Probability Basics, Density Estimation

These slides were assembled by Byron Boots, with only minor modifications from Eric Eaton's slides and grateful acknowledgement to the many others who made their course materials freely available online. Feel free to reuse or adapt these slides for your own academic purposes, provided that you include proper attribution.

The Joint Distribution

Recipe for making a joint distribution of *d* variables:

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

1. For each combination of values, say how probable it is.

2. If you subscribe to the axioms of probability, those numbers must sum to 1.

e.g., 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

Inferring Marginal Probabilities from the Joint

	alarm		¬alarm	
	earthquake	¬earthquake	earthquake	¬earthquake
burglary	0.01	0.08	0.001	0.009
¬burglary	0.01	0.09	0.01	0.79

$$P(alarm) = \sum_{b,e} P(alarm \wedge \text{Burglary} = b \wedge \text{Earthquake} = e)$$

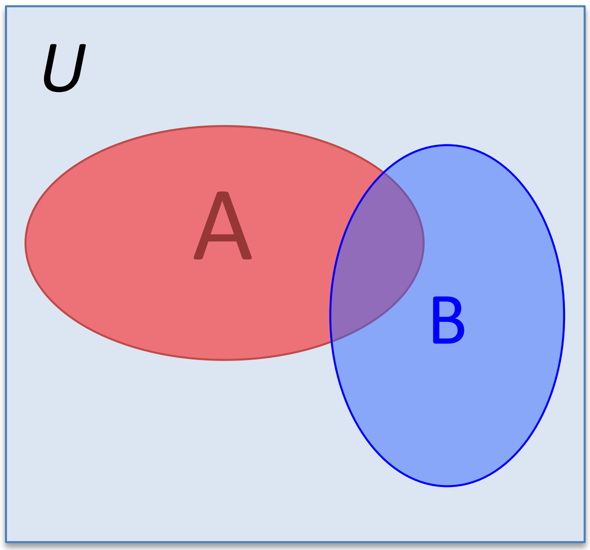
$$= 0.01 + 0.08 + 0.01 + 0.09 = 0.19$$

$$P(burglary) = \sum_{a,e} P(\text{Alarm} = a \wedge burglary \wedge \text{Earthquake} = e)$$

$$= 0.01 + 0.08 + 0.001 + 0.009 = 0.1$$

Conditional Probability

- $P(A \mid B)$ = Probability that A is true given B is true



What if we already know that *B* is true?

That knowledge changes the probability of *A*

- Because we know we're in a world where *B* is true

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

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

Example: Conditional Probabilities

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

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

P(Alarm, Burglary) =

	alarm	¬alarm
burglary	0.09	0.01
¬burglary	0.1	0.8

P(burglary | alarm) = P(burglary ∧ alarm) / P(alarm)
= 0.09 / 0.19 = 0.47

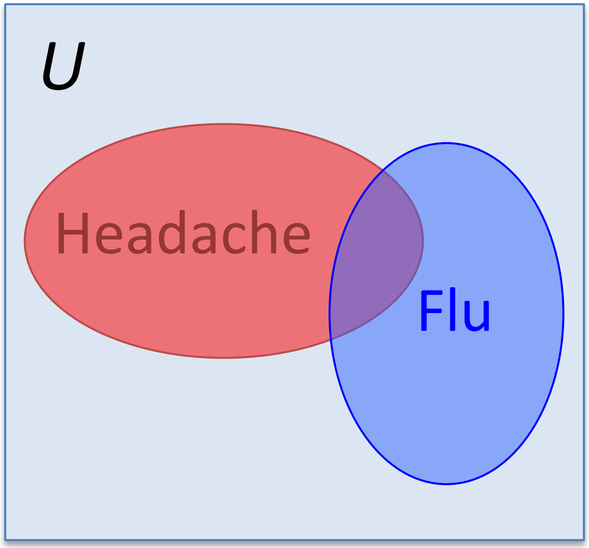
P(alarm | burglary) = P(burglary ∧ alarm) / P(burglary)
= 0.09 / 0.1 = 0.9

