

# 10-315 Intro to Machine Learning (PhD)

## Lecture 5: Naive Bayes

Leila Wehbe

Carnegie Mellon University

Machine Learning Department

Reading: **Generative and Discriminative Classifiers** by Tom Mitchell.

LECTURE OUTCOMES:

- Conditional Independence
- Naïve Bayes, Gaussian Naive Bayes
- Practical Examples

Links (use the version you need)

- **Notebook**
- **PDF slides**

Assume you want to build a classifier for new customers

<i>O</i> Is older than 35 years	<i>I</i> Has Personal Income	<i>S</i> Is a Student	<i>J</i> Birthday before July 1st	<i>Y</i> Buys computer
0	1	0	0	0
0	1	0	1	0
1	1	0	1	1
1	1	0	1	1
1	0	1	0	1
1	0	1	1	0
0	0	1	0	1
0	1	0	0	0
0	0	1	1	1
0	1	1	0	1
0	0	1	1	1
1	1	0	1	1
0	1	1	0	1
1	1	0	0	0

What to predict for the next customer? We want to find  $P(Y = 0 | O = 0, I = 0, S = 1, J = 1)$

<i>O</i> Is older than 35 years	<i>I</i> Has Personal Income	<i>S</i> Is a Student	<i>J</i> Birthday before July 1st	<i>Y</i> Buys computer
0	1	1	1	?

How many parameters must we estimate?

Suppose  $X = (X_1, X_2)$  where  $X_i$  and  $Y$  are boolean random variables

How many parameters do we need to estimate to know  $P(Y|X_1, X_2)$ ?

How many parameters must we estimate?

Suppose  $X = (X_1, X_2)$  where  $X_i$  and  $Y$  are boolean random variables  
How many parameters do we need to estimate to know  $P(Y|X_1, X_2)$ ?

$X_1$	$X_2$	$P(Y = 1 \mid X_1, X_2)$	$P(Y = 0 \mid X_1, X_2)$
0	1	0.1	0.9
1	0	0.24	0.76
0	1	0.54	0.46
1	1	0.23	0.77

4 parameters:  $P(Y = 0 \mid X_1, X_2) = 1 - P(Y = 1 \mid X_1, X_2)$

How many parameters must we estimate?

Suppose  $X = (X_1, X_2, \dots, X_d)$  where  $X_i$  and  $Y$  are boolean random variables

How many parameters do we need to estimate to know  $P(Y|X_1, \dots, X_d)$ ?

How many parameters must we estimate?

Suppose  $X = (X_1, X_2, \dots, X_d)$  where  $X_i$  and  $Y$  are boolean random variables

How many parameters do we need to estimate to know  $P(Y|X_1, \dots, X_d)$ ?

$X_1$	$X_2$	$X_3$	...	$X_d$	$P(Y = 1   X)$	$P(Y = 0   X)$
0	0	0	...	0	0.1	0.9
0	0	0	...	1	0.24	0.76
...	...	...	...	...	...	...
...	...	...	...	...	...	...
1	1	1	...	1	0.52	0.48

How many parameters must we estimate?

Suppose  $X = (X_1, X_2, \dots, X_d)$  where  $X_i$  and  $Y$  are boolean random variables  
 How many parameters do we need to estimate to know  $P(Y|X_1, \dots, X_d)$ ?

$X_1$	$X_2$	$X_3$	...	$X_d$	$P(Y = 1   X)$	$P(Y = 0   X)$
0	0	0	...	0	0.1	0.9
0	0	0	...	1	0.24	0.76
...	...	...	...	...	...	...
...	...	...	...	...	...	...
1	1	1	...	1	0.52	0.48

$2^d$  rows! ( $2^d$  parameters!).

If we have 30 boolean  $X_i$ 's:  $\sim 1$  billion rows!



Last Lecture we asked:

Can we just estimate  $P(Y|X)$  in this fashion and be done?

We might not have enough data. For example consider having 100 attributes of people:

- how many rows will we have?  $2^{100} > 10^{30}$
- how many people on earth?  $< 10^{10}$
- 99.99% of rows will not have training examples!

Can we use Bayes Rule to reduce the number of parameters?

We used bayes rule last lecture to express marginal and conditional probabilities of the data and parameter  $\theta$ :

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

We can also express the conditional distribution of  $Y$  given  $X$ :

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

BTW, this notation is a shorthand for:

$$(\forall i, j) P(Y = y_i|X = x_j) = \frac{P(X = x_j|Y = y_i)P(Y = y_i)}{P(X = x_j)}$$

Equivalently:

$$(\forall i, j) P(Y = y_i|X = x_j) = \frac{P(X = x_j|Y = y_i)P(Y = y_i)}{\sum_k P(X = x_i|Y = y_k)P(Y = y_k)}$$

Does Bayes Rule help reduce the number of parameters?

