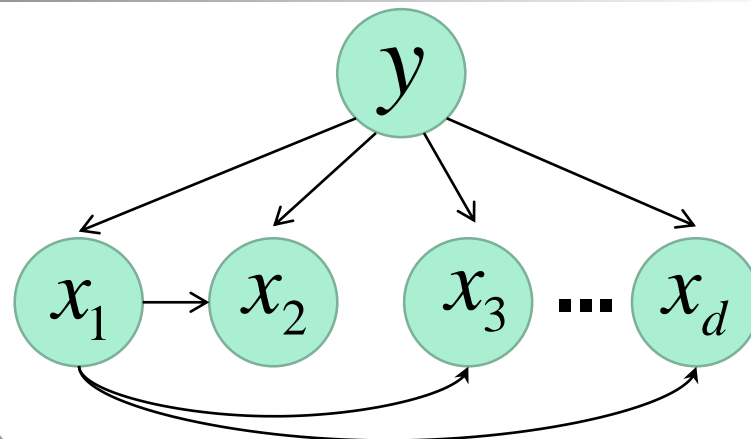


Semi-naïve Bayes classifiers

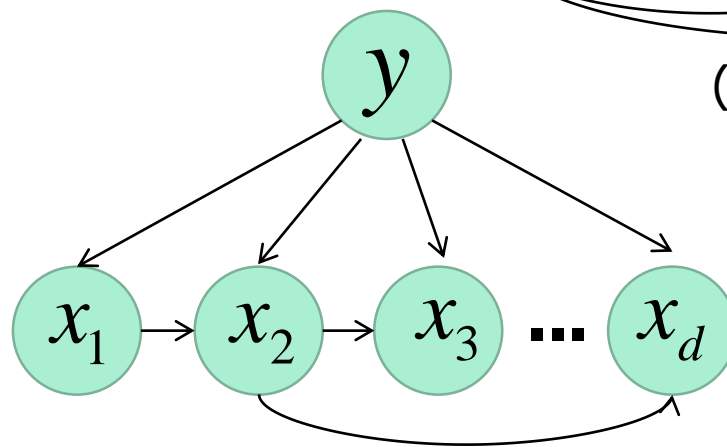
- maintain strong attribute dependencies
- One-Dependent Estimator
 - Each attribute is assumed to rely on at most one other attribute except the category
 - Super-Parent ODE
 - Tree Augmented naïve Bayes
 - Averaged One-Dependent Estimator



(a)NB



(b)SPODE



(c)TAN



What you should know...

- Optimal decision using Bayes Classifier
- Naïve Bayes classifier
 - What's the assumption
 - Why we use it
 - How do we learn it
 - Why is Bayesian estimation important
- Text classification
 - Bag of words model
- Gaussian NB
 - Features are still conditionally independent
 - Each feature has a Gaussian distribution given class