A [somewhat] Quick Overview of Probability

Shannon Quinn

CSCI 6900

[Some material pilfered from http://www.cs.cmu.edu/~awm/tutorials]

Probabilistic and Bayesian Analytics

Note to other teachers and users of these slides. Andrew would be delighted if you found this source material useful in giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. PowerPoint originals are available. If you make use of a significant portion of these slides in your own lecture, please include this message, or the following link to the source repository of Andrew's tutorials: http://www.cs.cmu.edu/~awm/tutorials . Comments and corrections gratefully received.

Andrew W. Moore
School of Computer Science
Carnegie Mellon University
www.cs.cmu.edu/~awm
awm@cs.cmu.edu
412-268-7599

Probability - what you need to really, really know

- Probabilities are cool

Probability - what you need to really, really know

- Probabilities are cool
- Random variables and events

Discrete Random Variables

- A is a Boolean-valued random variable if
  - A denotes an event,
  - there is uncertainty as to whether A occurs.
- Examples
  - A = The US president in 2023 will be male
  - A = You wake up tomorrow with a headache
  - A = You have Ebola
  - A = the $1{,}000{,}000{,}000{,}000^{th}$ digit of π is 7
- Define P(A) as "the fraction of possible worlds in which A is true"
  - We're assuming all possible worlds are equally probable

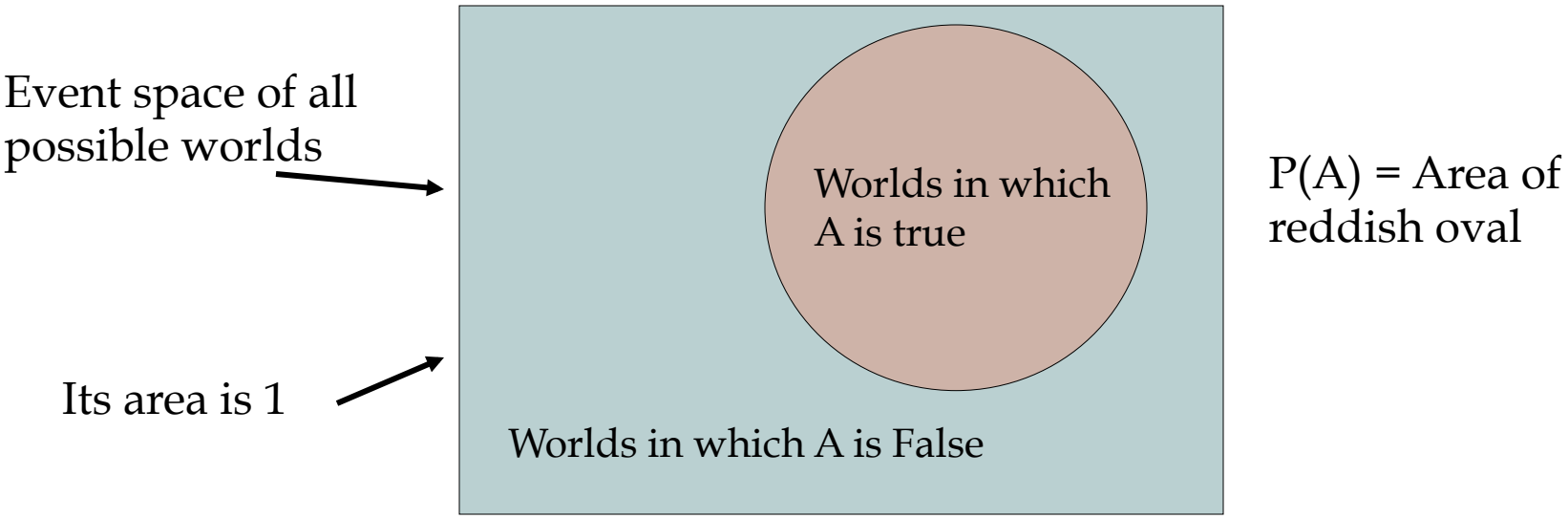
Discrete Random Variables

- A is a Boolean-valued random variable if
  - A denotes an event, a possible outcome of an "experiment"
  - there is uncertainty as to whether A occurs.

    the experiment is not deterministic

- Define P(A) as "the fraction of experiments in which A is true"
  - We're assuming all possible outcomes are equiprobable
- Examples
  - You roll two 6-sided die (the experiment) and get doubles (A=doubles, the outcome)
  - I pick two students in the class (the experiment) and they have the same birthday (A=same birthday, the outcome)

Visualizing A



Probability - what you need to really, really know

- Probabilities are cool
- Random variables and events
- There is One True Way to talk about uncertainty: the Axioms of Probability

The Axioms of Probability

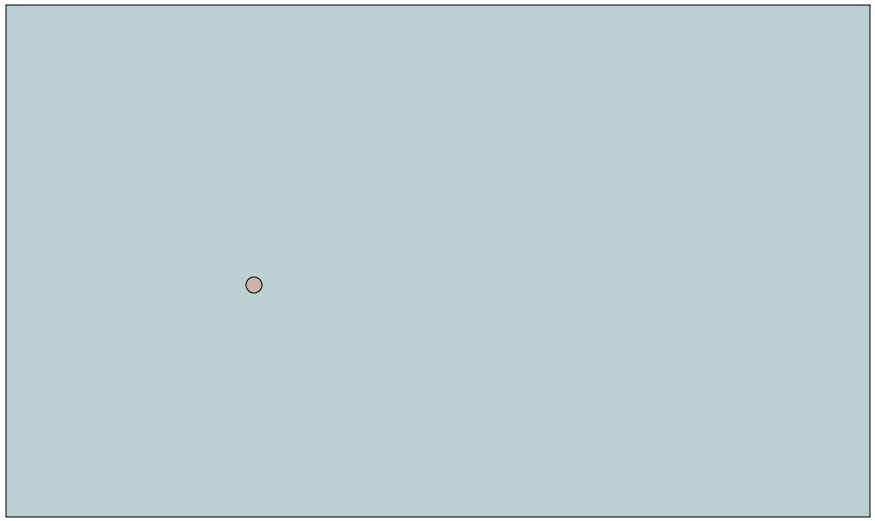
- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)

Events, random variables, …., probabilities

(This is Andrew's joke)

Interpreting the axioms

- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)

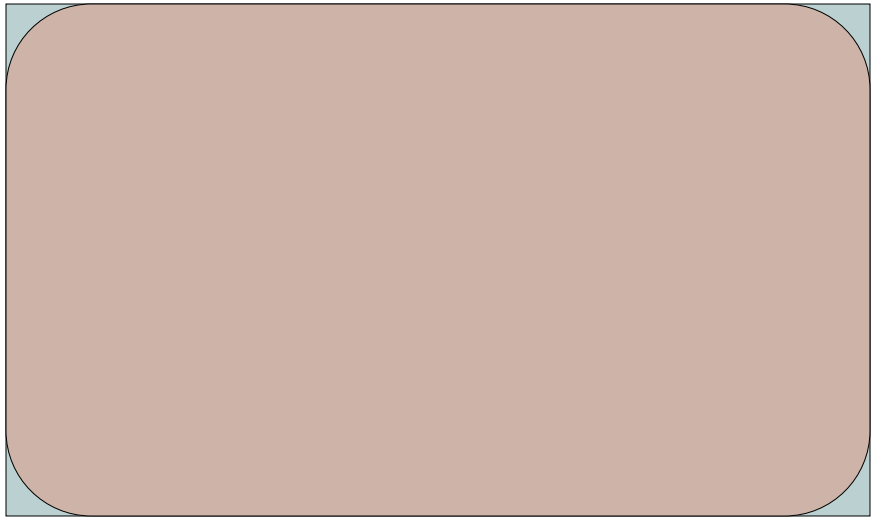


The area of A can't get any smaller than 0

And a zero area would mean no world could ever have A true

Interpreting the axioms

- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)

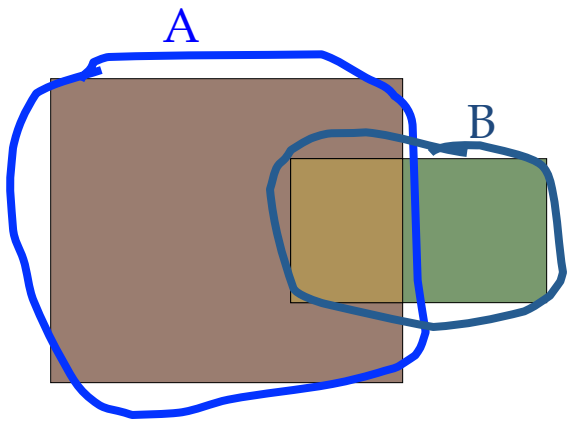


The area of A can't get any bigger than 1

And an area of 1 would mean all worlds will have A true

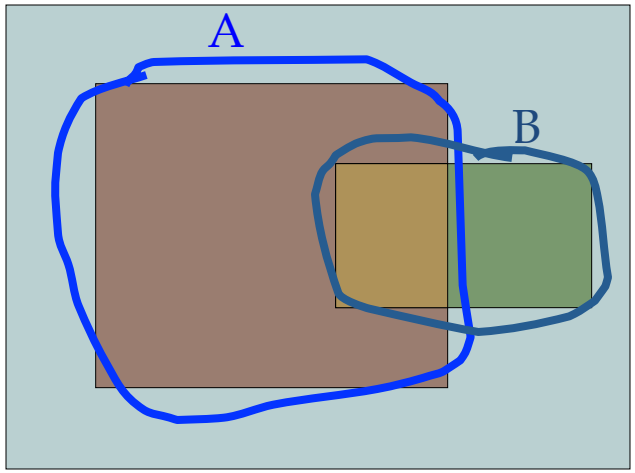
Interpreting the axioms

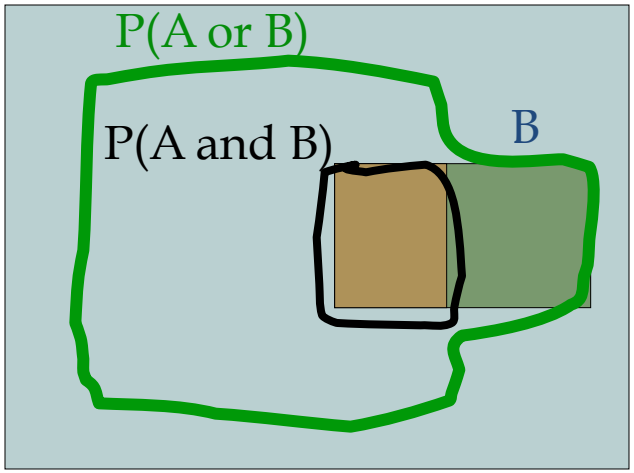
- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)