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- $P(X)$  is the same for both classes, so we might be able to avoid computing it (also, I can get it for free if I already know  $P(X|Y)$  and  $P(Y)$ ).
- $P(Y)$  is only 1 parameter.
- What about  $P(X|Y)$ ? i.e.  $P(X_1, X_2, \dots, X_d|Y)$ . How many parameters do we need?

How many parameters do we need for  $P(X_1, X_2, \dots, X_d | Y)$ ?

$Y$	$X_1$	$X_2$	$X_3$	...	$X_d$	$P(X   Y)$
0	0	0	0	...	0	0.1
0	0	0	0	...	1	0.02
0	...	...	...	...	...	...
0	1	1	1	...	1	0.16
1	0	0	0	...	0	0.03
1	0	0	0	...	1	0.2
1	...	...	...	...	...	...
1	1	1	1	...	1	0.01

- How many parameters for the red cells?  $2^d - 1$
- Should I compute parameters for the blue cells as well? yes,  $2^d - 1$

Does Bayes Rule help reduce the number of parameters?

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- $P(X)$  is the same for both classes, so we might be able to avoid computing it (also, I can get it for free if I already know  $P(X|Y)$  and  $P(Y)$ ).
- $P(Y)$  is only 1 parameter.
- $P(X_1, X_2, \dots, X_d|Y)$ :  $2(2^d - 1)$  parameters

**Still too many parameters!**

If we have 30 boolean  $X_i$ 's:  $\sim 2$  billion!

Solution:

- 1- Be smart about how to estimate probabilities from sparse data
  - maximum likelihood estimates
  - maximum a posteriori estimates
- 2- Be smart about how to represent joint distributions
  - Bayes networks, graphical models, conditional independence

Be smart about how to represent joint distributions

CONDITIONAL INDEPENDENCE:

**Definition:**  $X$  is conditionally independent of  $Y$  given  $Z$ , if the probability distribution governing  $X$  is independent of the value of  $Y$ , given the value of  $Z$ .

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

- Which we often write:

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

- For example:  $P(\text{thunder}|\text{rain}, \text{lightning}) = P(\text{thunder}|\text{lightning})$ 
  - Thunder is independent of rain **given lightning**.
  - Once we know there if there is or there is lightning, no more information is provided by the value of rain.
  - **This does not mean that thunder is independent of rain.**

## The Naïve Bayes Algorithm

- Naïve Bayes is a classifier that assumes conditional independence of the variables  $X_i$  given the label  $Y$ .
  - For example:  $P(X_1|X_2, Y) = P(X_1|Y)$
- How does this assumption simplify  $P(X_1, X_2|Y)$ ?

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y) \quad (\text{chain rule}) \quad (1)$$

$$= P(X_1|Y)P(X_2|Y) \quad (\text{conditional independence}) \quad (2)$$

- In general:

$$P(X_1, \dots, X_d|Y) = \prod_i P(X_i|Y) \quad (3)$$

- How many parameters do we need to describe  $P(X_1 \dots X_n|Y)$ ?
  - Without the conditional independence assumption:  $2(2^d - 1)$
  - With the conditional independence assumption:



## The Naïve Bayes Algorithm

- Naïve Bayes is a classifier that assumes conditional independence of the variables  $X_i$  given the label  $Y$ .
  - For example:  $P(X_1|X_2, Y) = P(X_1|Y)$
- How does this assumption simplify  $P(X_1, X_2|Y)$ ?

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y) \quad (\text{chain rule}) \quad (4)$$

$$= P(X_1|Y)P(X_2|Y) \quad (\text{conditional independence}) \quad (5)$$

- In general:

$$P(X_1, \dots, X_d|Y) = \prod_i P(X_i|Y) \quad (6)$$

- How many parameters do we need to describe  $P(X_1 \dots X_n|Y)$ ?
  - Without the conditional independence assumption:  $2(2^d - 1)$
  - With the conditional independence assumption:  $2d$

## The Naïve Bayes Algorithm

- Naïve Bayes assumes conditional independence of the  $X_i$ 's:

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

(more on this assumption soon!)

- Using Bayes rule with that assumption:

$$P(Y = y_k | X_1, \dots, X_d) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{P(X)}$$

- Train the algorithm
  - estimate  $P(X_i | Y = y_k)$  and  $P(Y = y_k)$
- To classify, pick the most probable  $Y^{\text{new}}$  for a new sample  $X^{\text{new}} = (X_1^{\text{new}}, X_2^{\text{new}}, \dots, X_d^{\text{new}})$  as:

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

## Naïve Bayes - Training and Prediction Phase - Discrete $x_i$

### Training:

- Estimate  $\pi_k \equiv P(Y = y_k)$ , get  $\hat{\pi}_k$
- Estimate  $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$ , get  $\hat{\theta}_{ijk}$ 
  - $\theta_{ijk}$  is estimated for each label  $y_k$ :
    - For each variable  $X_i$ :
      - For each value  $x_{ij}$  that  $X_i$  can take.
  - Example: if  $X_1$  is binary,  $P(X_1 | Y = 0)$  is a bernouilli distribution, where:
    - the probability of  $X_1 = 0$  given  $Y = 0$  is  $\theta_{100}$
    - the probability of  $X_1 = 1$  given  $Y = 0$  is  $\theta_{110}$
    - $\theta_{100} = 1 - \theta_{110}$ .

