EECS 391
Intro to AI

Probabilistic Reasoning

L11: Thu Oct 5, 2017

Review of concepts from last lecture

Making rational decisions when faced with uncertainty:

- *Probability*

  the precise representation of knowledge and uncertainty

- *Probability theory*

  how to optimally update your knowledge based on new information

- *Decision theory: probability theory + utility theory*

  how to use this information to achieve maximum expected utility

Today: Basic concepts

- random variables
- probability distributions (discrete) and probability densities (continuous)
- rules of probability
- joint and multivariate probability distributions and densities

Axioms of probability

- Axioms (Kolmogorov):

  $0 \leq P(A) \leq 1$

  P(true) = 1

  P(false) = 0

  P(A or B) = P(A) + P(B) − P(A and B)

- Corollaries:
  - A single random variable must sum to 1:

$$\sum_{i=1}^{n} P(D = d_i) = 1$$

  - The joint probability of a set of variables must also sum to 1.
  - If A and B are mutually exclusive:

    P(A or B) = P(A) + P(B)

De Finetti’s definition of probability

- Was there life on Mars?
- You promise to pay $1 if there is, and $0 if there is not.
- Suppose NASA will give us the answer tomorrow.
- Suppose you have an opponent
  - You set the odds (or the “subjective probability”) of the outcome
  - But your opponent decides which side of the bet will be yours
- de Finetti showed that the price you set has to obey the axioms of probability or you face certain loss, i.e. you’ll lose every time.

Rules of probability

- conditional probability

$$Pr(A|B) = \frac{Pr(A \,\mathrm{and}\, B)}{Pr(B)}, \qquad Pr(B) > 0$$

- corollary (Bayes' rule)

$$\begin{aligned} Pr(B|A)Pr(A) &= Pr(A \,\mathrm{and}\, B) = Pr(A|B)Pr(B) \\ \Rightarrow Pr(B|A) &= \frac{Pr(A|B)Pr(B)}{Pr(A)} \end{aligned}$$

Inference with the joint probability distribution

- The complete (probabilistic) relationship between variables is specified by the joint probability:

$$P(X_1, X_2, \ldots, X_n)$$
$$= P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

- All conditional and marginal distributions can be derived from this using the basic rules of probability, the sum rule and the product rule

$$P(X) = \sum_Y P(X,Y)$$ sum rule, "marginalization"

$$P(X,Y) = P(Y|X)P(X) = P(X|Y)P(Y)$$ product rule

$$P(Y|X) = \frac{P(X,Y)}{P(Y)}$$ corollary, conditional probability

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$ corollary, Bayes rule

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

Slides in this style are from Andrew 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