Interpreting the axioms

- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)





Simple addition and subtraction

Theorems from the Axioms

- $0 <= P(A) <= 1$, $P(True) = 1$, $P(False) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

➔ $P(\text{not } A) = P(\sim A) = 1 - P(A)$

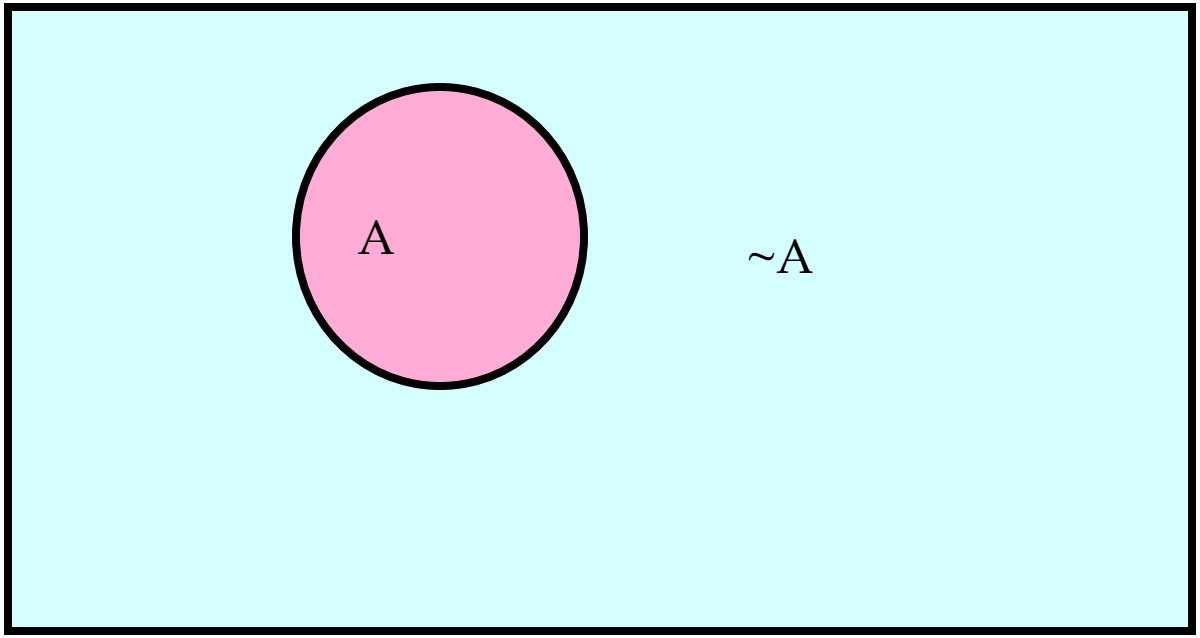
$P(A \text{ or } \sim A) = 1$ $\quad\quad$ $P(A \text{ and } \sim A) = 0$

$P(A \text{ or } \sim A) = P(A) + P(\sim A) - P(A \text{ and } \sim A)$

$1 = P(A) + P(\sim A) - 0$

Elementary Probability in Pictures

- P(~A) + P(A) = 1



Another important theorem

- 0 <= P(A) <= 1, P(True) = 1, P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)

➔ $P(A) = P(A \wedge B) + P(A \wedge \sim B)$

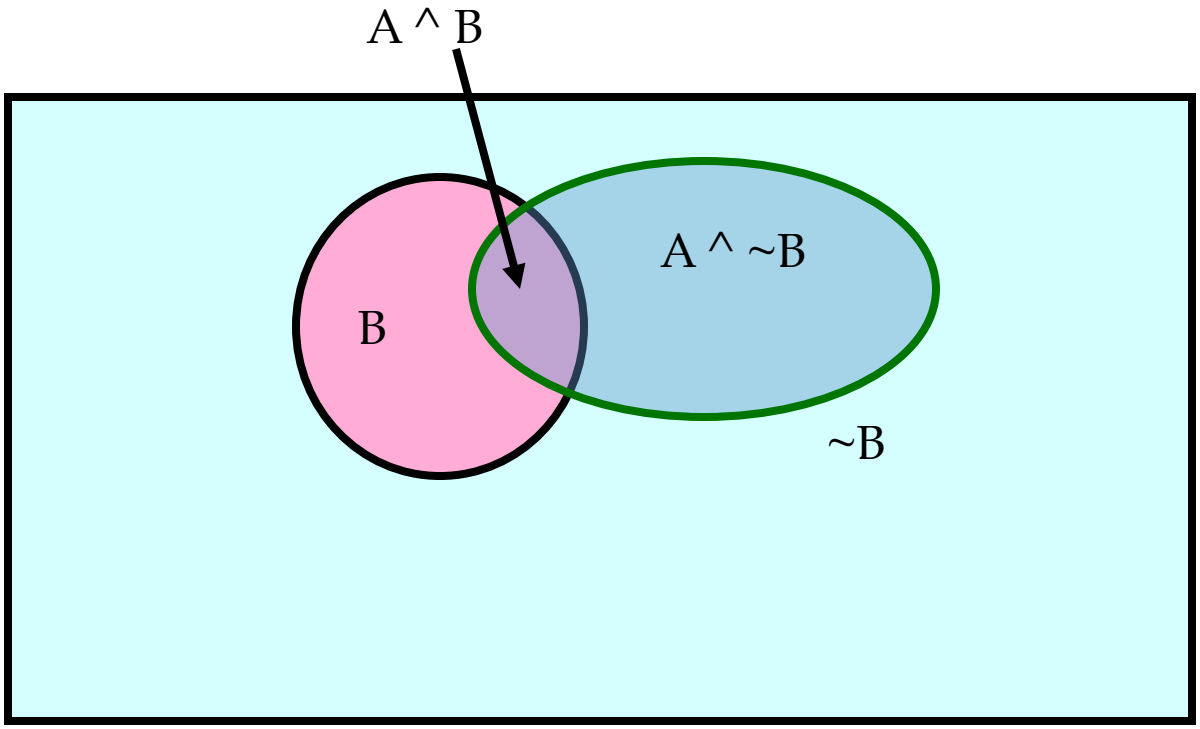
A = A and (B or ~B) = (A and B) or (A and ~B)

P(A) = P(A and B) + P(A and ~B) – P((A and B) and (A and ~B))

P(A) = P(A and B) + P(A and ~B) – P(A and A and B and ~B)

Elementary Probability in Pictures

- P(A) = P(A ^ B) + P(A ^ ~B)



Probability - what you need to really, really know

- Probabilities are cool
- Random variables and events
- The Axioms of Probability
- Independence

Independent Events

- Definition: two events A and B are *independent* if Pr(A and B)=Pr(A)*Pr(B).
- Intuition: outcome of A has no effect on the outcome of B (and vice versa).
  - We need to assume the different rolls are *independent* to solve the problem.
  - You frequently need to assume the independence of *something* to solve any learning problem.

