

Probability review

Some slides taken from:

Andrew Moore (<http://www.cs.cmu.edu/~awm/tutorials>)

and

Tom Mitchell (http://www.cs.cmu.edu/~tom/10701_sp11/slides/MLE_MAP_1-18-11-ann.pdf)

Discrete random variables

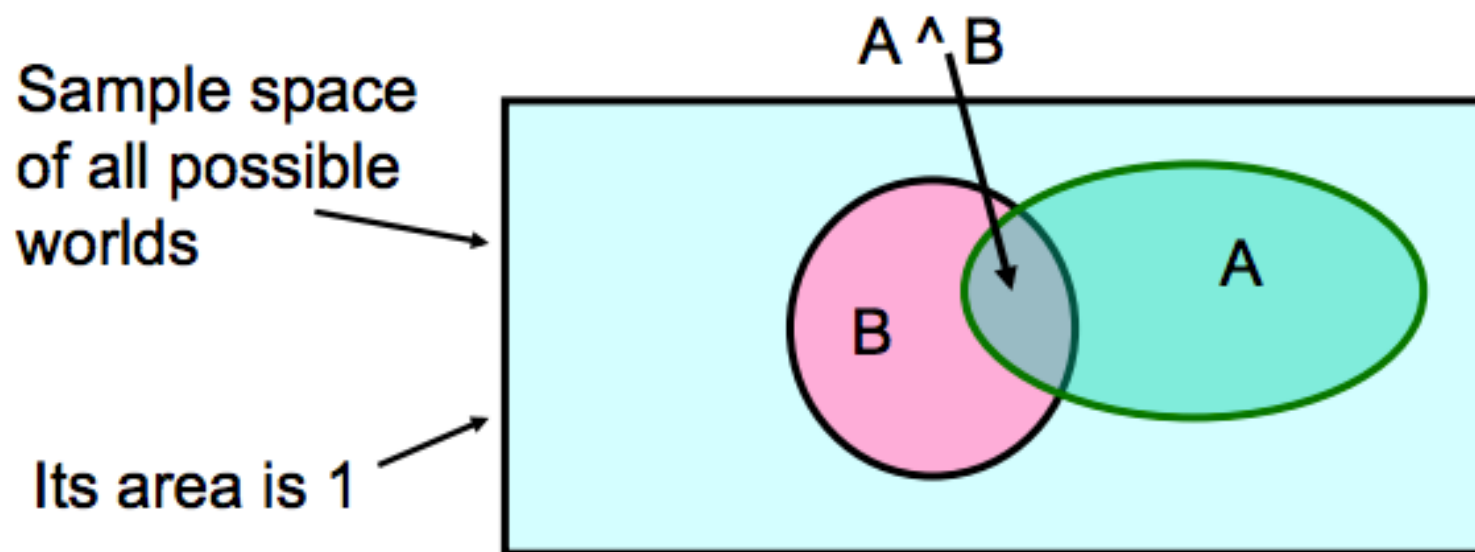
- A is a Boolean-valued random variable if A denotes an event, and there is some degree of uncertainty as to whether A occurs.

Examples

- A = The US president in 2023 will be male
- A = You wake up tomorrow with a headache
- A = You have Flu

Probabilities

- We write $P(A)$ as “the fraction of possible worlds in which A is true”



The axioms of probability

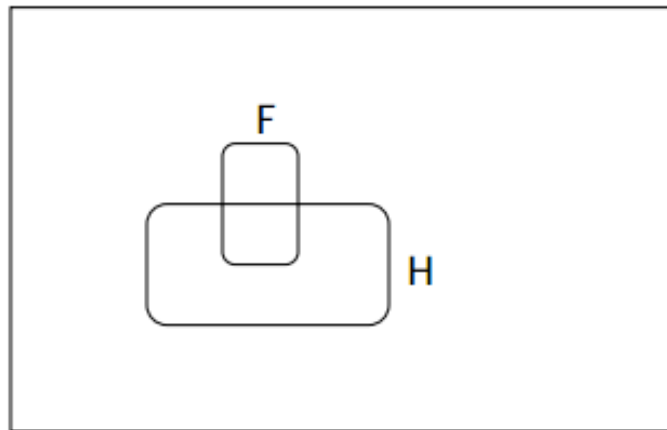
- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Conditional probability

- $P(A|B)$ = Fraction of worlds in which B is true that also have A true

H = "Have a headache"

F = "Coming down with Flu"