P(burglary ∧ alarm) = P(burglary | alarm) P(alarm)
= 0.47 * 0.19 = 0.09

Example: Inference from Conditional Probability

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

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



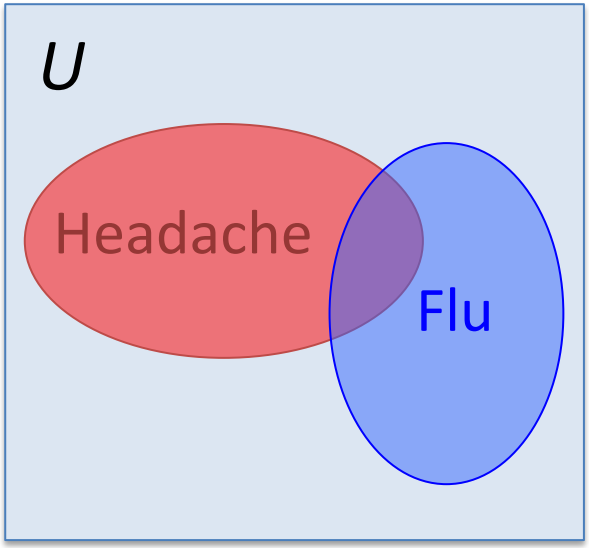
P(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 1/2

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

Example: Inference from Conditional Probability

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

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



P(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 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?

Example: Inference from Conditional Probability

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

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

P(headache) = 1/10

P(flu) = 1/40

P(headache | flu) = 1/2

Want to solve for:

$P(\text{headache} \wedge \text{flu}) = ?$

$P(\text{flu} \mid \text{headache}) = ?$

$$\begin{aligned} P(\text{headache} \wedge \text{flu}) &= P(\text{headache} \mid \text{flu}) \times P(\text{flu}) \\ &= 1/2 \times 1/40 = 0.0125 \end{aligned}$$

$$\begin{aligned} P(\text{flu} \mid \text{headache}) &= P(\text{headache} \wedge \text{flu}) \,/\, P(\text{headache}) \\ &= 0.0125 \,/\, 0.1 = 0.125 \end{aligned}$$

Based on example by Andrew Moore

Bayes' Rule

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

- Exactly the process we just used
- The most important formula in probabilistic machine learning

(Super Easy) Derivation:

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

these are the same

Just set equal...

$$P(A \mid B) \times P(B) = P(B \mid A) \times P(A)$$

and solve...

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**

Bayes' Rule

- Allows us to reason from **evidence** to **hypotheses**
- Another way of thinking about Bayes' rule:

$$P(\text{hypothesis} \mid \text{evidence}) = \frac{P(\text{evidence} \mid \text{hypothesis}) \times P(\text{hypothesis})}{P(\text{evidence})}$$

In the flu example:

P(headache) = 1/10 P(flu) = 1/40

P(headache | flu) = 1/2

Given evidence of headache, what is P(flu | headache) ?

Solve via Bayes rule!

Independence

- When two sets of propositions do not affect each others' probabilities, we call them **independent**
- Formal definition:

$$
\begin{aligned}
A \perp\!\!\!\perp B \quad &\leftrightarrow \quad P(A \wedge B) = P(A) \times P(B) \\
&\leftrightarrow \quad P(A \mid B) = P(A)
\end{aligned}
$$

For example, {moon-phase, light-level} might be independent of {burglary, alarm, earthquake}

- Then again, maybe not: Burglars might be more likely to burglarize houses when there's a new moon (and hence little light)
- But if we know the light level, the moon phase doesn't affect whether we are burglarized

Exercise: Independence

P(smart ∧ study ∧ prep)	smart		¬smart	
	study	**¬study**	**study**	**¬study**
prepared	0.432	0.16	0.084	0.008
¬prepared	0.048	0.16	0.036	0.072

Is *smart* independent of *study*?

Is *prepared* independent of *study*?

Exercise: Independence

$P(smart \wedge study \wedge prep)$	**smart**		**¬smart**	
	study	**¬study**	**study**	**¬study**
prepared	0.432	0.16	0.084	0.008
¬prepared	0.048	0.16	0.036	0.072

Is *smart* independent of *study*?