Some practical problems

Pearl Purple Cheaters Dice **$6.50**

This is a set of polyhedral dice that will roll high/low numbers. The set consists of 3 six-sided dice (1 high, 1 low, 1 standard), 2 ten-sided dice (1 high, 1 standard), and 2 twenty-sided dice (1 high, 1 standard).

on order

- You're the DM in a D&D game.
- Joe brings his own d20 and throws 4 critical hits in a row to start off
  - DM=dungeon master
  - D20 = 20-sided die
  - "Critical hit" = 19 or 20
- What are the odds of that happening with a fair die?
- Ci=critical hit on trial i, i=1,2,3,4
- P(C1 and C2 … and C4) = P(C1)*…*P(C4) = (1/10)^4

Multivalued Discrete Random Variables

- Suppose A can take on more than 2 values
- A is a random variable with arity k if it can take on exactly one value out of $\{v_1, v_2, .. v_k\}$
  - *Example: V={aaliyah, aardvark, …., zymurge, zynga}*
  - *Example: V={aaliyah_aardvark, …, zynga_zymgurgy}*

- Thus…

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$

$$P(A = v_1 \vee A = v_2 \vee A = v_k) = 1$$

Terms: Binomials and Multinomials

- Suppose A can take on more than 2 values
- A is a random variable with arity k if it can take on exactly one value out of $\{v_1, v_2, .. v_k\}$
  - *Example: V={aaliyah, aardvark, …., zymurge, zynga}*
  - *Example: V={aaliyah_aardvark, …, zynga_zymgurgy}*

- The distribution Pr(A) is a *multinomial*
- For *k*=2 the distribution is a *binomial*

More about Multivalued Random Variables

- Using the axioms of probability and assuming that A obeys. $P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$

$$P(A = v_1 \vee A = v_2 \vee A = v_k) = 1$$

- It's easy to prove that

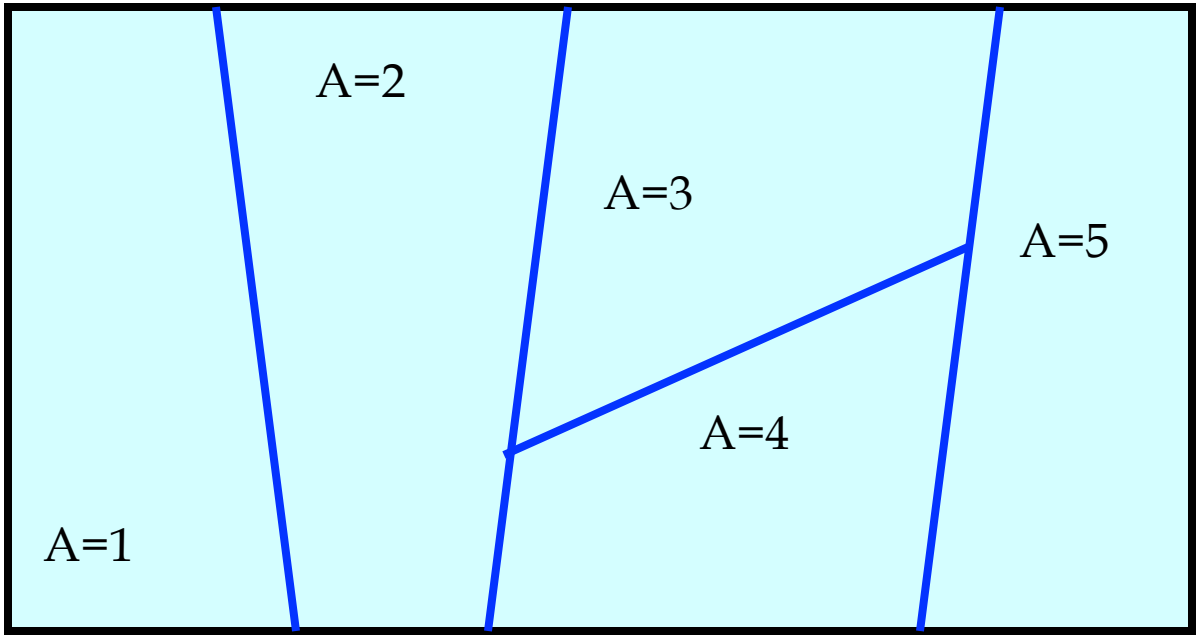
$$P(A = v_1 \vee A = v_2 \vee A = v_i) = \sum_{j=1}^{i} P(A = v_j)$$

- And thus we can prove

$$\sum_{j=1}^{k} P(A = v_j) = 1$$

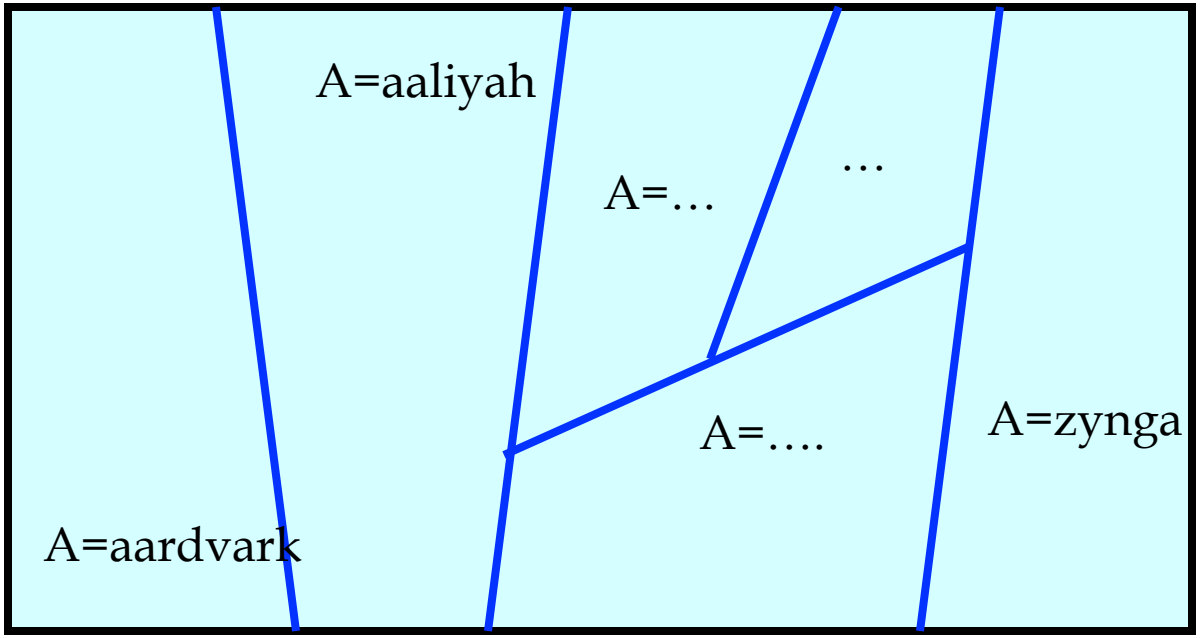
Elementary Probability in Pictures

$$\sum_{j=1}^{k} P(A = v_j) = 1$$



Elementary Probability in Pictures

$$\sum_{j=1}^{k} P(A = v_j) = 1$$



Probability - what you need to really, really know

- Probabilities are cool
- Random variables and events
- The Axioms of Probability
- Independence, binomials, multinomials
- Conditional probabilities

A practical problem

This is a set of polyhedral dice that will roll high/low numbers. The set consists of 3 six-sided dice (1 high, 1 low, 1 standard), 2 ten-sided dice (1 high, 1 standard), and 2 twenty-sided dice (1 high, 1 standard).

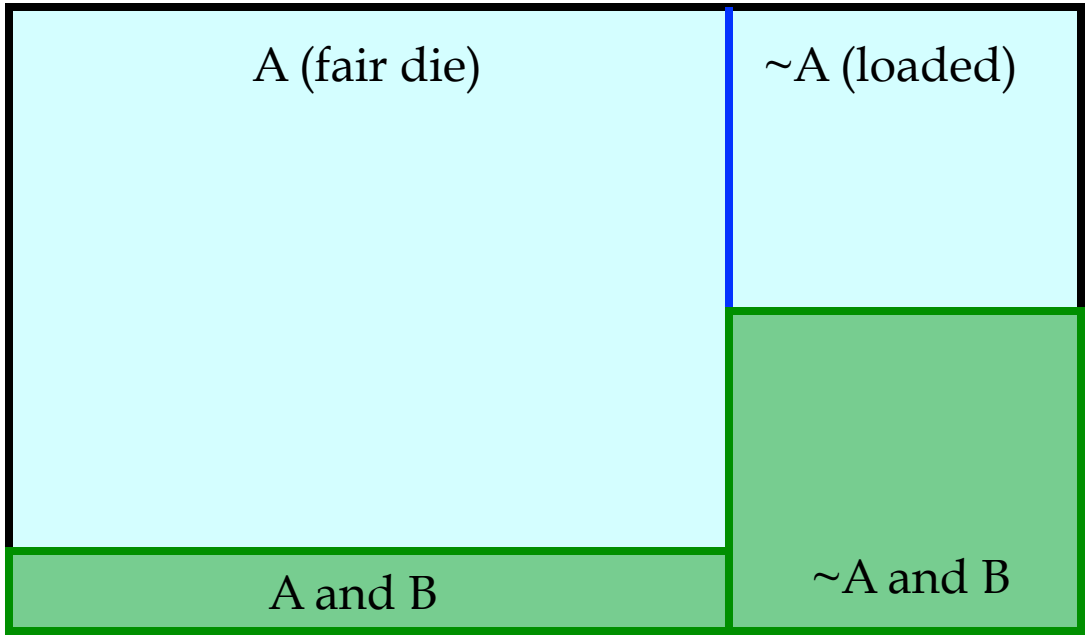
- I have lots of standard d20 die, lots of loaded die, all identical.
- Loaded die will give a 19/20 ("critical hit") half the time.
- In the game, someone hands me a random die, which is fair (A) or loaded (~A), with P(A) depending on how I mix the die. Then I roll, and either get a critical hit (B) or not (~B)
- . Can I mix the dice together so that P(B) is anything I want - say, p(B)= 0.137 ?