### Prediction: Classify $Y^{\text{new}}$

$$Y^{\text{new}} = \underset{y_k}{\operatorname{argmax}} P(Y = y_k) \prod_i P(X_i^{\text{new}} = x_i^{\text{new}} | Y = y_k) \quad (7)$$

$$= \underset{y_k}{\operatorname{argmax}} \hat{\pi}_k \prod_i \hat{\theta}_{i, X_i^{\text{new}}, k} \quad (8)$$

But... how do we estimate these parameters?

## Naïve Bayes - Training Phase - Discrete $X_i$ - Maximum (Conditional) Likelihood Estimation

- $P(X|Y = y_k)$  has parameters  $\theta_{ijk}$ , one for each value  $x_{ij}$  of each  $X_i$ .  $P(Y)$  has parameters  $\pi_k$ .
- To follow the MLE principle, we pick the parameters  $\pi$  and  $\theta$  that maximizes the (**conditional**) likelihood of the data given the parameters.
- To estimate:
  - Compute

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

- For each label  $y_k$ :
  - For each variable  $X_i$ :
    - For each value  $x_{ij}$  that  $X_i$  can take, compute:

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

Let's train!

<i>O</i> Is older than 35 years	<i>I</i> Has Personal Income	<i>S</i> Is a Student	<i>J</i> Birthday before July 1st	<i>Y</i> Buys computer
0	1	0	0	0
0	1	0	1	0
1	1	0	1	1
1	1	0	1	1
1	0	1	0	1
1	0	1	1	0
0	0	1	0	1
0	1	0	0	0
0	0	1	1	1
0	1	1	0	1
0	0	1	1	1
1	1	0	1	1
0	1	1	0	1
1	1	0	0	0

Let's train!

Removed one variable to simplify the problem + changed order of samples.

O= Is older than 35, S= Is a student, J = Birthday before July 1, Y= buys computer

<i>O</i>	<i>S</i>	<i>J</i>	<i>Y</i>
0	0	0	0
0	0	1	0
0	0	0	0
1	1	1	0
1	0	0	0
0	1	0	1
1	0	1	1
1	0	1	1
1	1	0	1
0	1	1	1
0	1	0	1
0	1	1	1
1	0	1	1
0	1	0	1

Let's train!

O= Is older than 35, S= Is a student, J = Birthday before July 1, Y= buys computer

Estimated Parameters		Data			
		O	S	J	Y
		0	0	0	0
		0	0	1	0
		0	0	0	0
$Y = 1$	$Y = 0$	1	1	1	0
$P(Y=1) =$	$P(Y=0) =$	1	0	0	0
$P(O=1 Y=1) =$	$P(O=1 Y=0) =$				
$P(O=0 Y=1) =$	$P(O=0 Y=0) =$	0	1	0	1
$P(S=1 Y=1) =$	$P(S=1 Y=0) =$	1	0	1	1
$P(S=0 Y=1) =$	$P(S=0 Y=0) =$	1	0	1	1
$P(J=1 Y=1) =$	$P(J=1 Y=0) =$	1	1	0	1
$P(J=0 Y=1) =$	$P(J=0 Y=0) =$	0	1	1	1
		0	1	0	1
		0	1	1	1
		1	0	1	1
		0	1	0	1

Let's train!

O= Is older than 35, S= Is a student, J = Birthday before July 1, Y= buys computer

Estimated Parameters		Data			
		<i>O</i>	<i>S</i>	<i>J</i>	<i>Y</i>
		0	0	0	0
		0	0	1	0
		0	0	0	0
<i>Y</i> = 1	<i>Y</i> = 0	1	1	1	0
$P(Y=1) = 9/14$	$P(Y=0) = 5/14$	1	0	0	0
$P(O=1 Y=1) =$	$P(O=1 Y=0) =$				
$P(O=0 Y=1) =$	$P(O=0 Y=0) =$	0	1	0	1
$P(S=1 Y=1) =$	$P(S=1 Y=0) =$	1	0	1	1
$P(S=0 Y=1) =$	$P(S=0 Y=0) =$	1	0	1	1
$P(J=1 Y=1) =$	$P(J=1 Y=0) =$	1	1	0	1
$P(J=0 Y=1) =$	$P(J=0 Y=0) =$	0	1	1	1
		0	1	0	1
		0	1	1	1
		1	0	1	1
		0	1	0	1



Let's train!