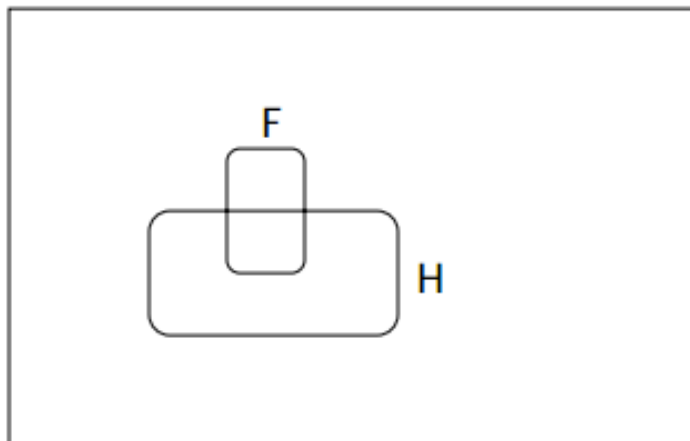
$$P(H) = 1/10$$

$$P(F) = 1/40$$

$$P(H|F) = 1/2$$

"Headaches are rare and flu is rarer, but if you're coming down with 'flu there's a 50-50 chance you'll have a headache."

Conditional probability



H = "Have a headache"
F = "Coming down with Flu"

$$P(H) = 1/10$$

$$P(F) = 1/40$$

$$P(H|F) = 1/2$$

$P(H|F)$ = Fraction of flu-inflicted worlds in which you have a headache

$$= \frac{\text{\#worlds with flu and headache}}{\text{\#worlds with flu}}$$

$$= \frac{\text{Area of "H and F" region}}{\text{Area of "F" region}}$$

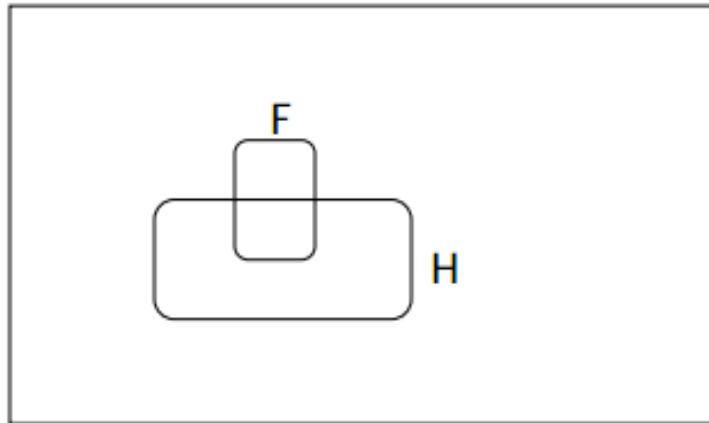
$$= \frac{P(H \wedge F)}{P(F)}$$

Conditional probability

Definition of Conditional Probability

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

Probabilistic Inference



H = "Have a headache"

F = "Coming down with Flu"

$$P(H) = 1/10$$

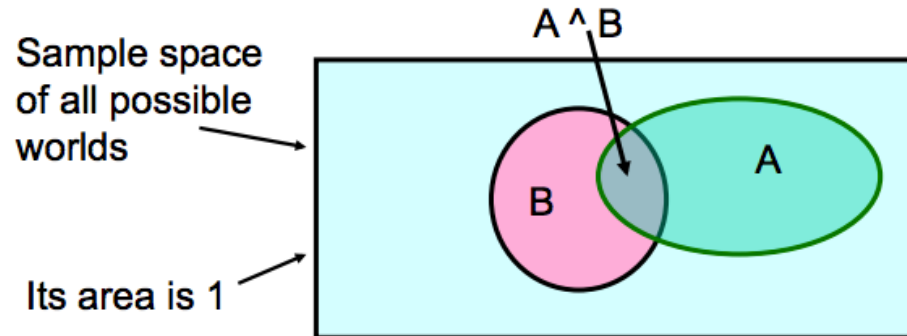
$$P(F) = 1/40$$

$$P(H|F) = 1/2$$

One day you wake up with a headache. You think: "Drat! 50% of flus are associated with headaches so I must have a 50-50 chance of coming down with flu"

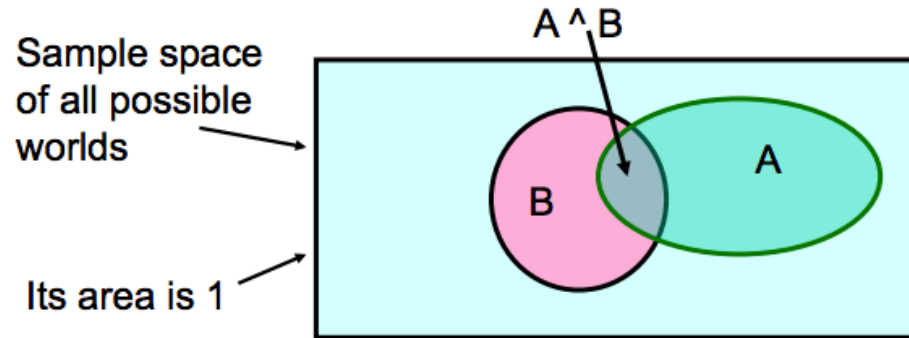
Is this reasoning good?