$P(B) = P(B \text{ and } A) + P(B \text{ and } \sim A) = 0.1*\lambda + 0.5*(1-\lambda) = 0.137$

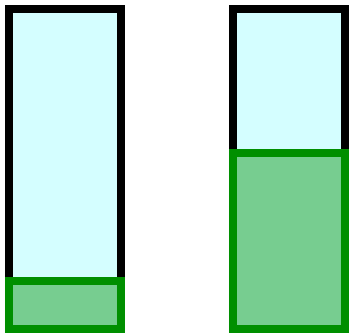
"mixture model" $\lambda = (0.5 - 0.137)/0.4 = 0.9075$

Another picture for this problem

It's more convenient to say
- "if you've picked a fair die then …" i.e. Pr(critical hit | fair die)=0.1
- "if you've picked the loaded die then…." Pr(critical hit | loaded die)=0.5



Conditional probability:
Pr(B | A) = P(B^A)/P(A)



P(B | A) P(B | ~A)

Definition of Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

Corollary: The Chain Rule

$$P(A \wedge B) = P(A \mid B)\, P(B)$$

Some practical problems

"marginalizing out" A

This is a set of polyhedral dice that will roll high/low numbers. The set consists of 3 six-sided dice (1 high, 1 low, 1 standard), 2 ten-sided dice (1 high, 1 standard), and 2 twenty-sided dice (1 high, 1 standard).

- I have 3 standard d20 dice, 1 loaded die.
- Experiment: (1) pick a d20 uniformly at random then (2) roll it. Let A=d20 picked is fair and B=roll 19 or 20 with that die. What is P(B)?

$P(B) = P(B \mid A)\, P(A) + P(B \mid \sim A)\, P(\sim A)$ $= 0.1 * 0.75 + 0.5 * 0.25 = 0.2$

posterior

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

prior

Bayes' rule

$$P(B \mid A) = \frac{P(A \mid B) * P(B)}{P(A)}$$

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

…by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter…. necessary to be considered by any that would give a clear account of the strength of *analogical* or *inductive reasoning*…

Probability - what you need to really, really know

- Probabilities are cool
- Random variables and events
- The Axioms of Probability
- Independence, binomials, multinomials
- Conditional probabilities
- Bayes Rule

Some practical problems

- Joe throws 4 critical hits in a row, is Joe cheating?
- A = Joe using cheater's die
- C = roll 19 or 20; P(C | A)=0.5, P(C | ~A)=0.1
- B = C1 and C2 and C3 and C4
- Pr(B | A) = 0.0625 P(B | ~A)=0.0001

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\sim A)P(\sim A)}$$

$$P(A|B) = \frac{0.0625 * P(A)}{0.0625 * P(A) + 0.0001 * (1 - P(A))}$$

What's the experiment and outcome here?

- Outcome A: Joe is cheating
- Experiment:
  - Joe picked a die uniformly at random from a bag containing 10,000 fair die and one bad one.
  - Joe is a D&D player picked uniformly at random from set of 1,000,000 people and n of them cheat with probability $p>0$.
  - I have no idea, but I don't like his looks. Call it P(A)=0.1

Some practical problems

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$\frac{P(A|B)}{P(\neg A|B)} = \frac{P(B|A)P(A)/P(B)}{P(B|\neg A)P(\neg A)/P(B)} = \frac{P(B|A)}{P(B|\neg A)} \times \frac{P(A)}{P(\neg A)}$$

$$= \frac{0.0625}{0.0001} \times \frac{P(A)}{P(\neg A)} = 6{,}250 \times \frac{P(A)}{P(\neg A)}$$

- Joe throws 4 critical hits in a row, is Joe cheating?
- A = Joe using cheater's die
- C = roll 19 or 20; P(C | A)=0.5, P(C | ~A)=0.1
- B = C1 and C2 and C3 and C4
- Pr(B | A) = 0.0625 P(B | ~A)=0.0001

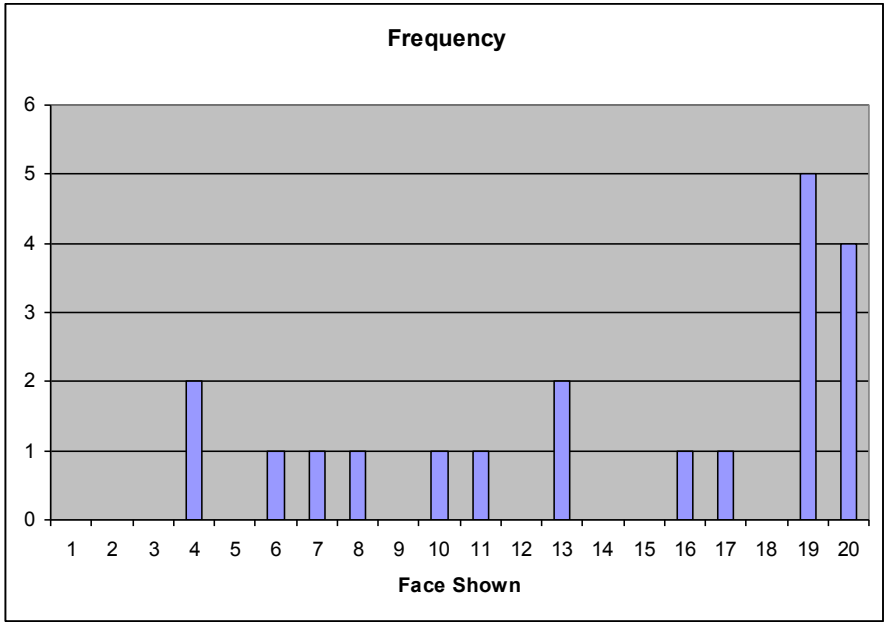
Moral: with enough evidence the prior P(A) doesn't really matter.

Probability - what you need to really, really know

- Probabilities are cool
- Random variables and events
- The Axioms of Probability
- Independence, binomials, multinomials
- Conditional probabilities
- Bayes Rule
- MLE's, smoothing, and MAPs

Some practical problems

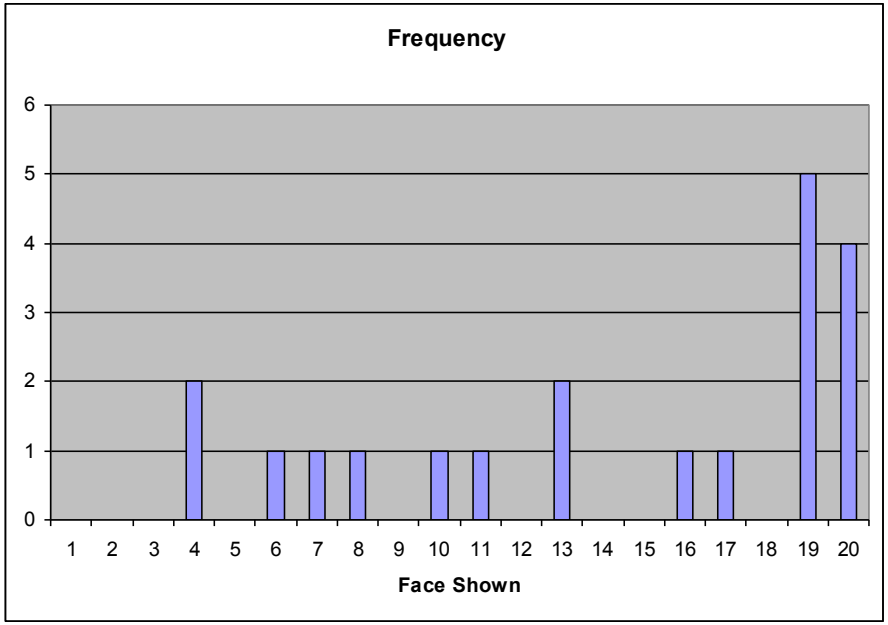
I bought a loaded d20 on EBay…but it didn't come with any specs. How can I find out how it behaves?



1. Collect some data (20 rolls)
2. Estimate Pr(i)=C(rolls of i)/C(any roll)

One solution

I bought a loaded d20 on EBay…but it didn't come with any specs. How can I find out how it behaves?



P(1)=0

P(2)=0

P(3)=0

P(4)=0.1

…

P(19)=0.25

P(20)=0.2

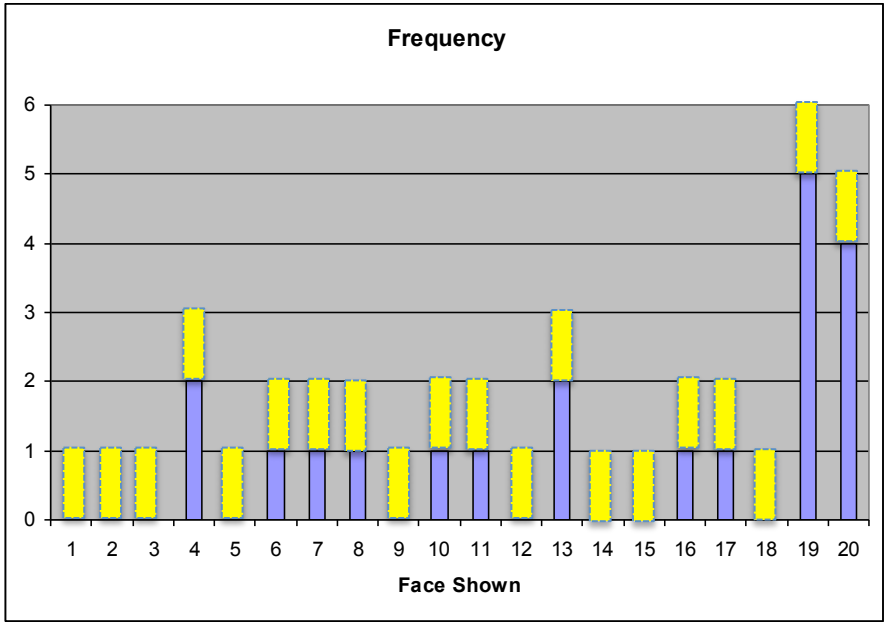
MLE = <u>maximum likelihood estimate</u>

But: Do I really think it's *impossible* to roll a 1,2 or 3?

