

Naïve Bayes Classifier

Intro to ML 466/566
Winter 2021

Administrivia

- Project Proposals due Thursday
- Assignment 1 out today, due in 2 weeks
 - Please start early
- Labs
 - Most weeks are just additional TA-led office hours
 - As deadlines approach, we may have content for labs
 - Also planning to have one virtual TA the lab before assignments/exams
 - Details forthcoming

From a while back...

- Estimating parameters
- Using Bayes Rule to incorporate prior beliefs
 - smoothing

Joint distribution

- Probability of >1 thing happening at the same time
 - Probability it will rain today and I forgot my umbrella
 - $P(\text{rain}=\text{true}, \text{umbrella}=\text{false})$

The Joint Distribution

Example: Boolean variables A,
B, C

Recipe for making a joint distribution of M
variables:

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).

A	B	C
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

The Joint Distribution

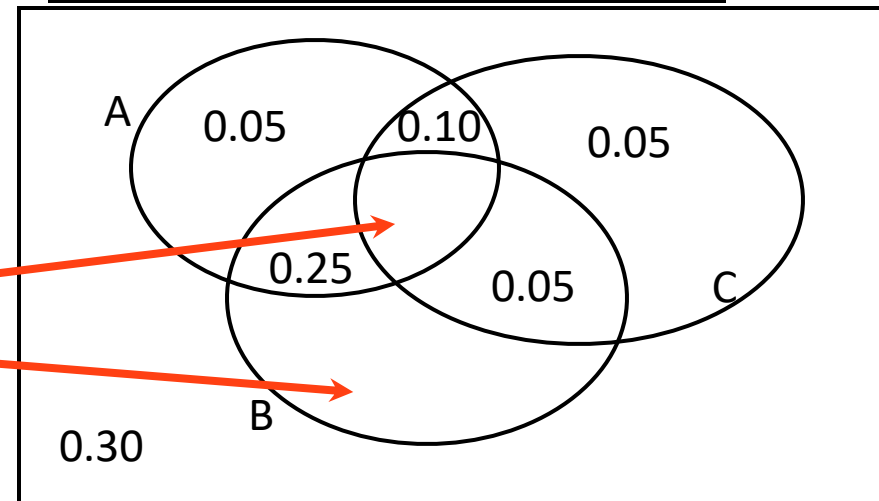
Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:









1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

What goes here?



Joint Probability Distribution

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0245895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

Once you have the joint distribution, you can ask for the probability of any logical expression involving your attribute

Using the Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

$$\begin{aligned} P(\text{Poor}) &= 0.2531 + 0.0422 + 0.3313 + 0.1341 \\ &= 0.7607 \end{aligned}$$

Inference with the Joint

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	<div></div>
		rich	0.0245895	<div></div>
	v1:40.5+	poor	0.0421768	<div></div>
		rich	0.0116293	<div></div>
Male	v0:40.5-	poor	0.331313	<div></div>
		rich	0.0971295	<div></div>
	v1:40.5+	poor	0.134106	<div></div>
		rich	0.105933	<div></div>

$$P(E_1 | E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

$$\begin{aligned}
 P(\text{hours_worked} > 40.5 \mid \text{Poor}) &= P(\text{hours_worked} > 40.5 \ \& \ \text{Poor}) / P(\text{Poor}) \\
 &= (0.2531 + 0.1341) / 0.7607 \\
 &= 0.5090
 \end{aligned}$$

Inference with the Joint

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	<div></div>
		rich	0.0245895	<div></div>
	v1:40.5+	poor	0.0421768	<div></div>
		rich	0.0116293	<div></div>
Male	v0:40.5-	poor	0.331313	<div></div>
		rich	0.0971295	<div></div>
	v1:40.5+	poor	0.134106	<div></div>
		rich	0.105933	<div></div>

$$P(E_1 | E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

- We *could* do prediction using the joint distribution!
- Say we want to **predict if someone is rich vs poor**, and we observe gender and hours worked.
- How do we decide what wealth value to predict?

Why not just model everything with a joint distribution?

Let's say there are 9 features & 1 class to predict. Our joint distribution would have how many rows?