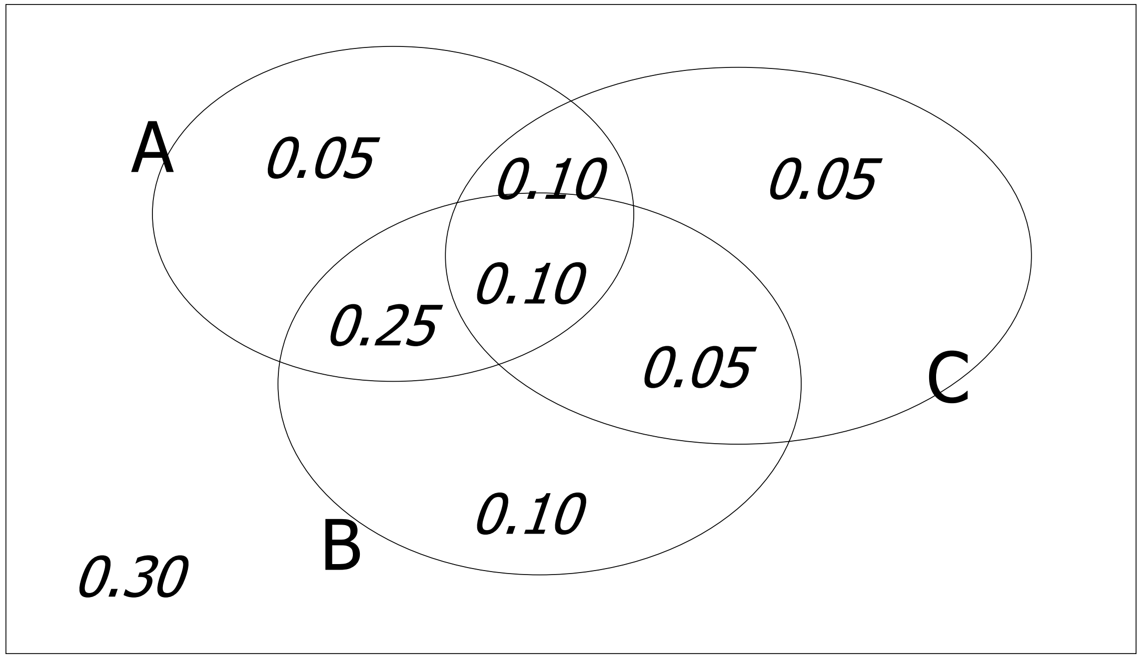
The Joint Distribution

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

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

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

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) = 0.7604

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

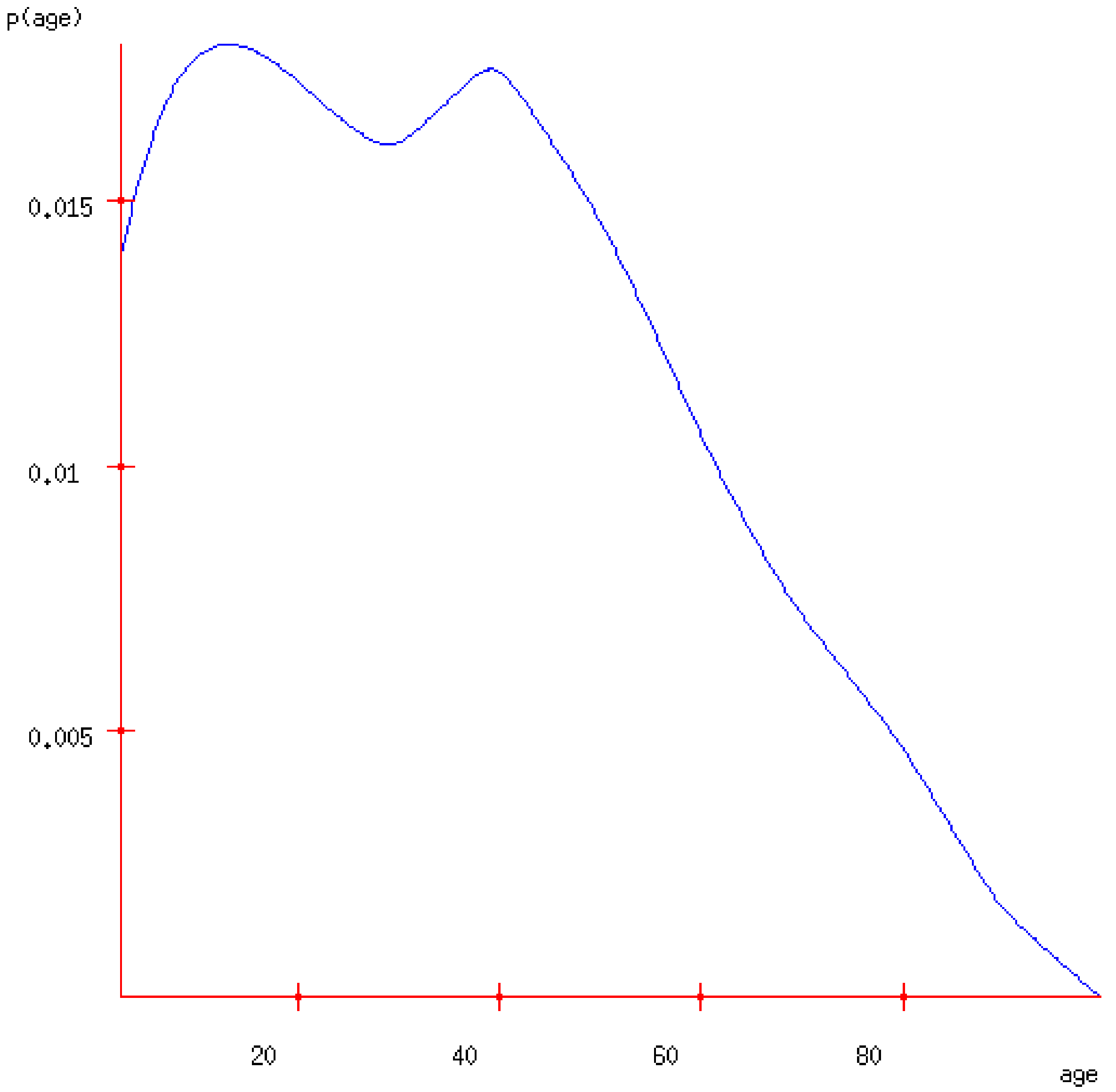
Inference with the Joint

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

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

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

A PDF of American Ages in 2000



more of Andrew's slides

What does p(x) mean?