Would you bet your house on it?

A better solution

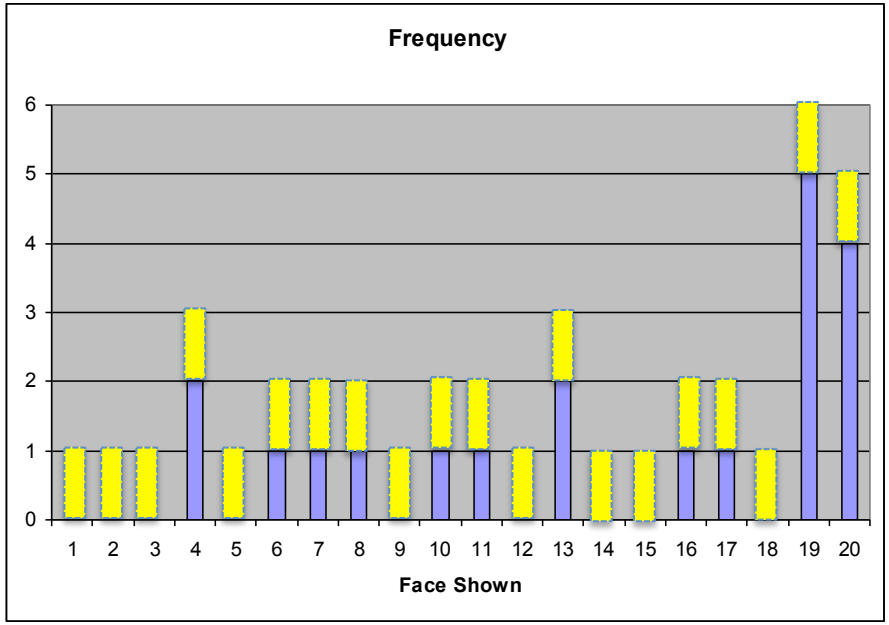
I bought a loaded d20 on EBay…but it didn't come with any specs. How can I find out how it behaves?



0. *Imagine* some data (20 rolls, each i shows up 1x)
1. Collect some data (20 rolls)
2. Estimate Pr(i)=C(rolls of i)/C(any roll)

A better solution

I bought a loaded d20 on EBay…but it didn't come with any specs. How can I find out how it behaves?



P(1)=1/40

P(2)=1/40

P(3)=1/40

P(4)=(2+1)/40

…

P(19)=(5+1)/40

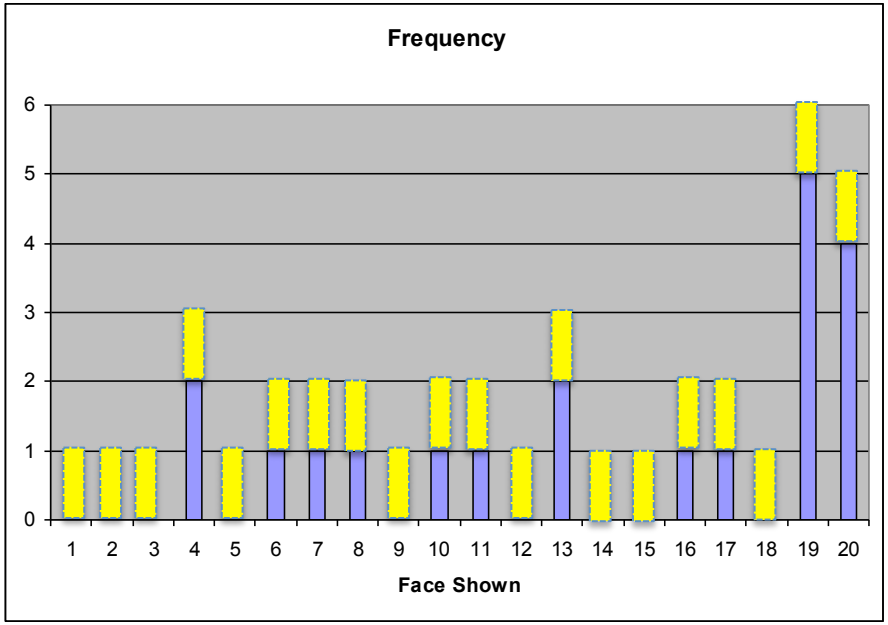
P(20)=(4+1)/40=1/8

$$\hat{\Pr}(i) = \frac{C(i)+1}{C(ANY)+C(IMAGINED)}$$

0.25 *vs.* 0.125 – really different! Maybe I should "imagine" less data?

A better solution?

Pearl Purple Cheaters Dice



P(1)=1/40

P(2)=1/40

P(3)=1/40

P(4)=(2+1)/40

…

P(19)=(5+1)/40

P(20)=(4+1)/40=1/8

$$\hat{\Pr}(i) = \frac{C(i)+1}{C(ANY)+C(IMAGINED)}$$

0.25 *vs.* 0.125 – really different! Maybe I should "imagine" less data?

A better solution?

Q: What if I used m rolls with a probability of $q=1/20$ of rolling any i?

$$\hat{\Pr}(i) = \frac{C(i)+1}{C(ANY)+C(IMAGINED)}$$

$$\Rightarrow \quad \hat{\Pr}(i) = \frac{C(i)+mq}{C(ANY)+m}$$

I can use this formula with m>20, or even with $m<20$ … say with $m=1$

A better solution

Q: What if I used m rolls with a probability of $q=1/20$ of rolling any i?

$$\hat{\Pr}(i) = \frac{C(i)+1}{C(ANY)+C(IMAGINED)}$$

$$\Rightarrow \quad \hat{\Pr}(i) = \frac{C(i)+mq}{C(ANY)+m}$$

If $m>>C(ANY)$ then your imagination q rules
If $m<<C(ANY)$ then your data rules BUT you never ever ever end up with $\Pr(i)=0$

Terminology – more later

This is called a *uniform Dirichlet* prior

C(i), C(ANY) are *sufficient statistics*

$$\hat{\Pr}(i) = \frac{C(i) + mq}{C(ANY) + m}$$

MLE = maximum likelihood estimate

MAP= maximum a posteriori estimate

Probability - what you need to really, really know

- Probabilities are cool
- Random variables and events
- The Axioms of Probability
- Independence, binomials, multinomials
- Conditional probabilities
- Bayes Rule
- MLE's, smoothing, and MAPs
- The joint distribution

Some practical problems

This is a set of polyhedral dice that will roll high/low numbers. The set consists of 3 six-sided dice (1 high, 1 low, 1 standard), 2 ten-sided dice (1 high, 1 standard), and 2 twenty-sided dice (1 high, 1 standard).

- I have 1 standard d6 die, 2 loaded d6 die.
- Loaded high: P(X=6)=0.50 Loaded low: P(X=1)=0.50
- Experiment: pick one d6 uniformly at random (A) and roll it. What is more likely – rolling a seven or rolling doubles?

Three combinations: HL, HF, FL

$$P(D) = P(D \wedge A=HL) + P(D \wedge A=HF) + P(D \wedge A=FL)$$
$$= P(D \mid A=HL)*P(A=HL) + P(D \mid A=HF)*P(A=HF) + P(A \mid A=FL)*P(A=FL)$$

Some practical problems

This is a set of polyhedral dice that will roll high/low numbers. The set consists of 3 six-sided dice (1 high, 1 low, 1 standard), 2 ten-sided dice (1 high, 1 standard), and 2 twenty-sided dice (1 high, 1 standard).

- I have 1 standard d6 die, 2 loaded d6 die.
- Loaded high: P(X=6)=0.50 Loaded low: P(X=1)=0.50
- Experiment: pick one d6 uniformly at random (A) and roll it. What is more likely – rolling a seven or rolling doubles?

Three combinations: HL, HF, FL

Roll 1

Roll 2	1	2	3	4	5	6
1	D					7
2		D			7	
3			D	7		
4			7	D		
5		7			D	
6	7					D

A brute-force solution

A	Roll 1	Roll 2	P	Comment
FL	1	1	1/3 * 1/6 * ½	doubles
FL				
FL				
…				
FL				
FL				
…				
FL				es
HL				
HL				
…				
HF				es
…				

A *joint probability table* shows P(X1=x1 and … and Xk=xk) for every possible combination of values x1,x2,…., xk

With this you can compute any P(A) where A is any boolean combination of the primitive events (Xi=Xk), e.g.

- P(doubles)
- P(seven or eleven)
- P(total is higher than 5)
- ….

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

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

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.