$P(study \wedge smart) = 0.432 + 0.048 = \boxed{0.48}$

$P(study) = 0.432 + 0.048 + 0.084 + 0.036 = 0.6$

$P(smart) = 0.432 + 0.048 + 0.16 + 0.16 = 0.8$

$P(study) \times P(smart) = 0.6 \times 0.8 = \boxed{0.48}$

So yes!

Is *prepared* independent of *study*?

Conditional Independence

- Absolute independence of *A* and *B*:

$$A \perp\!\!\!\perp B \quad \leftrightarrow \quad P(A \wedge B) = P(A) \times P(B)$$
$$\leftrightarrow \quad P(A \mid B) = P(A)$$

Conditional independence of *A* and *B* given *C*

$$A \perp\!\!\!\perp B \mid C \quad \leftrightarrow \quad P(A \wedge B \mid C) = P(A \mid C) \times P(B \mid C)$$

  - e.g., Moon-Phase and Burglary are ***conditionally independent given*** Light-Level

- This lets us decompose the joint distribution:

$$P(A \wedge B \wedge C) = P(A \mid C) \times P(B \mid C) \times P(C)$$

  – Conditional independence is weaker than absolute independence, but still useful in decomposing the full joint

Take Home Exercise: Conditional independence

P(smart ∧ study ∧ prep)	**smart**		**¬smart**	
	study	**¬study**	**study**	**¬study**
prepared	0.432	0.16	0.084	0.008
¬prepared	0.048	0.16	0.036	0.072

Is *smart* conditionally independent of *prepared*, given *study*?

Is *study* conditionally independent of *prepared*, given *smart*?

Summary: Essential Probability Concepts

- Marginalization: $P(B) = \sum_{v \in \text{values}(A)} P(B \wedge A = v)$
- Conditional Probability: $P(A \mid B) = \dfrac{P(A \wedge B)}{P(B)}$
- Bayes' Rule: $P(A \mid B) = \dfrac{P(B \mid A) \times P(A)}{P(B)}$
- Independence:

$$
\begin{aligned}
A \perp\!\!\!\perp B \quad &\leftrightarrow \quad P(A \wedge B) = P(A) \times P(B) \\
&\leftrightarrow \quad P(A \mid B) = P(A) \\
A \perp\!\!\!\perp B \mid C \quad &\leftrightarrow \quad P(A \wedge B \mid C) = P(A \mid C) \times P(B \mid C)
\end{aligned}
$$

Density Estimation

How Can We Obtain a Joint Distribution?

Option 1: Elicit it from an expert human

Option 2: Build it up from simpler probabilistic facts

- e.g, if we knew

  $P(a) = 0.7$ $\qquad$ $P(b|a) = 0.2$ $\qquad$ $P(b|\neg a) = 0.1$

  then, we could compute $P(a \wedge b)$

Option 3: Learn it from data...

Learning a Joint Distribution

Step 1:

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	?

Step 2:

Then, fill in each row with:

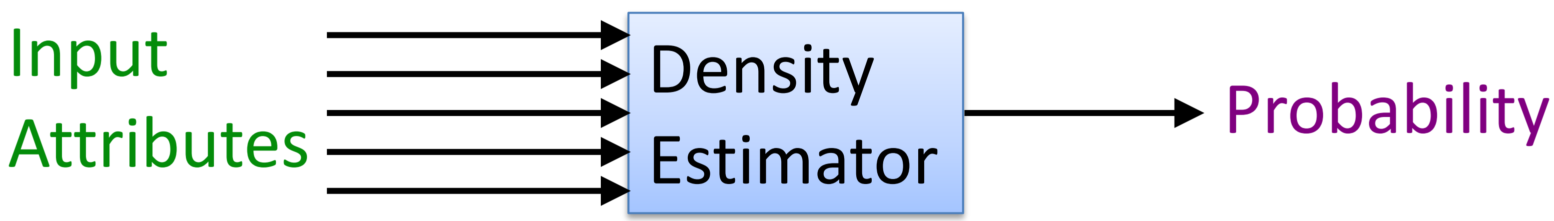
$$\hat{P}(\text{row}) = \frac{\text{records matching row}}{\text{total number of records}}$$