O= Is older than 35, S= Is a student, J = Birthday before July 1, Y= buys computer

Estimated Parameters		Data			
		O	S	J	Y
<div> <div>Y = 1</div> <div> <div>P(Y=1) = 9/14</div> <div>P(O=1 Y=1) = 4/9</div> <div>P(O=0 Y=1) = 5/9</div> <div>P(S=1 Y=1) = 6/9</div> <div>P(S=0 Y=1) = 3/9</div> <div>P(J=1 Y=1) = 5/9</div> <div>P(J=0 Y=1) = 4/9</div> </div> </div> <div> <div>Y = 0</div> <div> <div>P(Y=0) = 5/14</div> <div>P(O=1 Y=0) = 2/5</div> <div>P(O=0 Y=0) = 3/5</div> <div>P(S=1 Y=0) = 1/5</div> <div>P(S=0 Y=0) = 4/5</div> <div>P(J=1 Y=0) = 2/5</div> <div>P(J=0 Y=0) = 3/5</div> </div> </div>		0	0	0	0
		0	0	1	0
		0	0	0	0
		1	1	1	0
		1	0	0	0
		0	1	0	1
		1	0	1	1
		1	0	1	1
		1	1	0	1
		0	1	1	1
		0	1	0	1
		0	1	1	1
		1	0	1	1
		0	1	0	1

Let's predict!

O= Is older than 35, S= Is a student, J = Birthday before July 1, Y= buys computer

Estimated Parameters

New Data

$Y = 1$	$Y = 0$				
$P(Y=1) = 9/14$	$P(Y=0) = 5/14$				
$P(O=1 Y=1) = 4/9$	$P(O=1 Y=0) = 2/5$				
$P(O=0 Y=1) = 5/9$	$P(O=0 Y=0) = 3/5$	$O$	$S$	$J$	$Y$
$P(S=1 Y=1) = 6/9$	$P(S=1 Y=0) = 1/5$	0	1	1	?
$P(S=0 Y=1) = 3/9$	$P(S=0 Y=0) = 4/5$				
$P(J=1 Y=1) = 5/9$	$P(J=1 Y=0) = 2/5$				
$P(J=0 Y=1) = 4/9$	$P(J=0 Y=0) = 3/5$				

Let's predict!

O= Is older than 35, S= Is a student, J = Birthday before July 1, Y= buys computer

Estimated Parameters		New Data			
$Y = 1$		$Y = 0$			
$P(Y=1) = 9/14$	$P(Y=0) = 5/14$				
$P(O=1 Y=1) = 4/9$	$P(O=1 Y=0) = 2/5$				
$P(O=0 Y=1) = 5/9$	$P(O=0 Y=0) = 3/5$	$O$	$S$	$J$	$Y$
$P(S=1 Y=1) = 6/9$	$P(S=1 Y=0) = 1/5$	0	1	1	?
$P(S=0 Y=1) = 3/9$	$P(S=0 Y=0) = 4/5$				
$P(J=1 Y=1) = 5/9$	$P(J=1 Y=0) = 2/5$				
$P(J=0 Y=1) = 4/9$	$P(J=0 Y=0) = 3/5$				

$$Y^{\text{new}} = \underset{y_k}{\operatorname{argmax}} P(Y = y_k) P(O = 0, S = 1, J = 1 | Y = y_k) \quad (9)$$

$$= \underset{y_k}{\operatorname{argmax}} P(Y = y_k) P(O = 0 | Y = y_k) P(S = 1 | Y = y_k) P(J = 1 | Y = y_k) \quad (10)$$

$$P(Y=1) P(O=0 | Y=1) P(S = 1 | Y=1) P(J = 1 | Y=1) = 9/14 * 5/9 * 6/9 * 5/9 = 0.132$$

$$P(Y=0) P(O=0 | Y=0) P(S = 1 | Y=0) P(J = 1 | Y=0) = 5/14 * 3/5 * 1/5 * 2/5 = 0.017$$

Pick label 1!

Can also compute  $P(Y|X)$ :

Not required to make predictions, but we have everything we need:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$$P(X) = \sum_k P(X|Y = y_k)P(Y = y_k)$$

$$P(Y=1) P(O=0| Y=1)P(S = 1 | Y=1) P( J = 1 | Y=1) = 0.132$$

$$P(Y=0) P(O=0| Y=0)P(S = 1 | Y=0) P( J = 1 | Y=0) = 0.017$$

$$P(Y=1 | O=0, S = 1, J = 1 ) = 0.886$$

$$P(Y=0 | O=0, S = 1, J = 1 ) = 0.114$$

## Classification Accuracy

- To get an estimate of generalization performance, compute accuracy on held-out set (more about this soon).
  - Never train on your test data!
- Assume you train and use this algorithm to predict "Buy Computer?" with a large dataset, and obtain binary classification accuracy of 75%.
  - What does this mean? Is it good or bad?