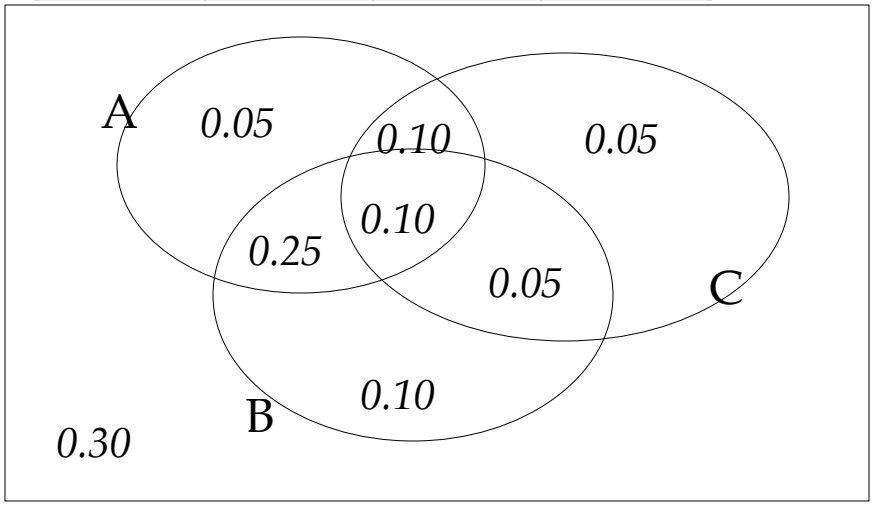
The Joint Distribution

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



Using the Joint

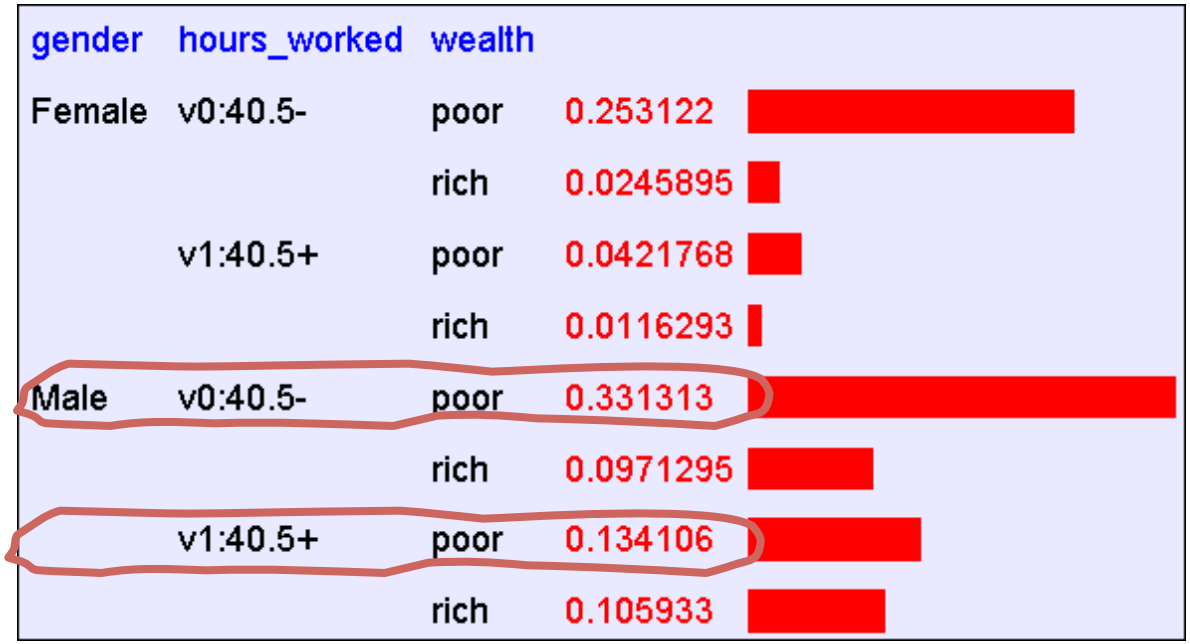
gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

One you have the JD you can ask for the probability of any logical expression involving your attribute

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

Abstract: Predict whether income exceeds $50K/yr based on census data. Also known as "Census Income" dataset. [Kohavi, 1996]
Number of Instances: 48,842
Number of Attributes: 14 (in UCI's copy of dataset); 3 (here)

Using the Joint

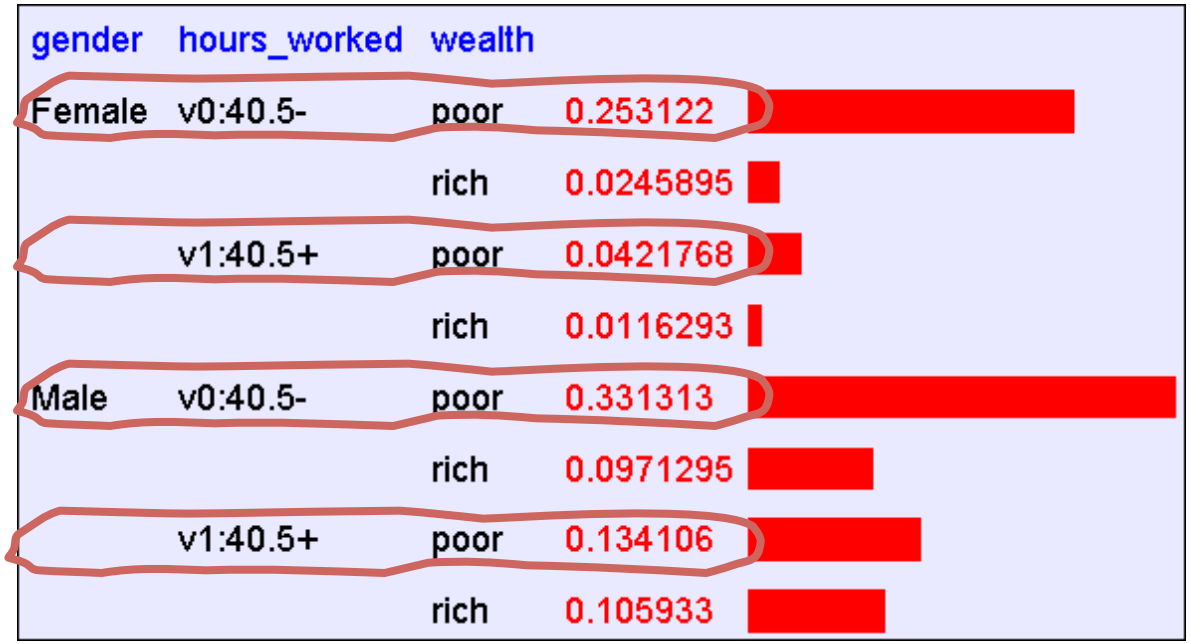


gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor Male) = 0.4654

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

Using the Joint



P(Poor) = 0.7604

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

Probability - what you need to really, really know

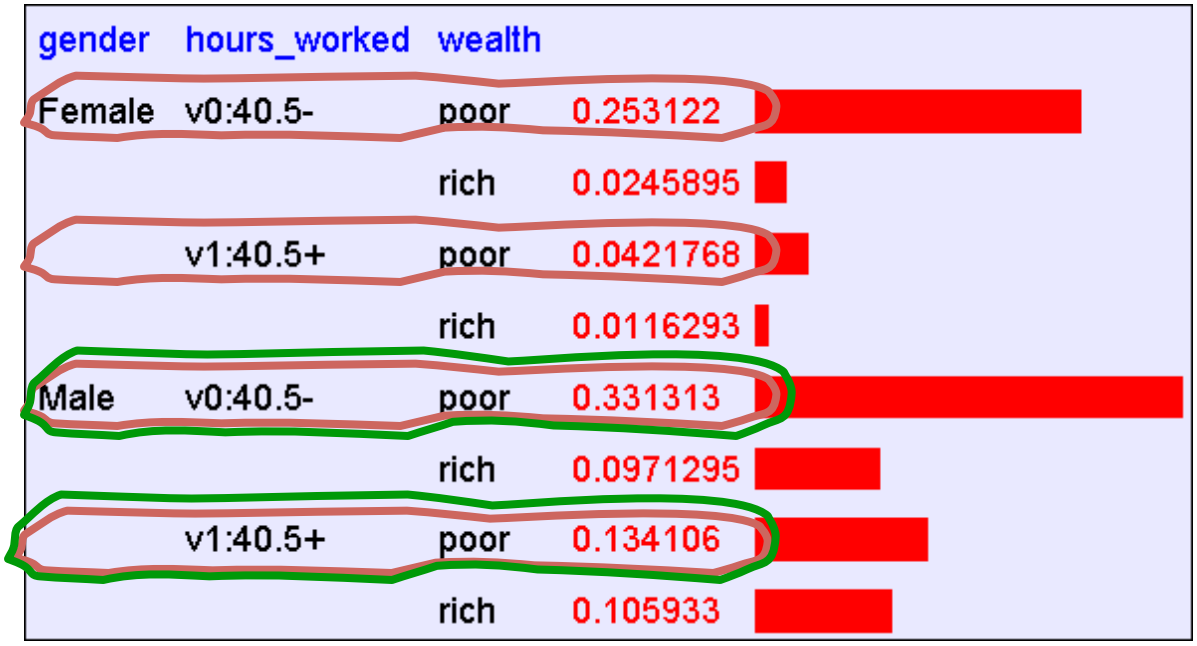
- Probabilities are cool
- Random variables and events
- The Axioms of Probability
- Independence, binomials, multinomials
- Conditional probabilities
- Bayes Rule
- MLE's, smoothing, and MAPs
- The joint distribution
- Inference

Inference with the Joint

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

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

Inference with the Joint



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

P(Male | Poor) = 0.4654 / 0.7604 = 0.612

Estimating the joint distribution

- Collect some data points
- Estimate the probability P(E1=e1 ^ … ^ En=en) as #(that row appears)/#(any row appears)
- ….