Deriving the Bayes Rule



- You learned that $A \wedge B = P(A) * P(B)$
- Not always! Only when A and B are independent
- An event A is independent of an event B if
 - $P(A|B) = P(A)$
 - “Knowing B, tells us nothing about A”

Deriving the Bayes Rule

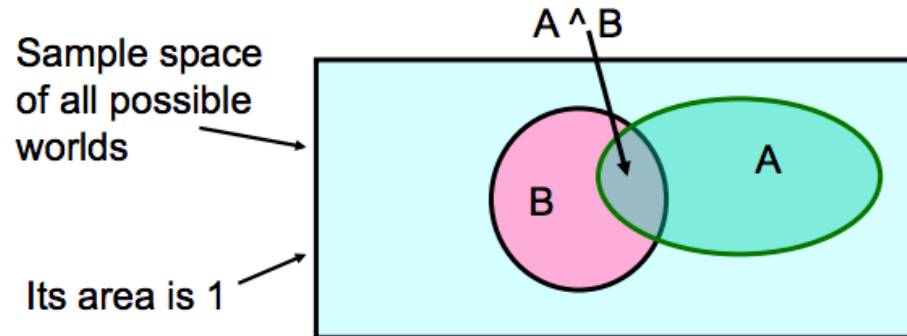


- More generally $B \wedge A = P(B|A) P(A)$

- The chain rule:

$$P(C \wedge B \wedge A) = P(C|B \wedge A) P(B|A) P(A)$$

Deriving the Bayes Rule



- More generally $A \wedge B = P(B|A) P(A)$
- $A \wedge B$ is commutative ($A \wedge B = B \wedge A$)
- Then $P(A|B) P(B) = P(B|A) P(A)$
- Bayes rule: $P(A|B) = P(B|A) P(A) / P(B)$

Bayes rule

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

This is Bayes Rule

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**



Case study 1: Cookie bowls

- Suppose there are two bowls of cookies. Bowl 1 contains 30 vanilla cookies and 10 chocolate cookies. Bowl 2 contains 20 of each.
- Now suppose you choose one of the bowls at random and, without looking, select a cookie at random. The cookie is vanilla. What is the probability that it came from Bowl 1?
- This is a conditional probability; we want $p(\text{Bowl 1} | \text{vanilla})$, but it is not obvious how to compute it.
- If I asked a different question—the probability of a vanilla cookie given Bowl 1—it would be easy:
 - $p(\text{vanilla} | \text{Bowl 1}) = 3/4$
 - Sadly, $p(A | B)$ is *not* the same as $p(B | A)$, but there is a way to get from one to the other: Bayes's theorem.

Case study 2: M&Ms

- In 1995, they introduced blue M&M's. Before then, the color mix in a bag of plain M&M's was 30% Brown, 20% Yellow, 20% Red, 10% Green, 10% Orange, 10% Tan. Afterward it was 24% Blue, 20% Green, 16% Orange, 14% Yellow, 13% Red, 13% Brown.
- Suppose I have two bags of M&M's, and I tell you that one is from 1994 and one from 1996. I won't tell you which is which, but I give you one M&M from each bag.
- One is yellow and one is green. What is the probability that the yellow one came from the 1994 bag?

Case study 3: Monty Hall

- Monty shows you three closed doors and tells you that there is a prize behind each door: one prize is a car, the other two are less valuable prizes like peanut butter and fake finger nails. The prizes are arranged at random. The objective of the game is to guess which door has the car. If you guess right, you get to keep the car.
- You pick a door, which we will call Door A. The other doors are B and C.
- Before opening the door you chose, Monty increases the suspense by opening either Door B or C, whichever does not have the car. (If the car is actually behind Door A, Monty can safely open B or C, so he chooses one at random.)
- Then Monty offers you the option to stick with your original choice or switch to the one remaining unopened door.
- The question is, should you “stick” or “switch” or does it make no difference?

What does all this have to do with function approximation?

- Learn $f: X \rightarrow Y$
- Similar to learning $P(Y|X)$

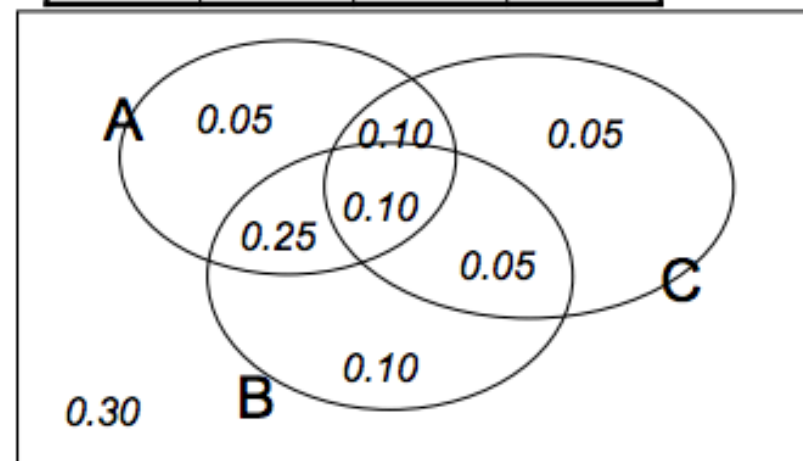
The Joint Distribution

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	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