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

Fraction of all records in which A and B are true but C is false

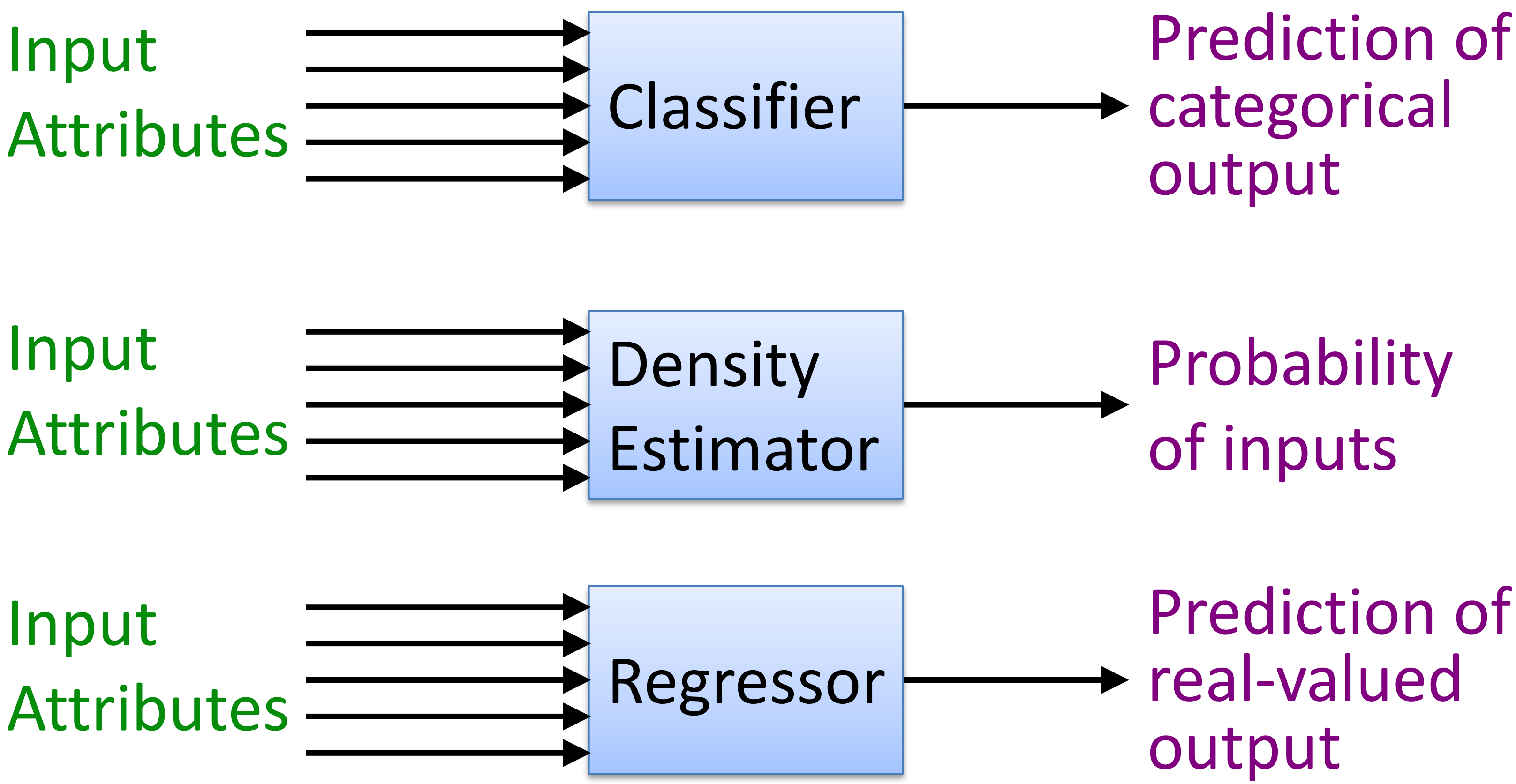
Density Estimation

- Our joint distribution learner is an example of something called **Density Estimation**
- A Density Estimator learns a mapping from a set of attributes to a probability