1024

For example, if all possibilities (rows of jpdf) are equally likely, our probability to estimate is $1/1024 \approx 0.001$

We will need many samples to get a good estimate of these probabilities. On average we would need 1000 samples to observe any row one time!

Even worse if some of the rows are less probable

Syntax

- **random variable** = **attribute**, e.g.

Weather is one of $\langle \text{sunny, rainy, cloudy, snow} \rangle$

Windy is one of $\langle \text{windy}, \neg \text{windy} \rangle$

- **Weather** and **Windy** are **discrete** random variables

- Domain values must be

- **exhaustive** and
- **mutually exclusive**

- Elementary propositions:

Weather = **sunny** (or simply **sunny**)

Windy = $\neg \text{windy}$ (or simply $\neg \text{windy}$)

Prior probability and distribution

- **Prior** or **unconditional probability** of a proposition is the **degree of belief** accorded to it in the absence of any other information.

$$P(\text{Weather} = \text{sunny}) = 0.7 \quad (\text{or abbrev. } P(\text{sunny}) = 0.7)$$

- **Probability distribution** gives values for all possible assignments:

$$P(\text{Weather} = \text{sunny}) = 0.7$$

$$P(\text{Weather} = \text{rain}) = 0.2$$

$$P(\text{Weather} = \text{cloudy}) = 0.08$$

$$P(\text{Weather} = \text{snow}) = 0.02$$

Conditional probability

- $P(\text{sunny} \mid \text{windy}) = 0.8$

i.e., probability of sunny given that *windy* is all I know

- **Definition** of conditional probability:

$$P(a \mid b) = P(a \wedge b) / P(b) \quad \text{if } P(b) > 0$$

- Alternative formulation:

$$P(a \wedge b) = P(a \mid b) P(b) = P(b \mid a) P(a)$$

Bayes' Rule

- From $P(a \wedge b) = P(a \mid b) P(b) = P(b \mid a) P(a)$

we get

Bayes' rule:

$$P(a \mid b) = P(b \mid a) P(a) / P(b)$$

- Useful for assessing **class** probability from **evidence** probability as:
 $P(\text{Class} \mid \text{Evidence}) = P(\text{Evidence} \mid \text{Class}) P(\text{Class}) / P(\text{Evidence})$

Onto Naïve Bayes

- We are interested in computing the probability of a class (c) given the features (evidence: e_1, e_2)