## Classification Accuracy

- What if 70% is the probability of  $Y = 1$ . Is 75% impressive?
  - What is an easy way to obtain 70% classification accuracy?
- What is chance performance?
- What is the accuracy if I flip an unbiased coin?
- If  $P(Y=1) > 0.5$ , then we can just predict 1 all the time!
  - What will be the accuracy?
- What happens if you predict  $Y=1$  with probability 0.7  $\Rightarrow$  this is called probability matching in cognitive science

## Naïve Bayes observation 1

- Usually the  $X_i$  are not conditionally independent:

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

- Even if the "naïve" conditional independence assumption is not true in the data, Naïve Bayes might still perform well and is used anyways
  - often the right classification, even when not the right probability (see [Domingos&Pazzani, 1996])
- To see the effect of the violation of this assumption, consider the extreme case in which  $X_i$  is a copy of  $X_k$ . What is the effect on  $P(Y|X)$ ?

- To see the effect of the violation of the conditional independence assumption, consider the extreme case in which  $X_i$  is a copy of  $X_k$ . What is the effect on  $P(Y|X)$ ?

$$P(Y=1) P(O=0|Y=1) P(S=1|Y=1) P(J=1|Y=1)$$

---


$$P(Y=1) P(O=0|Y=1) P(S=1|Y=1) P(J=1|Y=1) + P(Y=0) P(O=0|Y=0) P(S=1|Y=0) P(J=1|Y=0)$$



- To see the effect of the violation of the conditional independence assumption, consider the extreme case in which  $X_i$  is a copy of  $X_k$ . What is the effect on  $P(Y|X)$ ?

$$P(Y=1) P(O=0|Y=1) P(O'=0|Y=1) P(S=1|Y=1) P(J=1|Y=1)$$


---

$$P(Y=1) P(O=0|Y=1) P(O'=0|Y=1) P(S=1|Y=1) P(J=1|Y=1) + \dots$$

$$\dots P(Y=0) P(O=0|Y=0) P(O'=0|Y=0) P(S=1|Y=0) P(J=1|Y=0)$$

Consider for example that  $P(O = 1|Y = 1) > P(O = 1|Y = 0)$ , how does  $P(Y = 1|O = 1, O' = 1, S = 1, J = 1)$  with the duplicated variable compare to respect  $P(Y = 1|O = 1, S = 1, J = 1)$ ?

## Naïve Bayes observation 2

What if we have an irrelevant variable?

$$P(Y=1) P(O=0|Y=1) P(S=1|Y=1) P(J=1|Y=1)$$

---

$$P(Y=1) P(O=0|Y=1) P(S=1|Y=1) P(J=1|Y=1) + P(Y=0) P(O=0|Y=0) P(S=1|Y=0) P(J=1|Y=0)$$

Assume  $J$  is independent of  $Y$ :  $P(J|Y = 0) = P(J|Y = 1) = P(J)$

Does it hurt classification?

- If we have the correct estimates, then performance is not affected.
- If we have noisy estimates, performance is affected.

### Naïve Bayes observation 3

- Another way to view Naïve Bayes with Boolean  $Y$  and  $X_i$ s is:
- Decision rule: is this quantity greater or less than 1?

$$\frac{P(Y = 1|X_1 \dots X_d)}{P(Y = 0|X_1 \dots X_d)} = \frac{P(Y = 1)P(X_1 \dots X_d|Y = 1)}{P(Y = 0)P(X_1 \dots X_d|Y = 0)} \quad > \text{ or } < 1? \quad (11)$$

- Practical concern: What happens when  $d$  is large?

### Naïve Bayes observation 3

- Another way to view Naïve Bayes with Boolean  $Y$  and  $X_i$ s is:
- Decision rule: is this quantity greater or less than 1?

$$\frac{P(Y = 1|X_1 \dots X_d)}{P(Y = 0|X_1 \dots X_d)} = \frac{P(Y = 1)P(X_1 \dots X_d|Y = 1)}{P(Y = 0)P(X_1 \dots X_d|Y = 0)} \quad > \text{ or } < 1? \quad (12)$$

- Taking the log of this ratio prevents **underflow** and expresses the decision rule in a useful way (we will see later more reasons why it's useful).

$$\ln \frac{P(Y = 1|X_1 \dots X_d)}{P(Y = 0|X_1 \dots X_d)} = \ln \frac{P(Y = 1)}{P(Y = 0)} + \sum_i \ln \frac{P(X_i|Y = 1)}{P(X_i|Y = 0)} \quad > \text{ or } < 0? \quad (13)$$

- Since  $X_i$ s are boolean, we can simplify the notation:
  - $\theta_{ik} = \hat{P}(X_i = 1|Y = k)$  and  $1 - \theta_{ik} = \hat{P}(X_i = 0|Y = k)$

$$\ln \frac{P(Y = 1|X_1 \dots X_d)}{P(Y = 0|X_1 \dots X_d)} = \ln \frac{P(Y = 1)}{P(Y = 0)} + \sum_i \ln \frac{P(X_i|Y = 1)}{P(X_i|Y = 0)} \quad (14)$$

$$= \ln \frac{P(Y = 1)}{P(Y = 0)} + \sum_i X_i \ln \frac{\theta_{i1}}{\theta_{i0}} + \sum_i (1 - X_i) \ln \frac{1 - \theta_{i1}}{1 - \theta_{i0}} \quad (15)$$

## Naïve Bayes observation 4

- If unlucky, our MLE estimate for  $P(X_i|Y = y_k)$  might be zero.
  - for example,  $X_i = \text{birthdate}$ .  $x_i = \text{Jan\_25\_1992}$ .