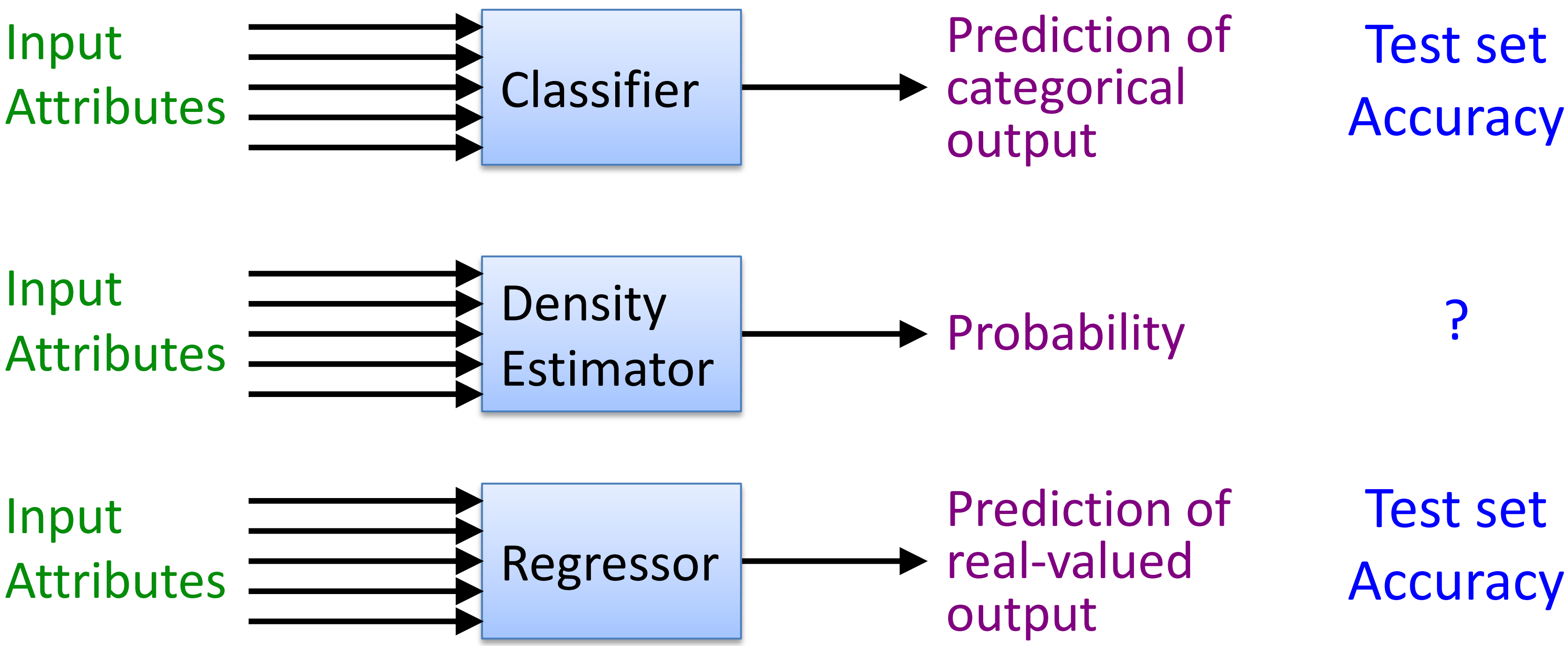
Density Estimation

Compare it against the two other major kinds of models:



Evaluating Density Estimation

Test-set criterion for estimating performance on future data



Evaluating a Density Estimator

- Given a record **x**, a density estimator *M* can tell you how likely the record is:

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

- The density estimator can also tell you how likely the dataset is:
  - Under the assumption that all records were independently generated from the Density Estimator's JD (that is, i.i.d.)