$$P(c | e_1, e_2)$$

Bayes' rule -- more vars

$$P(c | e_1, e_2) = \frac{P(c, e_1, e_2)}{P(e_1, e_2)} = \alpha P(c, e_1, e_2)$$

$$\alpha = 1 / P(e_1, e_2)$$

$$= \alpha P(e_1, e_2, c)$$

$$= \alpha P(e_1 | e_2, c) P(e_2, c)$$

$$= \alpha P(e_1 | e_2, c) P(e_2 | c) P(c)$$

$$= \alpha P(e_1 | c) P(e_2 | c) P(c)$$

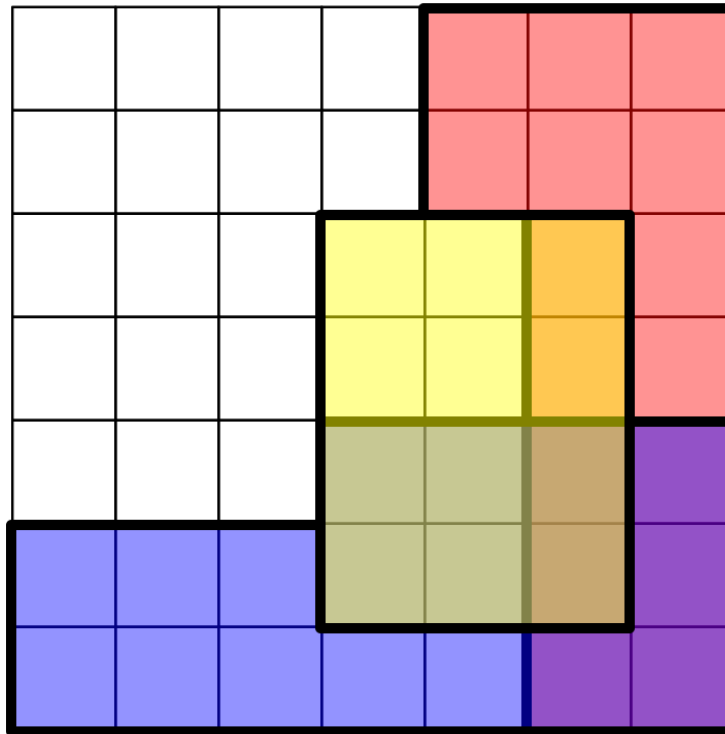
Conditional
Independence

- Although the e_1 might not be independent of e_2 , it might be that *given the class* they are independent.
- E.g.
 - e_1 is 'abilityInReading'
 - e_2 is 'lengthOfArms'
- There is indeed a dependence of **abilityInReading** to **lengthOfArms**. People with longer arms read better than those with short arms....
- However, given a *class variable*, say 'Age', the **abilityInReading** is independent of **lengthOfArms**.

Conditional Independence

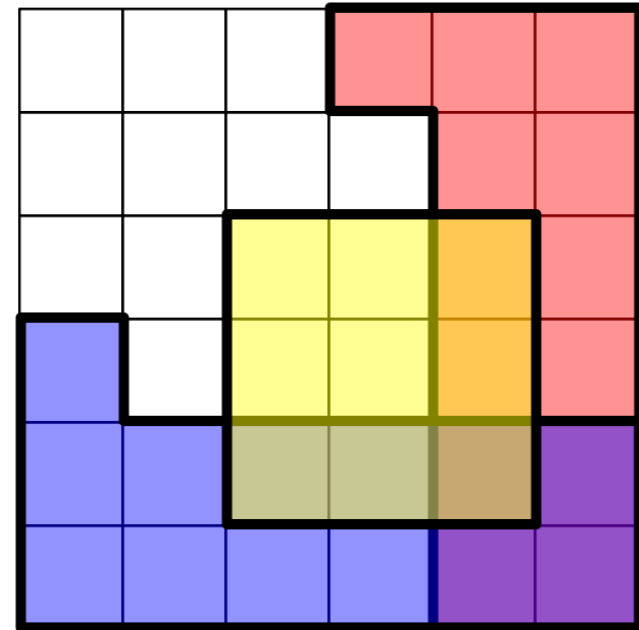
- We say A and B are conditionally independent given C if
 - $P(A \text{ and } B \mid C) = P(A \mid C) * P(B \mid C)$
 - $P(A \mid B \text{ and } C) = P(A \mid C)$

Conditional Independence



$$P(R \text{ and } B | Y) = P(R | Y)P(B | Y)$$

$$2/12 = (4/12)(6/12)$$



$$P(R \text{ and } B | Y) = P(R | Y)P(B | Y)$$

$$1/9 = (3/9)(3/9)$$

Exercise: Are R and B independent given not Y?

$$P(R \text{ and } B | \text{not } Y) = P(R | \text{not } Y)P(B | \text{not } Y) ?$$

Naive Bayes

$$P(c \mid e_1, \dots, e_n) = \alpha P(e_1 \mid c) \dots P(e_n \mid c) P(c)$$

- Assumption:
Attributes are conditionally independent (given the class value)
- Although based on assumption that is almost never correct, this scheme works well in practice!

Weather Data

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

■ A new day:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Naïve Bayes for classification

- Classification learning: what's the probability of the class given an instance?
- **Instance-Tuple (Evidence):** $E_1=e_1, E_2=e_1, \dots, E_n=e_n$
- **Class** $C = \{c, \dots\}$
- Naïve Bayes assumption: evidence can be split into independent parts (i.e. attributes of instance!)

$$\begin{aligned} P(c|E) &= P(c \mid e_1, e_2, \dots, e_n) \\ &= P(e_1|c) P(e_2|c) \dots P(e_n|c) P(c) / P(e_1, e_2, \dots, e_n) \end{aligned}$$