[A. Moore]

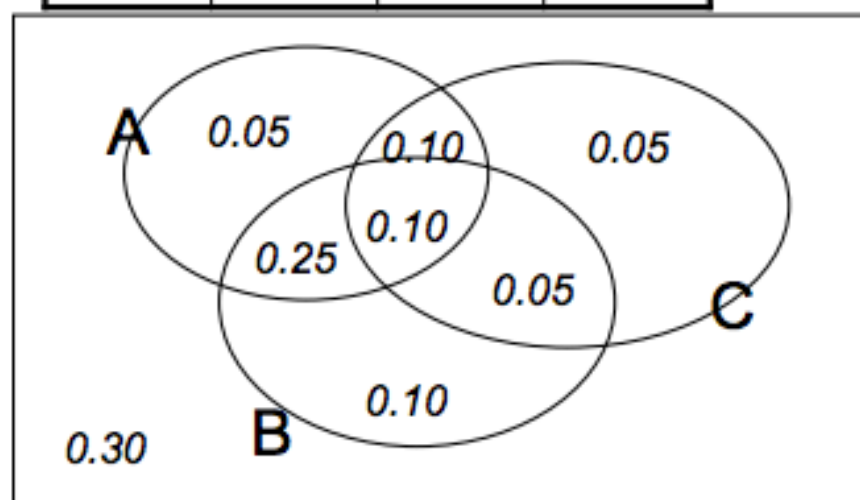
The Joint Distribution

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



[A. Moore]

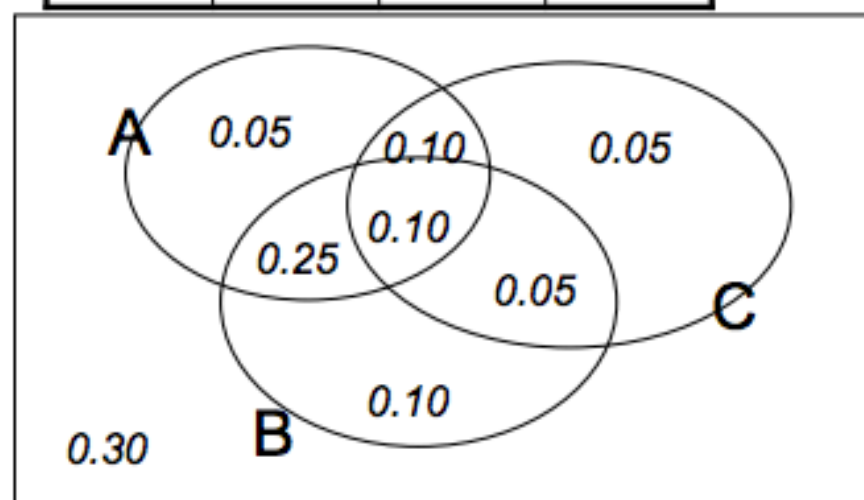
The Joint Distribution

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



[A. Moore]

Learning a joint distribution

Build a JD table for your attributes in which the probabilities are unspecified

A	B	C	Prob
0	0	0	?
0	0	1	?
0	1	0	?
0	1	1	?
1	0	0	?
1	0	1	?
1	1	0	?
1	1	1	?

Fraction of all records in which
A and B are True but C is False

The fill in each row with

$$\hat{P}(\text{row}) = \frac{\text{records matching row}}{\text{total number of records}}$$

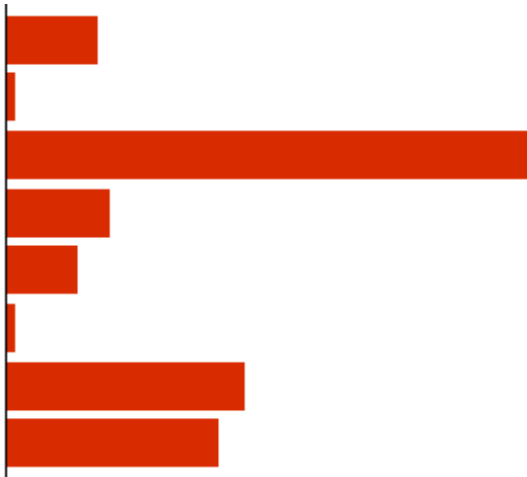
A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

Case study: breast cancer

- The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer
- Number of Instances: 306
- Attribute Information:
 - Number of positive axillary nodes detected (numerical)
 - Age of patient at time of operation (numerical)
- Class:
 - Survival status (-1,1)

Nodes	Age	Survived
4	53	-1
1	58	-1
3	33	-1
9	41	-1
24	45	-1
12	45	-1
1	52	1
1	38	1
2	35	1
1	42	1