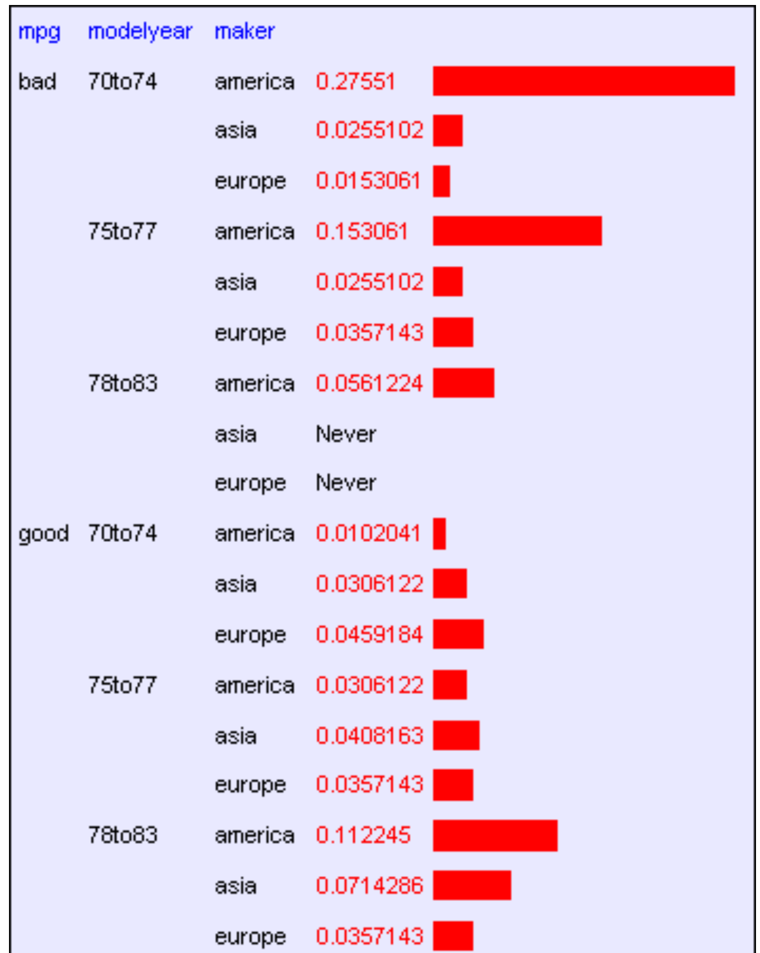
$$\hat{P}(\underbrace{\mathbf{x}_1 \wedge \mathbf{x}_2 \wedge \ldots \wedge \mathbf{x}_n}_{\text{dataset}} \mid M) = \prod_{i=1}^{n} \hat{P}(\mathbf{x}_i \mid M)$$

Example Small Dataset: Miles Per Gallon

From the UCI repository (thanks to Ross Quinlan)

- 192 records in the training set

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



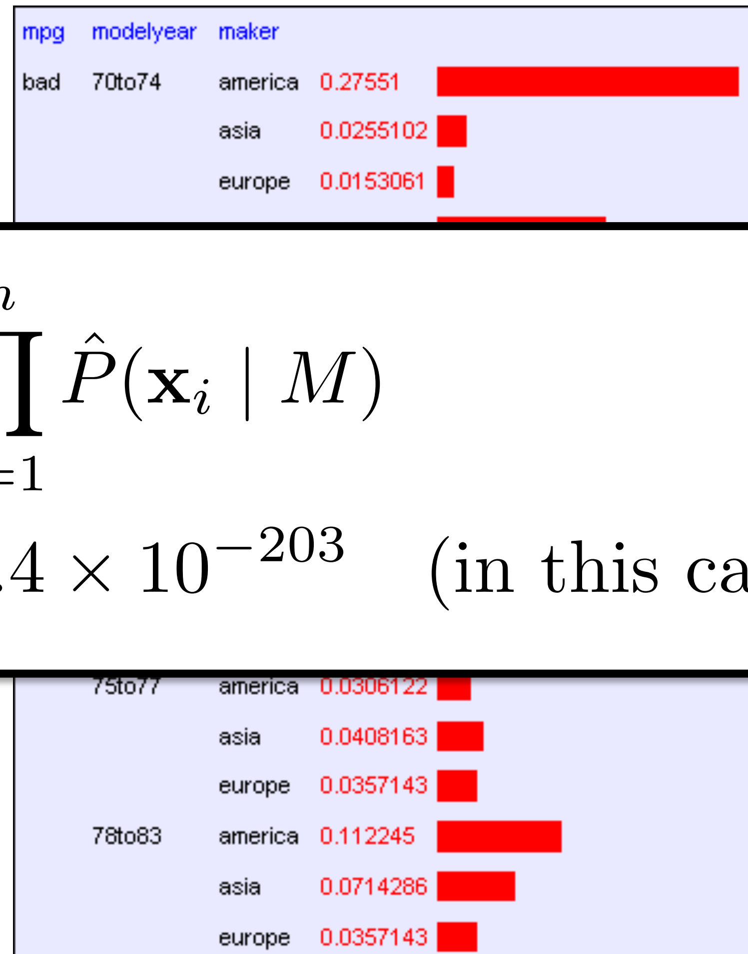
Slide by Andrew Moore

Example Small Dataset: Miles Per Gallon

From the UCI repository (thanks to Ross Quinlan)