The weather data example

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← *Evidence E*

$$\begin{aligned}
 &P(\text{Play}=\text{yes} \mid E) = \\
 &\quad P(\text{Outlook}=\text{Sunny} \mid \text{play}=\text{yes}) * \\
 &\quad P(\text{Temp}=\text{Cool} \mid \text{play}=\text{yes}) * \\
 &\quad P(\text{Humidity}=\text{High} \mid \text{play}=\text{yes}) * \\
 &\quad P(\text{Windy}=\text{True} \mid \text{play}=\text{yes}) * \\
 &\quad P(\text{play}=\text{yes}) / P(E) = \\
 &= \quad (2/9) * \\
 &\quad (3/9) * \\
 &\quad (3/9) * \\
 &\quad (3/9) * \\
 &\quad (9/14) / P(E) = 0.0053 / P(E)
 \end{aligned}$$

Don't worry for the $1/P(E)$; It's alpha, the normalization constant.

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

The weather data example

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← *Evidence E*

$$P(\text{Play=no} \mid E) =$$

$$P(\text{Outlook=Sunny} \mid \text{play=no}) *$$

$$P(\text{Temp=Cool} \mid \text{play=no}) *$$

$$P(\text{Humidity=High} \mid \text{play=no}) *$$

$$P(\text{Windy=True} \mid \text{play=no}) *$$

$$P(\text{play=no}) / P(E) =$$

$$= (3/5) *$$

$$(1/5) *$$

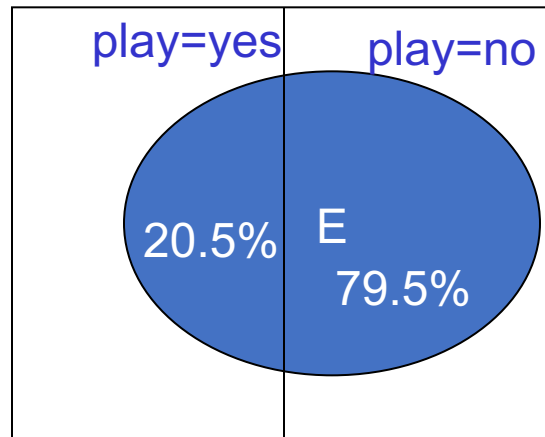
$$(4/5) *$$

$$(3/5) *$$

$$(5/14) / P(E) = 0.0206 / P(E)$$

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Normalization constant



$P(\text{play=yes} \mid E) + P(\text{play=no} \mid E) = 1$ i.e.

$$0.0053 / P(E) + 0.0206 / P(E) = 1 \quad \text{i.e.}$$

$$P(E) = 0.0053 + 0.0206$$

So,

$$P(\text{play=yes} \mid E) = 0.0053 / (0.0053 + 0.0206) = \mathbf{20.5\%}$$

$$P(\text{play=no} \mid E) = 0.0206 / (0.0053 + 0.0206) = \mathbf{79.5\%}$$

The “zero-frequency problem”

- What if an attribute value doesn't occur with every class value (e.g. “Humidity = High” for class “Play=Yes”)?
 - Probability $P(\text{Humidity=High} | \text{play=yes})$ will be zero!
- $P(\text{Play=“Yes”} | E)$ will also be zero!
 - No matter how likely the other values are!

- Remedy:
 - Add **1** to the count for every attribute value-class combination (Laplace estimator);
 - Add **k** (# of possible attribute values) to the denominator. (see example on the right).

$$P(\text{play=yes} | E) =$$

$$P(\text{Outlook=Sunny} | \text{play=yes}) *$$

$$P(\text{Temp=Cool} | \text{play=yes}) *$$

$$P(\text{Humidity=High} | \text{play=yes}) *$$

$$P(\text{Windy=True} | \text{play=yes}) *$$

$$P(\text{play=yes}) / P(E) =$$

$$= (2/9) * (3/9) * (3/9) * (3/9) * (9/14) / P(E) = 0.0053 / P(E)$$

It will be instead:

Number of possible values for 'Outlook'