- Why worry about just one parameter out of many?

$$P(Y = y_k)P(X_1|Y = y_k)P(X_2|Y = y_k) \dots P(X_d|Y = y_k)$$

- What happens if one of the  $P(X_i|Y = y_k)$  is zero?
  - What can be done to address this?

## Naïve Bayes - Training Phase - Discrete $X_i$

### METHOD 1: MAXIMUM (CONDITIONAL) LIKELIHOOD ESTIMATION

- $P(X|Y = y_k)$  has parameters  $\theta_{ijk}$ , one for each value  $x_{ij}$  of each  $X_i$ .
- To follow the MLE principle, we pick the parameters  $\theta$  that maximizes the **conditional** likelihood of the data given the parameters.

### METHOD 2: MAXIMUM A POSTERIORI PROBABILITY ESTIMATION

- To follow the MAP principle, pick the parameters  $\theta$  with maximum posterior probability given the conditional likelihood of the data and the prior on  $\theta$ .

## Naïve Bayes - Training Phase - Discrete $X_i$

### METHOD 1: MAXIMUM (CONDITIONAL) LIKELIHOOD ESTIMATION

To estimate:

- Compute

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

- For each label  $y_k$ :
  - For each variable  $X_i$ :
    - For each value  $x_{ij}$  that  $X_i$  can take, compute:

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

## METHOD 2: MAXIMUM A POSTERIORI PROBABILITY ESTIMATION (BETA OR DIRICHLET PRIORS)

- $K$ : the number of values  $Y$  can take
- $J$ : the number of values  $X$  can take (we assume here that all  $X_i$  have the same number of possible values, but this can be changed)
- Example prior for  $\pi_k$  where  $K > 2$ :
  - Dirichlet( $\beta_\pi, \beta_\pi, \dots, \beta_\pi$ ) prior. (optionally, you can choose different values for each parameter to encode a different weighting).
  - if  $K = 2$  this becomes a Beta prior.
- Example prior for  $\theta_{ijk}$  where  $J > 2$ :
  - Dirichlet( $\beta_\theta, \beta_\theta, \dots, \beta_\theta$ ) prior. (optionally, you can choose different values for each parameter to encode a different weighting, you can choose a different prior per  $X_i$  or even per label  $y_k$ ).
  - if  $J = 2$  this becomes a Beta prior.



## METHOD 2: MAXIMUM A POSTERIORI PROBABILITY ESTIMATION (BETA OR DIRICHLET PRIORS)

- $K$ : the number of values  $Y$  can take
- $J$ : the number of values  $X$  can take

These priors will act as imaginary examples that smooth the estimated distributions and prevent zero values.

To estimate:

- Compute

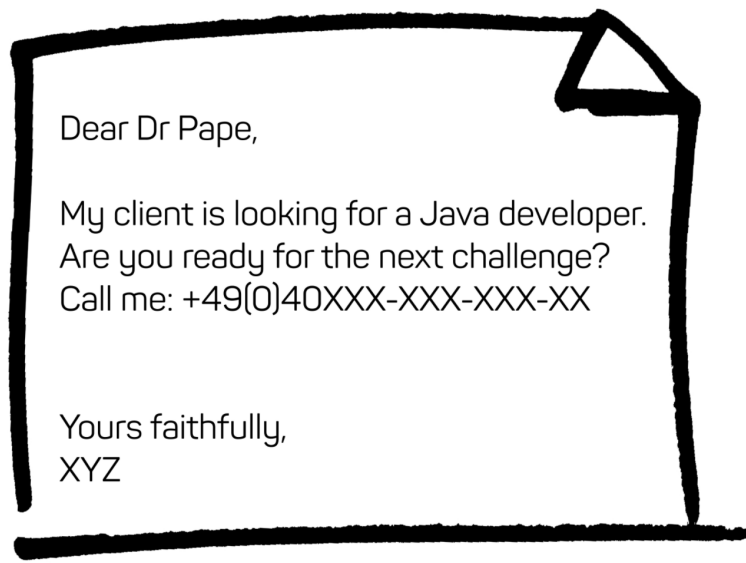
$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D(Y = y_k) + (\beta_\pi - 1)}{|D| + K(\beta_\pi - 1)}$$

- For each label  $y_k$ :
  - For each variable  $X_i$ :
    - For each value  $x_{ij}$  that  $X_i$  can take, compute:

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D(X_i = x_{ij} \wedge Y = y_k) + (\beta_\theta - 1)}{\#D(Y = y_k) + J(\beta_\theta - 1)}.$$

## Example: Text classification

- Classify which emails are spam?