- 192 records in the training set

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



$$\hat{P}(\text{dataset} \mid M) = \prod_{i=1}^{n} \hat{P}(\mathbf{x}_i \mid M)$$

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

Log Probabilities

- For decent sized data sets, this product will underflow

$$\hat{P}(\text{dataset} \mid M) = \boxed{\prod_{i=1}^{n}} \hat{P}(\mathbf{x}_i \mid M)$$

- Therefore, since probabilities of datasets get so small, we usually use log probabilities

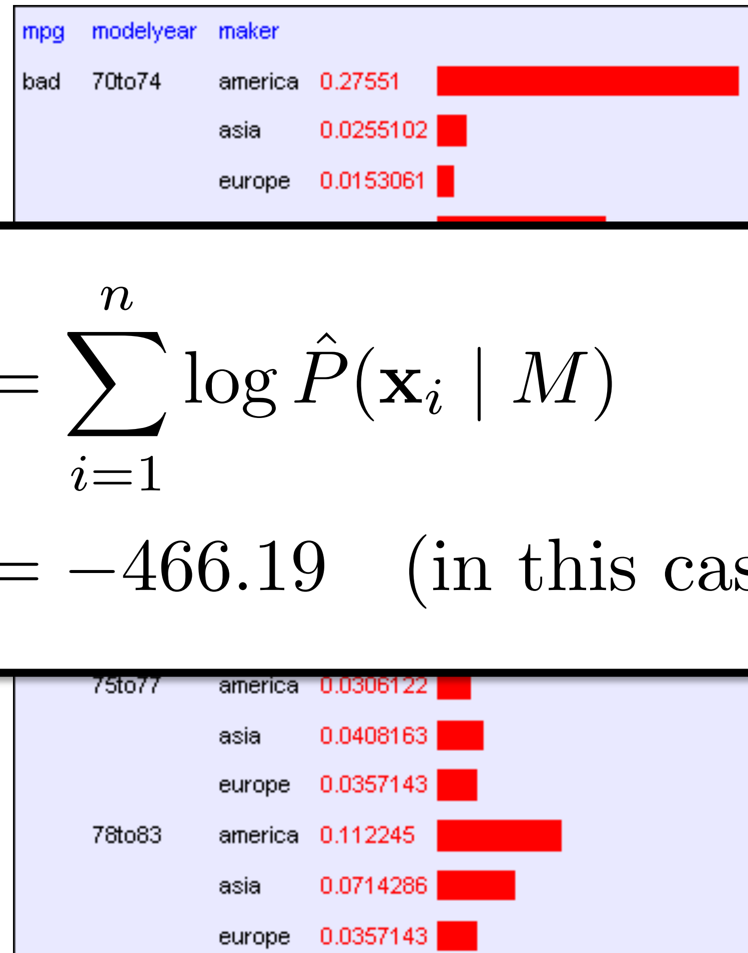
$$\log \hat{P}(\text{dataset} \mid M) = \log \prod_{i=1}^{n} \hat{P}(\mathbf{x}_i \mid M) = \sum_{i=1}^{n} \log \hat{P}(\mathbf{x}_i \mid M)$$

Example Small Dataset: Miles Per Gallon

From the UCI repository (thanks to Ross Quinlan)