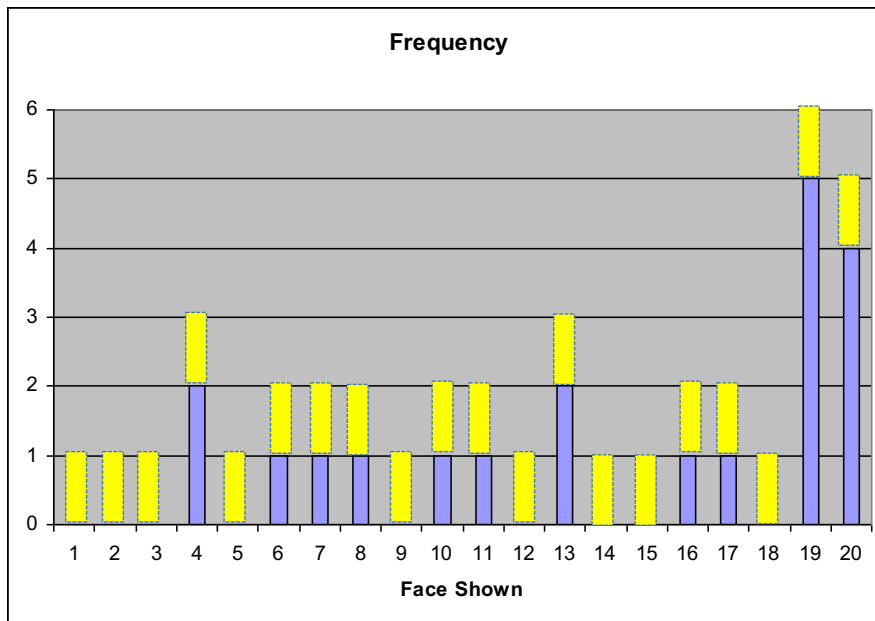
$$= ((2+1)/(9+3)) * ((3+1)/(9+3)) * ((3+1)/(9+2)) * ((3+1)/(9+2)) * ((9+1)/(14+2)) / P(E) = 0.0069 / P(E)$$

Number of possible values for 'Windy'

Number of possible values for 'play'

The “zero-frequency problem”

- What we just did is called smoothing
- “Hallucinating” training data, like we saw in MAP estimate



Dealing with continuous attributes

- Usual assumption: attributes have a normal or Gaussian probability distribution (given the class).
- Probability density function for the normal distribution is:

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

- We approximate μ by the sample mean:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- We approximate σ^2 by the sample variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Weather Data with Numeric Attrib.

outlook	temperature	humidity	windy	play
sunny	85	85	FALSE	no
sunny	80	90	TRUE	no
overcast	83	86	FALSE	yes
rainy	70	96	FALSE	yes
rainy	68	80	FALSE	yes
rainy	65	70	TRUE	no
overcast	64	65	TRUE	yes
sunny	72	95	FALSE	no
sunny	69	70	FALSE	yes
rainy	75	80	FALSE	yes
sunny	75	70	TRUE	yes
overcast	72	90	TRUE	yes
overcast	81	75	FALSE	yes
rainy	71	91	TRUE	no

We compute similarly:
 $f(\text{Temperature}=66 \mid \text{no})$

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

$$f(\text{Temperature}=66 \mid \text{yes}) \\ = e(-((66-m)^2 / 2*\text{var})) / \\ \text{sqrt}(2*3.14*\text{var})$$

$$m = \\ (83+70+68+64+69+75+75+72 \\ +81) / 9 = 73$$

$$\text{var} = ((83-73)^2 + (70-73)^2 + \\ (68-73)^2 + (64-73)^2 + (69- \\ 73)^2 + (75-73)^2 + (75-73)^2 \\ + (72-73)^2 + (81-73)^2) / (9-1) \\ = 38$$

$$f(\text{Temperature}=66 \mid \text{yes}) \\ = e(-((66-73)^2 / (2*38))) / \\ \text{sqrt}(2*3.14*38) = .034$$