**SPAM**

**vs.**



**HAM**

### image by Daniel Pape

- Classify which emails promise an attachment?
- Classify which web pages are student home pages?

How shall we represent text documents for Naïve Bayes?

How can we express  $X$ ?

- $Y$  discrete valued. e.g., Spam or not
- $X = ?$

How can we express  $X$ ?

- $Y$  discrete valued. e.g., Spam or not
- $X = (X_1, X_2, \dots, X_d)$  with  $d$  the number of words in English.

What are the limitations with this representation?

How can we express  $X$ ?

- $Y$  discrete valued. e.g., Spam or not
- $X = (X_1, X_2, \dots, X_d)$  with  $d$  the number of words in English.

What are the limitations with this representation?

- Some words always present
- Some words very infrequent
- Doesn't count how often a word appears
- Conditional independence assumption is false...

## Alternative Featurization

- $Y$  discrete valued. e.g., Spam or not
- $X = (X_1, X_2, \dots, X_d) = \text{document}$
- $X_i$  is a random variable describing the word at position  $i$  in the document
- possible values for  $X_i$  : any word  $w_k$  in English
- $X_i$  represents the  $i$ th word position in document
- $X_1 = \text{"I"}, X_2 = \text{"am"}, X_3 = \text{"pleased"}$

How many parameters do we need to estimate  $P(X|Y)$ ? (say 1000 words per document, 10000 words in english)

## Alternative Featurization

- $Y$  discrete valued. e.g., Spam or not
- $X = (X_1, X_2, \dots, X_d) = \text{document}$
- $X_i$  is a random variable describing the word at position  $i$  in the document
- possible values for  $X_i$  : any word  $w_k$  in English
- $X_i$  represents the  $i$ th word position in document
- $X_1 = \text{"I"}, X_2 = \text{"am"}, X_3 = \text{"pleased"}$

How many parameters do we need to estimate  $P(X|Y)$ ? (say 1000 words per document, 10000 words in english)

**Conditional Independence Assumption** very useful:

- reduce problem to only computing  $P(X_i|Y)$  for every  $X_i$ .

## "Bag of Words" model

- Additional assumption: position doesn't matter!
  - (this is not true of language, but can be a useful assumption for building a classifier)
- assume the  $X_i$  are IID:

$$P(X_i|Y) = P(X_j|Y)(\forall i, j)$$

- we call this "Bag of Words"





Art installation in Gates building (now removed)

## "Bag of Words" model

- Additional assumption: position doesn't matter!
  - (this is not true of language, but can be a useful assumption for building a classifier)
- assume the  $X_i$  are IID:

$$P(X_i|Y) = P(X_j|Y)(\forall i, j)$$

- we call this "Bag of Words"
- **Since all  $X_i$ s have the same distribution, we only have to estimate one parameter per word, per class.**
- **$P(X|Y = y_k)$  is a multinomial distribution:**

$$P(X|Y = y_k) \propto \theta_{1k}^{\alpha_{1k}} \theta_{2k}^{\alpha_{2k}} \dots \theta_{dk}^{\alpha_{dk}}$$

## Review of distributions

$$P(X_i = w_j) \begin{cases} \theta_1, & \text{if } X_i = w_1 \\ \theta_2, & \text{if } X_i = w_2 \\ \dots & \\ \theta_k, & \text{if } X_i = w_K \end{cases} \quad (16)$$

Probability of observing a document with  $\alpha_1$  count of  $w_1$ ,  $\alpha_2$  count of  $w_2$  ... is a multinomial:

$$\frac{|D|!}{\alpha_1! \dots \alpha_J!} \theta_1^{\alpha_1} \theta_2^{\alpha_2} \theta_3^{\alpha_3} \dots \theta_J^{\alpha_J}$$

## Review of distributions

Dirichlet Prior examples:

- if constant across classes and words:

$$P(\theta) = \frac{\theta^{\beta_\theta} \theta^{\beta_\theta}, \dots \theta^{\beta_\theta}}{\text{Beta}(\beta_\theta, \beta_\theta, \dots, \beta_\theta)}$$

- if constant across classes but different for different words:

$$P(\theta) = \frac{\theta^{\beta_1} \theta^{\beta_2}, \dots \theta^{\beta_J}}{\text{Beta}(\beta_1, \beta_2, \dots, \beta_J)}$$

- if different for different classes  $k$  and words:

$$P(\theta_k) = \frac{\theta^{\beta_{1k}} \theta^{\beta_{2k}}, \dots \theta^{\beta_{Jk}}}{\text{Beta}(\beta_{1k}, \beta_{2k}, \dots, \beta_{Jk})}$$

MAP estimates for Bag of words:

(Dirichlet is the conjugate prior for a multinomial likelihood function)

$$\theta_{jk} = \frac{\alpha_{jk} + \beta_{jk} - 1}{\sum_m (\alpha_{mk} + \beta_{mk} - 1)} \quad (17)$$

Again the prior acts like hallucinated examples

What  $\beta$ s should we choose?