Joint distribution

Only 1 node	Up to 40	Survived 5+	0.0719	
		Died within 5	0.0065	
	40+	Survived 5+	0.4183	
		Died within 5	0.0817	
2 or more nodes	Up to 40	Survived 5+	0.0556	
		Died within 5	0.0065	
	40+	Survived 5+	0.1895	
		Died within 5	0.1699	

Using the joint

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

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

Only 1 node	Up to 40	Survived 5+	0.0719
		Died within 5	0.0065
	40+	Survived 5+	0.4183
		Died within 5	0.0817
2 or more nodes	Up to 40	Survived 5+	0.0556
		Died within 5	0.0065
	40+	Survived 5+	0.1895
		Died within 5	0.1699

$$\begin{aligned} P(\text{finding only 1 node}) &= \\ &0.0719 + 0.0065 + 0.4183 + 0.0817 \\ &= 0.5784 \end{aligned}$$

Using the joint

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

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

Only 1 node	Up to 40	Survived 5+	0.0719
		Died within 5	0.0065
	40+	Survived 5+	0.4183
		Died within 5	0.0817
2 or more nodes	Up to 40	Survived 5+	0.0556
		Died within 5	0.0065
	40+	Survived 5+	0.1895
		Died within 5	0.1699

P(finding only 1 node and survived)=
 0.0719+0.4183
 = 0.4902

Inference with the joint

Only 1 node	Up to 40	Survived 5+	0.0719
		Died within 5	0.0065
	40+	Survived 5+	0.4183
		Died within 5	0.0817
2 or more nodes	Up to 40	Survived 5+	0.0556
		Died within 5	0.0065
	40+	Survived 5+	0.1895
		Died within 5	0.1699

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

Inference with the joint

Only 1 node	Up to 40	Survived 5+	0.0719
		Died within 5	0.0065
	40+	Survived 5+	0.4183
		Died within 5	0.0817
2 or more nodes	Up to 40	Survived 5+	0.0556
		Died within 5	0.0065
	40+	Survived 5+	0.1895
		Died within 5	0.1699

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

$$P(\text{survived} \mid 2+\text{ nodes}) = \frac{0.0556 + 0.1895}{0.0556 + 0.0065 + 0.1895 + 0.1699} = 0.5052$$

Inference with the joint

Only 1 node	Up to 40	Survived 5+	0.0719
		Died within 5	0.0065
	40+	Survived 5+	0.4183
		Died within 5	0.0817
2 or more nodes	Up to 40	Survived 5+	0.0556
		Died within 5	0.0065
	40+	Survived 5+	0.1895
		Died within 5	0.1699

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

$$P(\text{survived} | 1 \text{ node}) = \frac{0.0719 + 0.4183}{0.0719 + 0.0065 + 0.4183 + 0.0817} = 0.7951$$

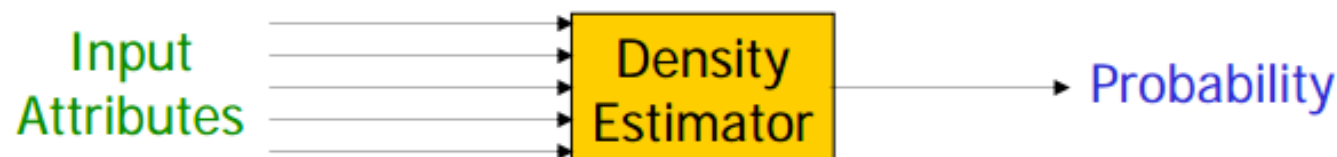
Learning and the Joint

- Suppose we want to learn the function
 $f: \langle \text{Nodes}, \text{Age} \rangle \rightarrow \text{Survival}$
- Equivalently, $P(\text{Survival} \mid \text{Nodes}, \text{Age})$

One solution: learn joint distribution from data,
calculate $P(\text{Survival} \mid \text{Nodes}, \text{Age})$