Weather Data

outlook	temperature	humidity	windy	play
sunny	85	85	FALSE	no
sunny	80	90	TRUE	no
overcast	83	86	FALSE	yes
rainy	70	96	FALSE	yes
rainy	68	80	FALSE	yes
rainy	65	70	TRUE	no
overcast	64	65	TRUE	yes
sunny	72	95	FALSE	no
sunny	69	70	FALSE	yes
rainy	75	80	FALSE	yes
sunny	75	70	TRUE	yes
overcast	72	90	TRUE	yes
overcast	81	75	FALSE	yes
rainy	71	91	TRUE	no

We compute similarly:

$f(\text{Humidity}=90 \mid \text{no})$

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

$f(\text{Humidity}=90 \mid \text{yes})$

$$= e^{-((90-m)^2 / 2 \cdot \text{var})} / \sqrt{2 \cdot 3.14 \cdot \text{var}}$$

$m =$

$$(86+96+80+65+70+80+70+90+75) / 9 = 79$$

$$\begin{aligned} \text{var} = & ((86-79)^2 + (96-79)^2 + (80-79)^2 + (65-79)^2 + (70-79)^2 \\ & + (80-79)^2 + (70-79)^2 + (90-79)^2 + (75-79)^2) / (9-1) \\ = & 104 \end{aligned}$$

$f(\text{Humidity}=90 \mid \text{yes})$

$$= e^{-((90-79)^2 / (2 \cdot 104))} / \sqrt{2 \cdot 3.14 \cdot 104} = .022$$

Classifying a new day

- A new day E:

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

$$P(\text{play}=\text{yes} \mid E) =$$

$$P(\text{Outlook}=\text{sunny} \mid \text{play}=\text{yes}) *$$

$$P(\text{Temp}=66 \mid \text{play}=\text{yes}) *$$

$$P(\text{Humidity}=90 \mid \text{play}=\text{yes}) *$$

$$P(\text{Windy}=\text{true} \mid \text{play}=\text{yes}) *$$

$$P(\text{play}=\text{yes}) / P(E) =$$

$$= (2/9) * (0.034) * (0.022) * (3/9) \\ * (9/14) / P(E) = 0.000036 / P(E)$$

$$P(\text{play}=\text{no} \mid E) =$$

$$P(\text{Outlook}=\text{sunny} \mid \text{play}=\text{no}) *$$

$$P(\text{Temp}=66 \mid \text{play}=\text{no}) *$$

$$P(\text{Humidity}=90 \mid \text{play}=\text{no}) *$$

$$P(\text{Windy}=\text{true} \mid \text{play}=\text{no}) *$$

$$P(\text{play}=\text{no}) / P(E) =$$

$$= (3/5) * (0.0291) * (0.038) * (3/5) \\ * (5/14) / P(E) = 0.000136 / P(E)$$

After normalization: $P(\text{play}=\text{yes} \mid E) = 20.9\%$, $P(\text{play}=\text{no} \mid E) = 79.1\%$

Probability densities

- Relationship between probability and density:

$$\Pr[c - \frac{\varepsilon}{2} < x < c + \frac{\varepsilon}{2}] \approx \varepsilon * f(c)$$

- But: this doesn't change calculation of a posteriori probabilities because ε cancels out after normalization

Discussion of Naïve Bayes

- Naïve Bayes works surprisingly well (even if independence assumption is clearly violated)
- Because classification doesn't require very accurate probability estimates so long as *maximum* probability is assigned to correct class

Tax Data – Naive Bayes

Classify: (_, No, Married, 95K, ?)

(Using Laplace normalization)

<i>Tid</i>	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Tax Data – Naive Bayes

Classify: (, No, Married, 95K, ?)