Gender	Hours	Wealth
g1	h1	w1
g2	h2	w2
..	…	…
gN	hN	wN

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

Estimating the joint distribution

- For each combination of values **r:**
  - Total = C[**r**] = 0
- For each data row $\mathbf{r_i}$
  - C[$\mathbf{r_i}$] ++
  - Total ++

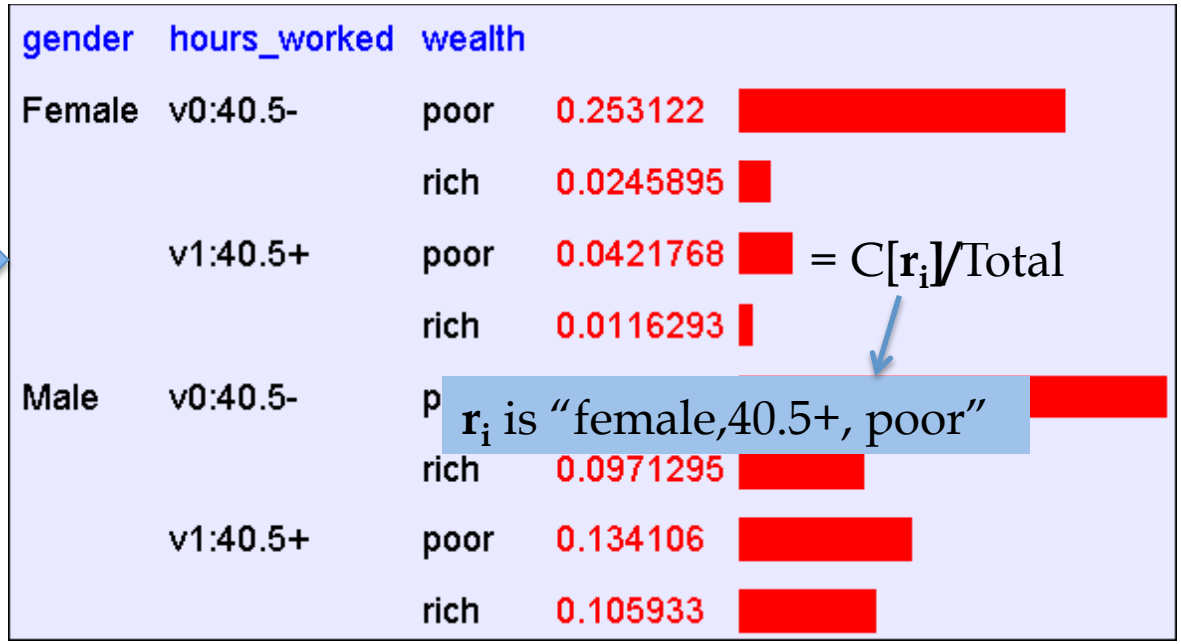
Complexity? $O(2^d)$

d = #attributes (all binary)

Complexity? $O(n)$

n = total size of input data

Gender	Hours	Wealth
g1	h1	w1
g2	h2	w2
..	…	…
gN	hN	wN



Estimating the joint distribution

- For each combination of values **r:**
  - Total = C[**r**] = 0
- For each data row $\mathbf{r_i}$
  - C[$\mathbf{r_i}$] ++
  - Total ++

Complexity? $O(\prod_{i=1}^{d} k_i)$

k_i = arity of attribute i

Complexity? O(n)

n = total size of input data

Gender	Hours	Wealth
g1	h1	w1
g2	h2	w2
..	…	…
gN	hN	wN

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

Estimating the joint distribution

- For each combination of values **r:**
  - Total = C[**r**] = 0
- For each data row $\mathbf{r_i}$
  - C[$\mathbf{r_i}$] ++
  - Total ++

Complexity? $O(\prod_{i=1}^{d} k_i)$

k_i = arity of attribute i

Complexity? O(n)

n = total size of input data

Gender	Hours	Wealth
g1	h1	w1
g2	h2	w2
..	…	…
gN	hN	wN

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

Estimating the joint distribution

- For each data row $\mathbf{r_i}$
  - If $\mathbf{r_i}$ not in hash tables C,Total:
    - **Insert** $C[\mathbf{r_i}] = 0$
  - $C[\mathbf{r_i}]$ ++
  - Total ++

Complexity? O(m)

m = size of the model

Complexity? O(n)

n = total size of input data

Gender	Hours	Wealth
g1	h1	w1
g2	h2	w2
..	…	…
gN	hN	wN

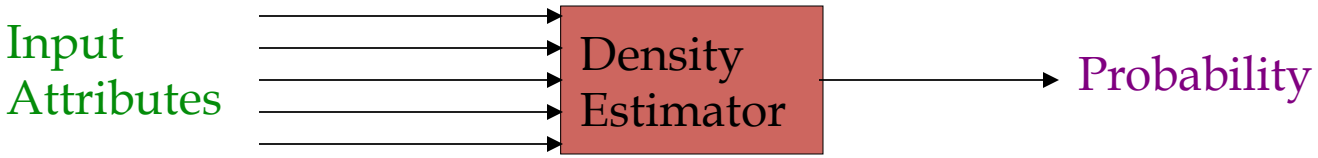
gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Probability - what you need to really, really know

- Probabilities are cool
- Random variables and events
- The Axioms of Probability
- Independence, binomials, multinomials
- Conditional probabilities
- Bayes Rule
- MLE's, smoothing, and MAPs
- The joint distribution
- Inference
- Density estimation and classification

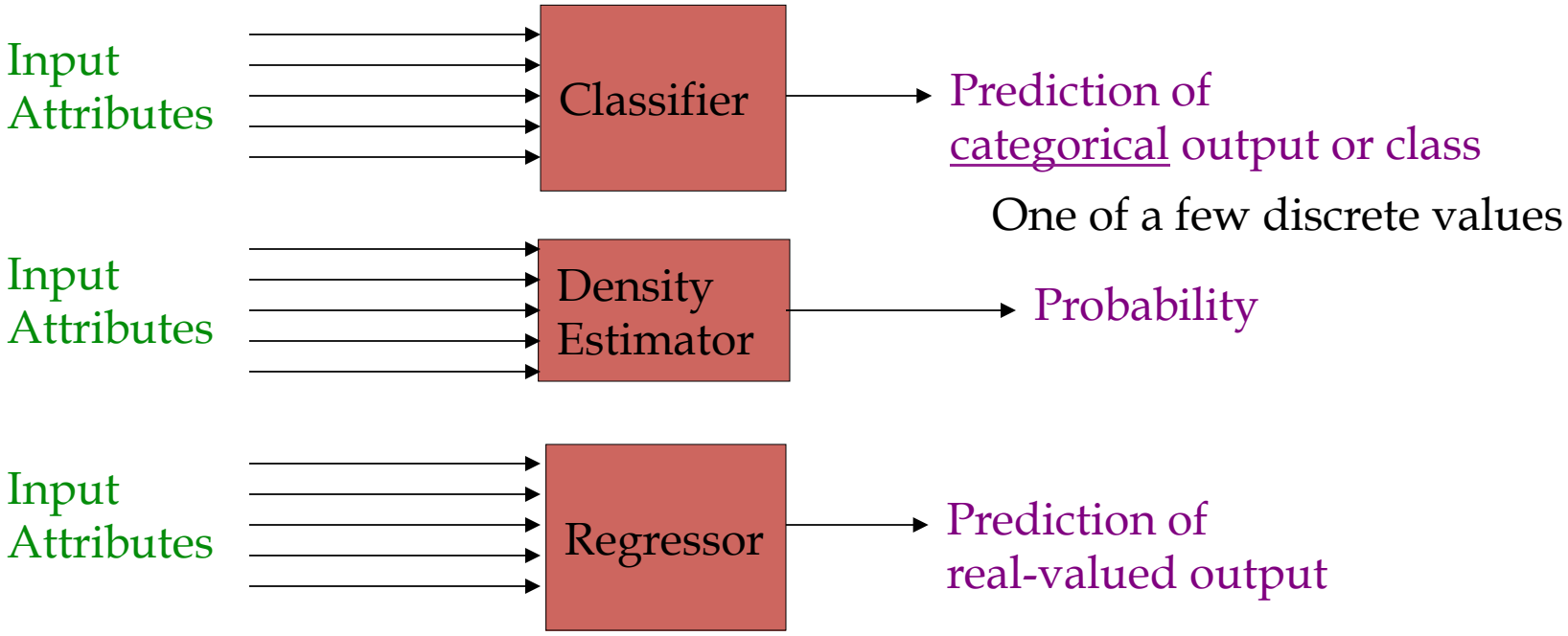
Density Estimation

- Our Joint Distribution learner is our first example of something called Density Estimation
- A Density Estimator learns a mapping from a set of attributes values to a Probability