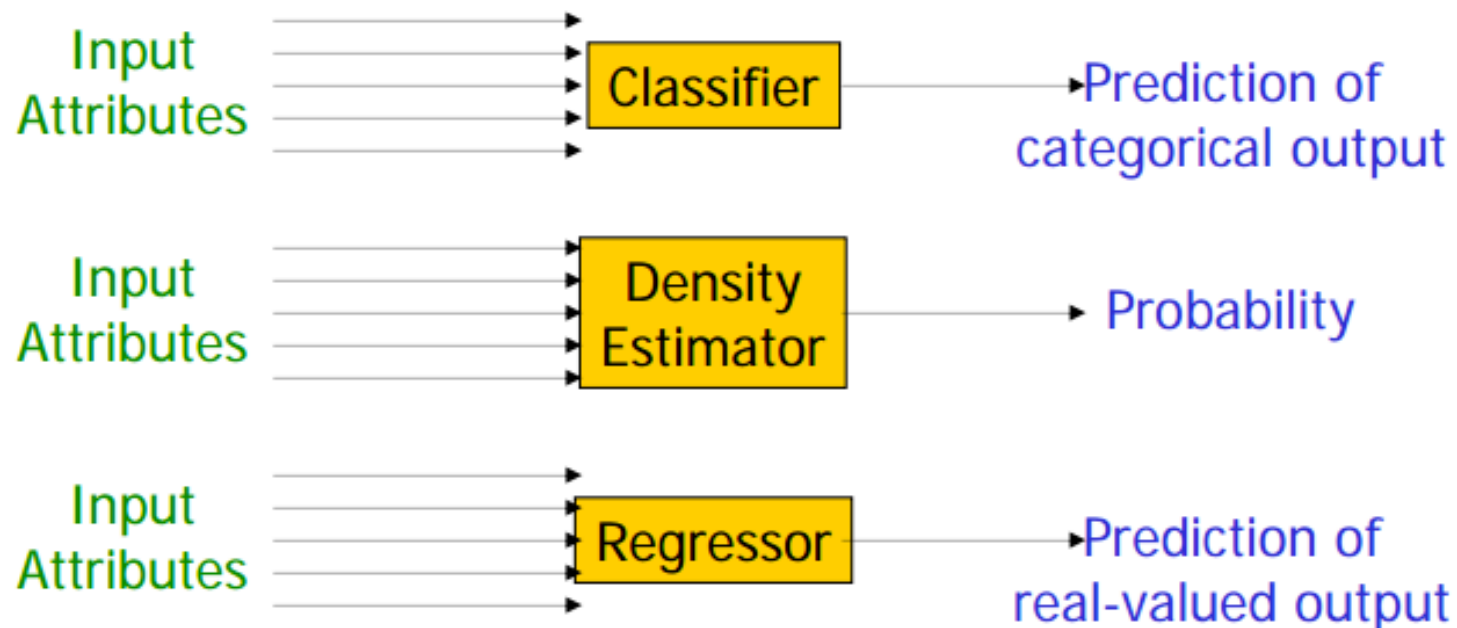
Density Estimation

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



Density Estimation

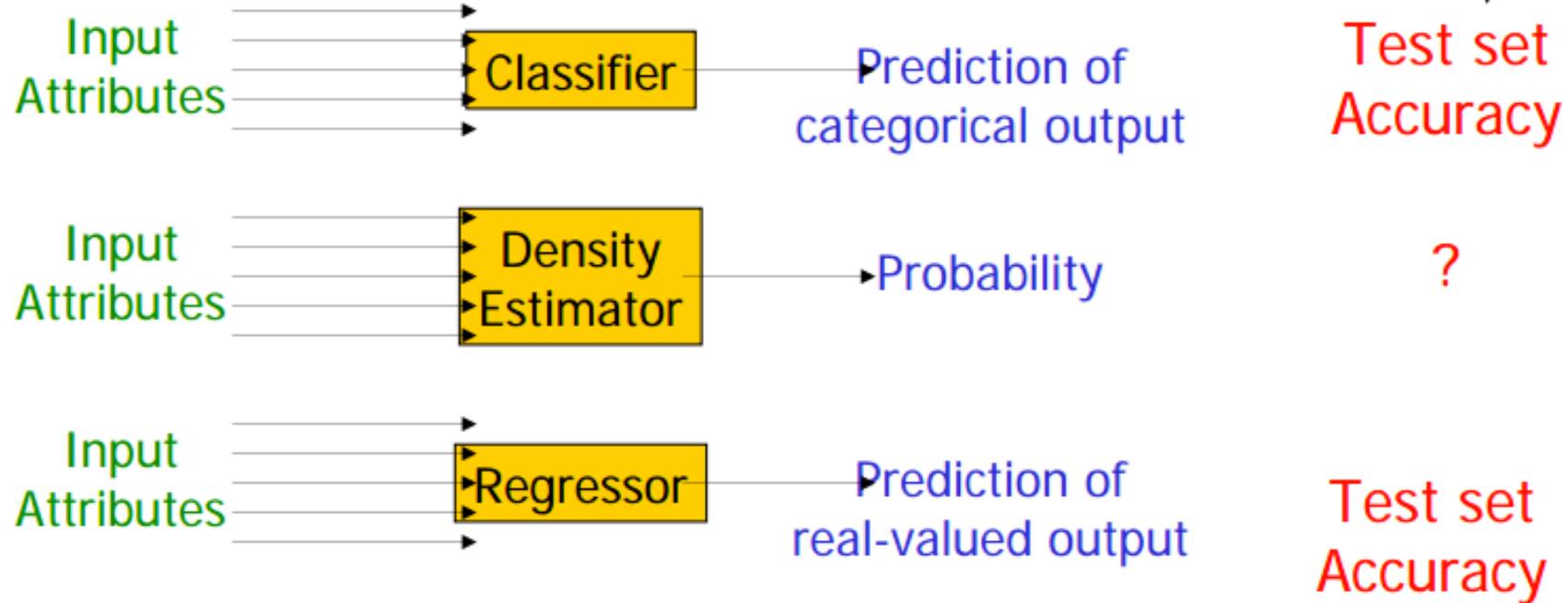
- Compare it against the two other major kinds of models:



Evaluating Density Estimation

Test-set criterion for estimating performance on future data*

** See the Decision Tree or Cross Validation lecture for more detail*



Evaluating a density estimator

- Given a record \mathbf{x} , a density estimator M can tell you how likely the record is:

$$\hat{P}(\mathbf{x}|M)$$

- Given a dataset with R records, a density estimator can tell you how likely the dataset is:

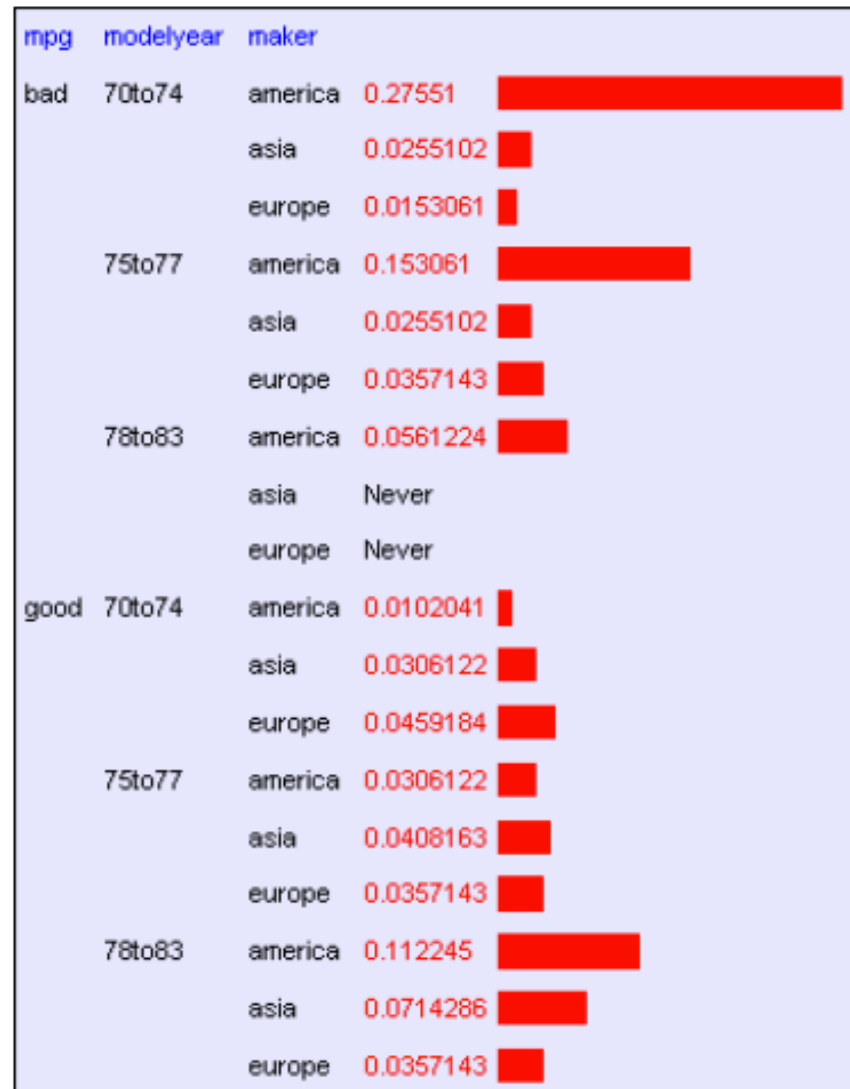
(Under the assumption that all records were **independently** generated from the Density Estimator's JD)

$$\hat{P}(\text{dataset}|M) = \hat{P}(\mathbf{x}_1 \wedge \mathbf{x}_2 \dots \wedge \mathbf{x}_R|M) = \prod_{k=1}^R \hat{P}(\mathbf{x}_k|M)$$

A small dataset: Miles Per Gallon

192
Training
Set
Records

mpg	modelyear	maker
good	75to78	asia
bad	70to74	america
bad	75to78	europa
bad	70to74	america
bad	70to74	america
bad	70to74	asia
bad	70to74	asia
bad	75to78	america
:	:	:
:	:	:
:	:	:
bad	70to74	america
good	79to83	america
bad	75to78	america
good	79to83	america
bad	75to78	america
good	79to83	america
good	79to83	america
bad	70to74	america
good	75to78	europa
bad	75to78	europa