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$$P(\text{Yes}) = 3/10 = 0.3$$

$$P(\text{Refund}=\text{No} | \text{Yes}) = (3+1)/(3+2) = 0.8$$

$$P(\text{Status}=\text{Married} | \text{Yes}) = (0+1)/(3+3) = 0.17$$

$$f(\text{income} | \text{Yes}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Approximate μ with: $(95+85+90)/3 = 90$

Approximate σ^2 with:

$$\frac{((95-90)^2 + (85-90)^2 + (90-90)^2)}{(3-1)} = 25$$

$$f(\text{income}=95 | \text{Yes}) =$$

$$e^{-((95-90)^2 / (2*25))} / \sqrt{2*3.14*25} = .048$$

$$P(\text{Yes} | E) = \alpha * .8 * .17 * .048 * .3 = \alpha * 0.0019584$$

Tax Data

<i>Tid</i>	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Classify: (, No, Married, 95K, ?)

$$P(\text{No}) = 7/10 = .7$$

$$P(\text{Refund}=\text{No}|\text{No}) = (4+1)/(7+2) = .556$$

$$P(\text{Status}=\text{Married}|\text{No}) = (4+1)/(7+3) = .5$$

$$f(\text{income} | \text{No}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Approximate μ with:

$$(125+100+70+120+60+220+75)/7 = 110$$

Approximate σ^2 with:

$$((125-110)^2 + (100-110)^2 + (70-110)^2 + (120-110)^2 + (60-110)^2 + (220-110)^2 + (75-110)^2) / (7-1) = 2975$$

$$f(\text{income}=95|\text{No}) =$$

$$e(-((95-110)^2 / (2*2975)))$$

$$/\text{sqrt}(2*3.14* 2975) = .00704$$

$$P(\text{No} | \text{E}) = \alpha * .556 * .5 * .00704 * 0.7 =$$

$$\alpha * .00137$$

Tax Data

Classify: (, No, Married, 95K, ?)

$$P(\text{Yes} | E) = \alpha * .0019584$$

$$P(\text{No} | E) = \alpha * .00137$$

$$\alpha = 1/ (.0019584 + .00137) = 300.44$$

$$P(\text{Yes}|E) = 300.44 * .0019584 = 0.59$$

$$P(\text{No}|E) = 300.44 * .00137 = 0.41$$

We predict “**Yes.**”

<i>Tid</i>	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Benefits

- Why use Naïve Bayes
 - works well
 - fewer parameters
- How many params for binary classification with N binary features?
 - $P(e_1 | C) \dots P(e_N | C) * P(C)$
- For the full joint distribution?
 - $P(e_1 \dots e_N | C) * P(C)$
- In practice, take the log before multiplying

Text Classification

- Assign a document to a category (e.g. spam/not spam)
- Naïve Bayes models are often used for this task.
- Evidence variables are the presence or absence of each word in the language.
 - Bag of words

Bernoulli Naïve Bayes

- Training: Given **training data** a set of documents that **have been** assigned to categories, training is just counting

of course $P(\neg W | C) = 1 - P(W | C)$

- Prior probability $P(\text{Category})$
 - Fraction of all the training documents that are of that category
- Conditional probabilities $P(\text{Word}_i | \text{Category})$
 - Fraction of docs of category c that contain word Word_i .
- Conditional probabilities $P(\neg \text{Word}_i | \text{Category})$
 - Fraction of docs of category c that **don't** contain word Word_i .
- Also, $P(\text{Word}_i | \text{Category} = \neg c)$
 - Fraction of docs **not** of category c that contain word Word_i .
- $P(\neg \text{Word}_i | \text{Category} = \neg c)$
 - Fraction of docs **not** of category c that **don't** contain word Word_i .

Bernoulli Naïve Bayes

- Now we can use Naïve Bayes for classifying a new document:

$$P(\text{Category} = c \mid \text{Word}_1 = \text{true}, \dots, \text{Word}_n = \text{false}) = \alpha * P(\text{Category} = c) \prod_{i=1}^n P(\text{Word}_i ? \mid \text{Category} = c)$$

$$P(\text{Category} = \neg c \mid \text{Word}_1 = \text{true}, \dots, \text{Word}_n = \text{false}) = \alpha * P(\text{Category} = \neg c) \prod_{i=1}^n P(\text{Word}_i ? \mid \text{Category} = \neg c)$$

- $\text{Word}_1, \dots, \text{Word}_n$ are all of the words in **any** document
 - (i.e. all the words we know about)
- $P(\text{Word}_i ? \mid \text{Category} = c)$ is
 - $P(\text{Word}_i \mid \text{Category} = c)$ if word i does appear in test doc
 - $P(\neg \text{Word}_i \mid \text{Category} = c)$ if word i does not appear in test doc
- α is the normalization constant.