- 192 records in the training set

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



$$\log \hat{P}(\text{dataset} \mid M) = \sum_{i=1}^{n} \log \hat{P}(\mathbf{x}_i \mid M)$$

$$= -466.19 \quad \text{(in this case)}$$

Slide by Andrew Moore

Pros/Cons of the Joint Density Estimator

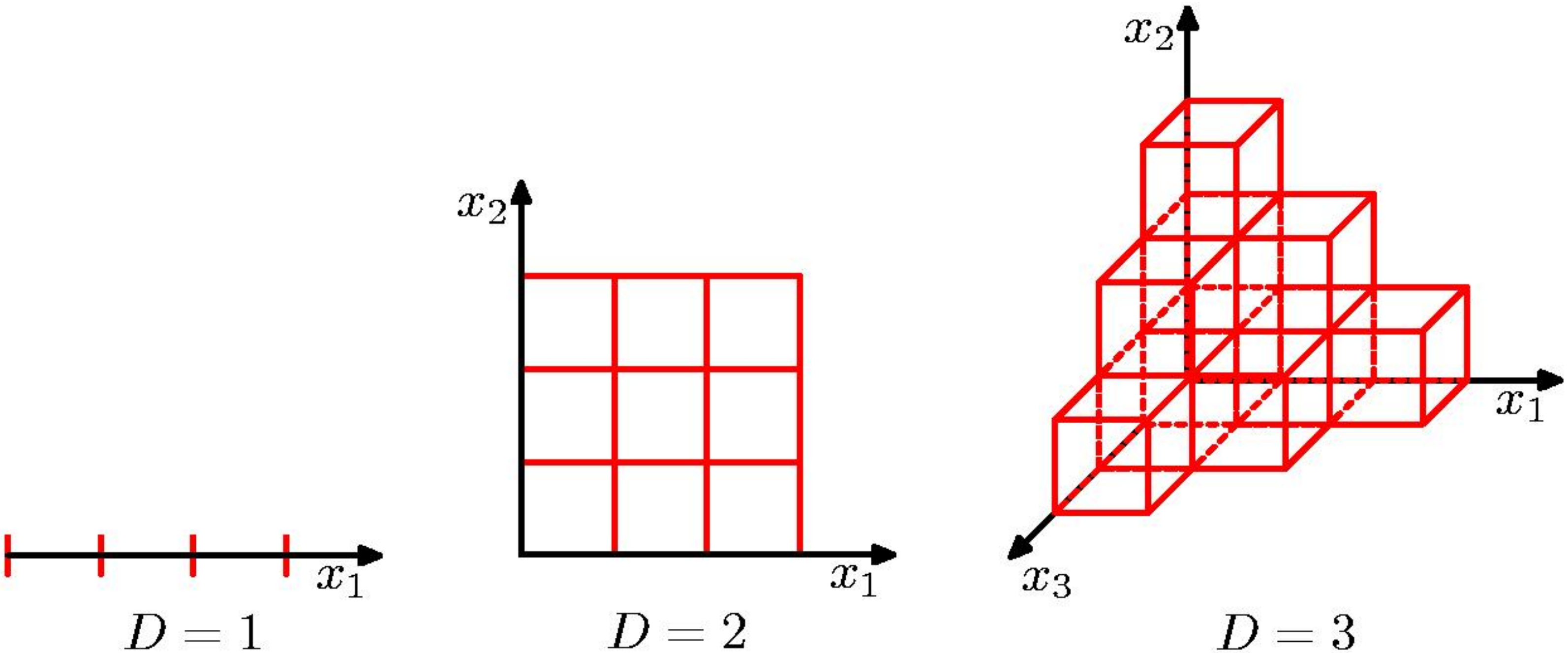
The Good News:

- We can 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
  - Ingredient for Bayes Classifiers (coming very soon...)

The Bad News:

- Density estimation by directly learning the joint is impractical, may result in adverse behavior

Curse of Dimensionality



The Joint Density Estimator on 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 much 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 overfittiest of them all!

Overfitting Density Estimators

If this ever happens, the joint PDE learns there are certain combinations that are impossible

mpg	modelyear	maker	
bad	70to74	america	0.27551
		asia	0.0255102
		europe	0.0153061
	75to77	america	0.153061
		asia	0.0255102
		europe	0.0357143
	78to83	america	0.0561224
		asia	Never
		europe	Never
good	70to74	america	0.0102041
		asia	0.0306122

$$\log \hat{P}(\text{dataset} \mid M) = \sum_{i=1}^{n} \log \hat{P}(\mathbf{x}_i \mid M)$$

$$= -\infty \quad \text{if for any } i,\ \hat{P}(\mathbf{x}_i \mid M) = 0$$

Slide by Andrew Moore