A small dataset: Miles Per Gallon

192
Training
Set

mpg	modelyear	maker
good	75to78	asia
bad	70to74	america
bad	75to78	europa
bad	70to74	america
bad	70to74	america
bad	70to74	asia
bad	70to74	asia

mpg	modelyear	maker		
bad	70to74	america	0.27551	<div></div>
		asia	0.0255102	<div></div>
		europa	0.0153061	<div></div>
	75to77	america	0.153061	<div></div>
		asia	0.0255102	<div></div>
		europa	0.0357143	<div></div>

$$\hat{P}(\text{dataset}|M) = \hat{P}(\mathbf{x}_1 \wedge \mathbf{x}_2 \dots \wedge \mathbf{x}_R|M) = \prod_{k=1}^R \hat{P}(\mathbf{x}_k|M)$$

$$= (\text{in this case}) = 3.4 \times 10^{-203}$$

70to74	america	0.112245	<div></div>
	asia	0.0714286	<div></div>
	europa	0.0357143	<div></div>

Log Probabilities

Since probabilities of datasets get so small we usually use log probabilities

$$\log \hat{P}(\text{dataset}|M) = \log \prod_{k=1}^R \hat{P}(\mathbf{x}_k|M) = \sum_{k=1}^R \log \hat{P}(\mathbf{x}_k|M)$$

A small dataset: Miles Per Gallon

192
Training
Set

mpg	modelyear	maker
good	75to78	asia
bad	70to74	america
bad	75to78	europa
bad	70to74	america
bad	70to74	america
bad	70to74	asia
bad	70to74	asia

mpg	modelyear	maker		
bad	70to74	america	0.27551	<div></div>
		asia	0.0255102	<div></div>
		europa	0.0153061	<div></div>
	75to77	america	0.153061	<div></div>
		asia	0.0255102	<div></div>
		europa	0.0357143	<div></div>

$$\log \hat{P}(\text{dataset}|M) = \log \prod_{k=1}^R \hat{P}(\mathbf{x}_k|M) = \sum_{k=1}^R \log \hat{P}(\mathbf{x}_k|M)$$

= (in this case) = -466.19

70to74	america	0.112245	<div></div>
	asia	0.0714286	<div></div>
	europa	0.0357143	<div></div>

Summary: The good news

- We have a way to learn a Density Estimator from data.
- Density estimators can do many good things...
 - Can sort the records by probability, and thus spot weird records (anomaly detection)
 - Can do inference: $P(E1 | E2)$
 - Automatic Doctor / Help Desk etc

Ingredient for Bayes Classifiers (see later)

Summary: The bad news

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

Using a test set

	Set Size	Log likelihood
Training Set	196	-466.1905
Test Set	196	-614.6157

An independent test set with 196 cars has a worse log likelihood

(actually it's a billion quintillion quintillion quintillion quintillion times less likely)

....Density estimators can overfit. And the full joint density estimator is the overfittest of them all!

Overfitting Density Estimators

If **this** ever happens, it means there are certain combinations that we learn are impossible

mpg	modelyear	maker		
bad	70to74	america	0.27551	<div></div>
		asia	0.0255102	<div></div>
		europa	0.0153061	<div></div>
	75to77	america	0.153061	<div></div>
		asia	0.0255102	<div></div>
		europa	0.0357143	<div></div>
	78to83	america	0.0561224	<div></div>
		asia	Never	<div></div>
		europa	Never	<div></div>
good	70to74	america	0.0102041	<div></div>
		asia	0.000402	<div></div>
		europa	0.0357143	<div></div>

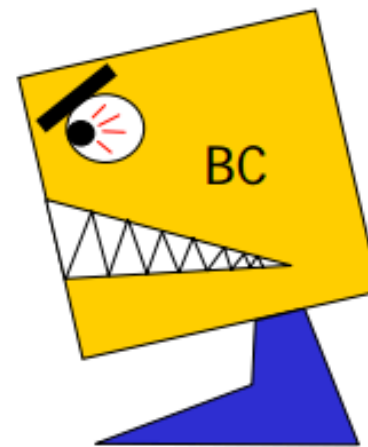
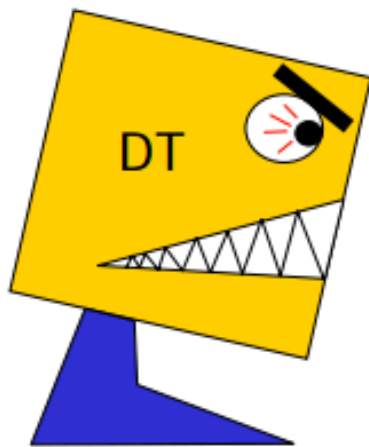
$$\log \hat{P}(\text{testset}|M) = \log \prod_{k=1}^R \hat{P}(\mathbf{x}_k|M) = \sum_{k=1}^R \log \hat{P}(\mathbf{x}_k|M)$$

$$= -\infty \text{ if for any } k \hat{P}(\mathbf{x}_k|M) = 0$$

We need Density Estimators that are
less prone to overfitting

Bayes Classifiers

- A formidable and sworn enemy of decision trees



But first... MLE and MAP