Vectorizing Bernoulli NB at classification time

$$\hat{y} = \operatorname{argmax}_y \left(P(y) \prod_{i=1}^p P(x_i|y) \right)$$

- For each y_i

$$P(y_i) \prod_{i=1}^p P(x_i|y_i)$$

$$\log(P(y_i)) + \sum_{i=1}^p \log(P(x_i|y_i))$$

If word x_i is in the document, use $p(x_i|y_i)$, else use $1 - p(x_i|y_i)$

If you have a vector $v = [p(x_1|y_1), \dots, p(x_p|y_p)]$ and a binary vector w where each element w_i is 1 if word x_i is in the document, how could you simplify the above sum?

Comments on the assignment

- Avoid for loops and use vectorized expressions
- Use numpy functions as much as possible (sum, log, mean, etc...)
 - numpy is built on top of high performance linear algebra libraries. When you use those fxns, you are calling efficient C/Fortran code behind the scenes!
- In particular for the assignment you don't need to train a new model for each smoothing value, because the data behind the counts doesn't change!

Multinomial Naïve Bayes

- Don't use probability for words that don't appear in a document

$$P(\text{Category} = c \mid \text{Word}_1 = \text{true}, \dots, \text{Word}_m = \text{true}) =$$

$$\alpha * P(\text{Category} = c) \prod_{i=1}^m P(\text{Word}_i \mid \text{Category} = c)$$

- Now the product is over **only the m words that do appear** in the document we are testing on
 - (all words will be equal to true)
- If the same word appears more than one time (say t times), we count its probability more than one time (t times).

Multinomial Naïve Bayes

- We also redefine the $P(w_i | y)$
- N_{yi} is the total # of times feature i appeared for any instance with label y

$$P(w_i | y) = \frac{N_{yi} + \alpha_s}{N_y + p * \alpha_s}$$

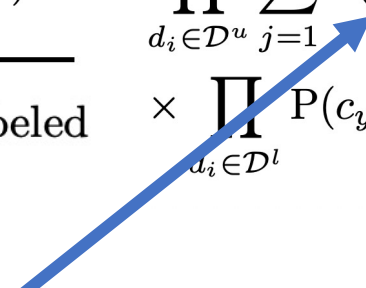
- $N_y = \sum_{i=1}^p N_{yi}$ total # of times any feature appeared in an instance labeled y
- p is the total number of features,
- and α_s is the additive smoothing parameter.
 - Different from normalization constant

Naïve Bayes and unlabeled data

- Labeling data is (often) costly
- In some scenarios, there is far more unlabeled data vs labeled data
- Naïve Bayes gives a nice framework within which to reason about unlabeled data

$$P(\mathcal{D}|\theta) = \prod_{d_i \in \mathcal{D}^u} \sum_{j=1}^{|\mathcal{C}|} P(c_j|\theta)P(d_i|c_j; \theta) \\ \times \prod_{d_i \in \mathcal{D}^l} P(c_{y_i}|\theta)P(d_i|c_{y_i}; \theta).$$

Naïve Bayes and unlabeled data

$$P(\mathcal{D}|\theta) = \prod_{d_i \in \mathcal{D}^u} \sum_{j=1}^{|\mathcal{C}|} P(c_j|\theta)P(d_i|c_j; \theta) \times \prod_{d_i \in \mathcal{D}^l} P(c_{y_i}|\theta)P(d_i|c_{y_i}; \theta).$$


-
- Build an initial classifier by calculating $\hat{\theta}$ from the labeled documents only (Equations 6 and 7).
 - Loop while classifier parameters change:
 - Use the current classifier to calculate probabilistically-weighted labels for the unlabeled documents (Equation 8).
 - Recalculate the classifier parameters $\hat{\theta}$ given the probabilistically assigned labels (Equations 6 and 7).
-

Table 1: The Algorithm.

- This is an example of the EM algorithm, which we will discuss later this semester

Questions?