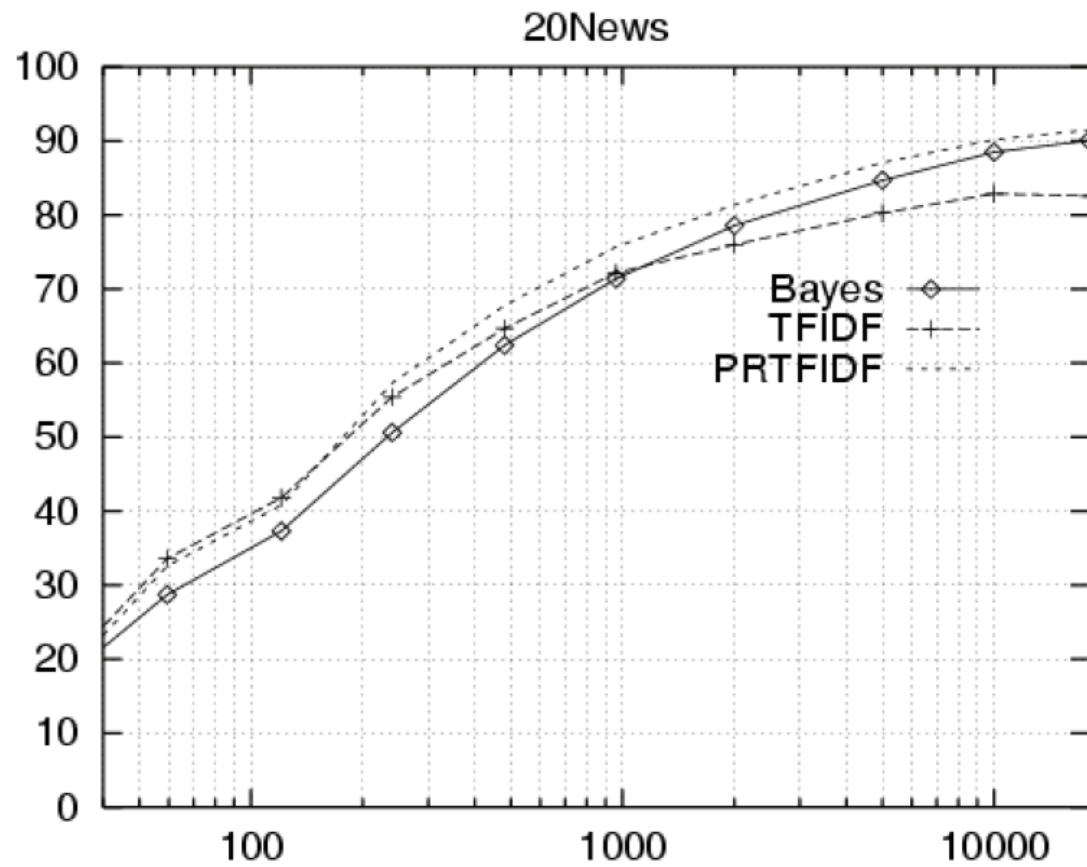
## Example: Twenty NewsGroups

For code and data, see [www.cs.cmu.edu/~tom/mlbook.html](http://www.cs.cmu.edu/~tom/mlbook.html) click on "Software and Data".

Can group labels into groups that share priors:

- comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.max.hardware, comp.windows.x
- misc.forsale
- rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey
- alt.atheism,
- soc.religion.christian,
- talk.religion.misc, talk.politics.mideast, talk.politics.misc, talk.politics.guns,
- sci.space, sci.crypt, sci.electronics, sci.med
- Naïve Bayes: 89% classification accuracy

## Learning curve for 20 Newsgroups



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

Even if incorrect assumption, performance can be very good

Even when taking half of the email

- Assumption doesn't hurt the particular problem?
- Redundancy?
- Leads less examples to train? Converges faster to asymptotic performance? (Ng and Jordan)



Even if incorrect assumption, performance can be very good

Even when taking half of the email

- Assumption doesn't hurt the particular problem?
- Redundancy?
- Leads less examples to train? Converges faster to asymptotic performance? (Ng and Jordan)

More recently, algorithms such as Transformers have become very good

- are able to capture the sequential aspect of language and produce more complex representations.
- They do have many parameters, but nowhere as much as we mentioned before ( $10000^{1000}$ ).