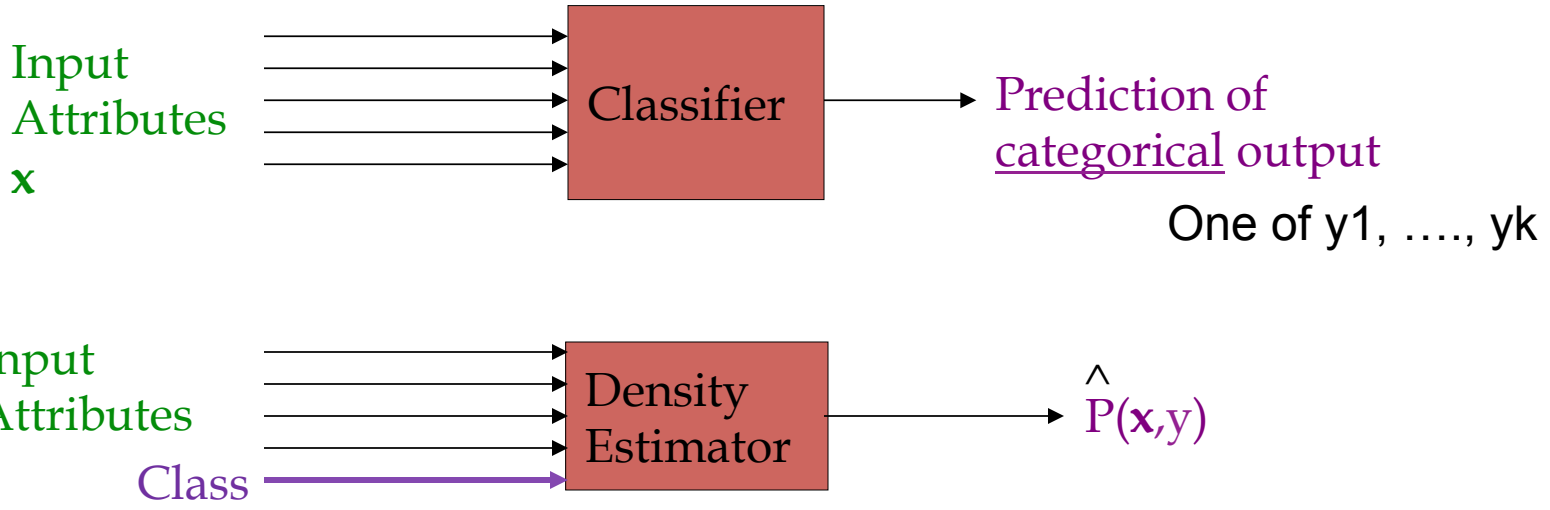


Density Estimation

- Compare it against the two other major kinds of models:



Density Estimation ➔ Classification



To classify **x**

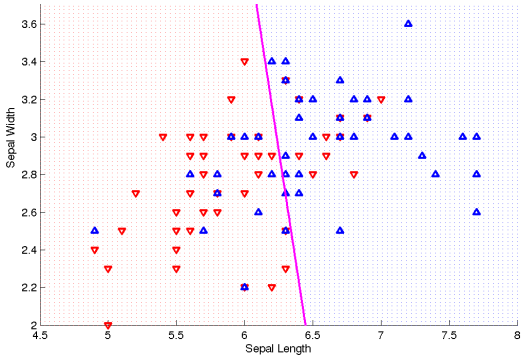
1. Use your estimator to compute $\hat{P}(\mathbf{x},y1)$, …., $\hat{P}(\mathbf{x},yk)$
2. Return the class y* with the highest predicted probability

Binary case: predict POS if $\hat{P}(\mathbf{x})>0.5$

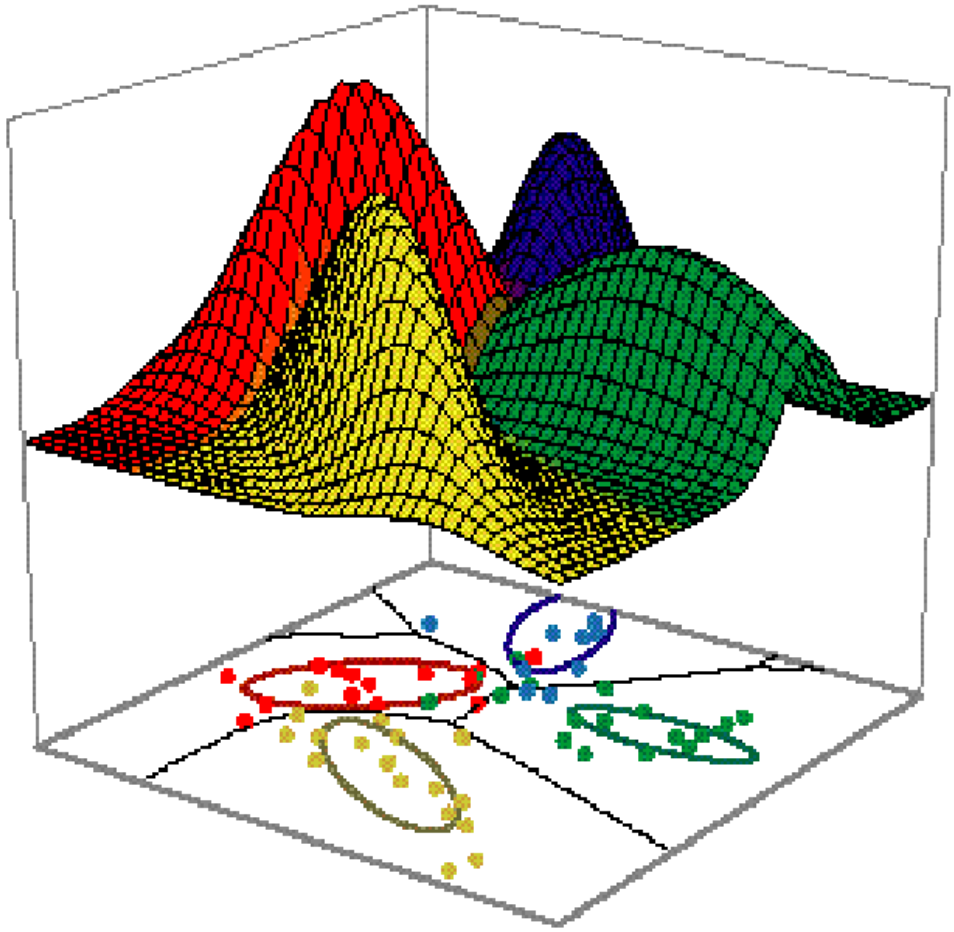
Ideally is correct with $\hat{P}(x,y^*) = \hat{P}(x,y^*)/(\hat{P}(\mathbf{x},y1) + \ldots + \hat{P}(\mathbf{x},yk))$

Classification vs Density Estimation

Classification



Density Estimation



Bayes Classifiers

- If we can do inference over $\Pr(X_1 \ldots, X_d, Y)$…
- … in particular compute $\Pr(X_1 \ldots X_d \mid Y)$ and $\Pr(Y)$.
  - And then we can use Bayes' rule to compute

$$\Pr(Y \mid X_1, \ldots, X_d) = \frac{\Pr(X_1, \ldots, X_d \mid Y) \Pr(Y)}{\Pr(X_1, \ldots, X_d)}$$

Summary: The Bad News

- Density estimation by directly learning the joint is trivial, mindless and dangerous

Andrew's joke

- Density estimation by directly learning the joint is hopeless unless you have some combination of
  - Very few attributes
  - Attributes with low "arity"
  - Lots and lots of data
- Otherwise you can't estimate all the row frequencies

Part of a Joint Distribution

A	B	C	D	E	p
is	the	effect	of	the	0.00036
is	the	effect	of	a	0.00034
.	The	effect	of	this	0.00034
to	this	effect	:	“	0.00034
be	the	effect	of	the	…
	…	…	…	…	…
not	the	effect	of	any	0.00024
	…	…	…	…	…
does	not	affect	the	general	0.00020
does	not	affect	the	question	0.00020
any	manner	affect	the	principle	0.00018

Probability - what you need to really, really know

- Probabilities are cool
- Random variables and events
- The Axioms of Probability
- Independence, binomials, multinomials
- Conditional probabilities
- Bayes Rule
- MLE's, smoothing, and MAPs
- The joint distribution
- Inference
- Density estimation and classification
- Naïve Bayes density estimators and classifiers

Naïve Density Estimation

The problem with the Joint Estimator is that it just mirrors the training data.

(and is also really, really hard to compute)

We need something which generalizes more usefully.

The naïve model generalizes strongly:

Assume that each attribute is distributed independently of any of the other attributes.

Naïve Distribution General Case

- Suppose $X_1, X_2, \ldots, X_d$ are independently distributed.

$$\Pr(X_1 = x_1, \ldots, X_d = x_d) = \Pr(X_1 = x_1) \cdot \ldots \cdot \Pr(X_d = x_d)$$

- So if we have a Naïve Distribution we can construct any row of the implied Joint Distribution on demand.

- How do we learn this?

Learning a Naïve Density Estimator

$$P(X_i = x_i) = \frac{\#\text{records with } X_i = x_i}{\#\text{records}} \qquad \text{MLE}$$

$$P(X_i = x_i) = \frac{\#\text{records with } X_i = x_i + mq}{\#\text{records} + m} \qquad \text{Dirichlet (MAP)}$$

Another trivial learning algorithm!

Probability - what you need to really, really know

- Probabilities are cool
- Random variables and events
- The Axioms of Probability
- Independence, binomials, multinomials
- Conditional probabilities
- Bayes Rule
- MLE's, smoothing, and MAPs
- The joint distribution
- Inference
- Density estimation and classification
- Naïve Bayes density estimators and classifiers
- **Conditional independence…more on this tomorrow!**