- It does *not* mean a probability!
- First of all, it's not a value between 0 and 1.
- It's just a value, and an arbitrary one at that.
- The likelihood of p(a) can only be compared *relatively* to other values p(b)
- It indicates the relative probability of the integrated density over a small delta:

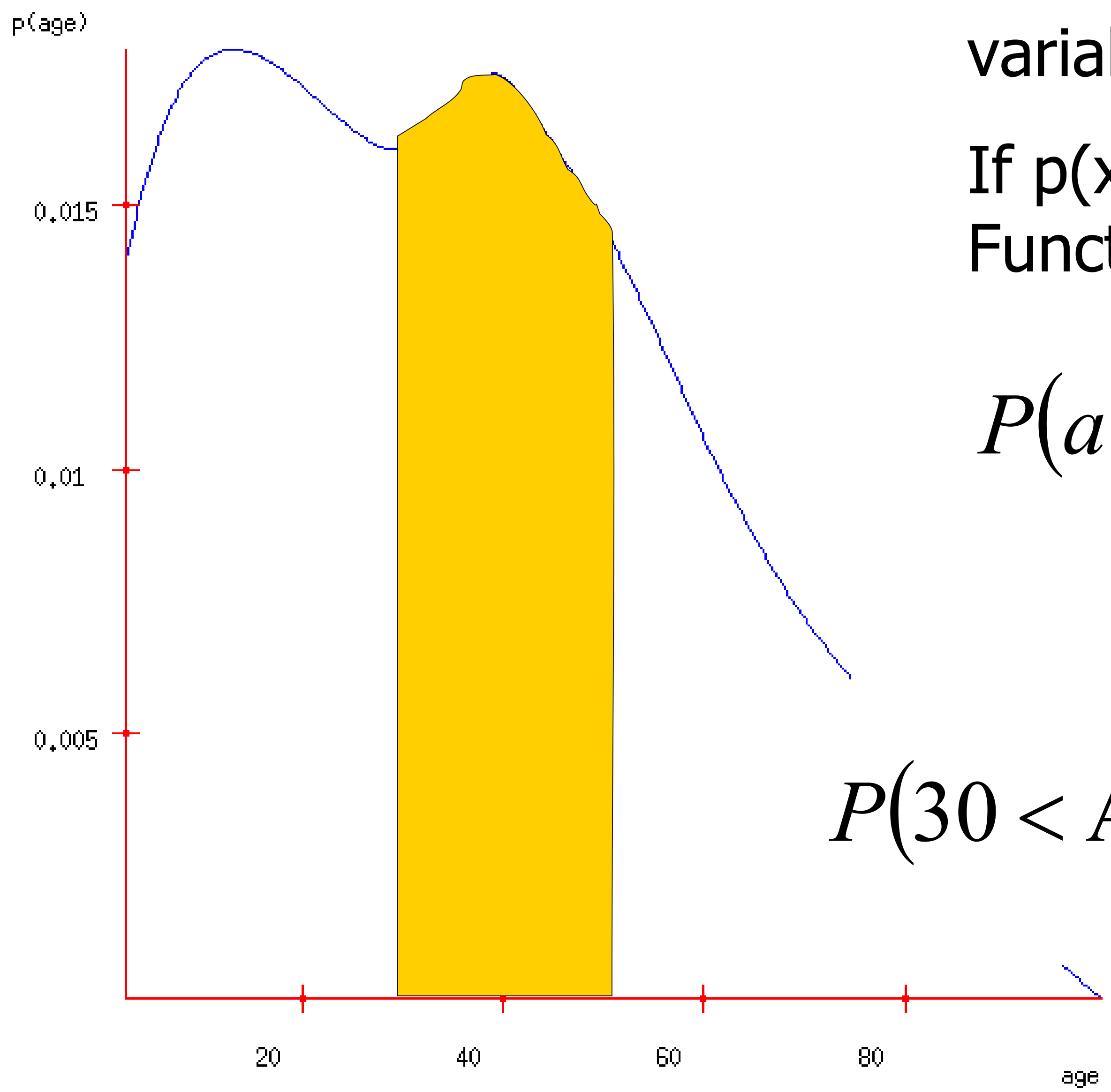
If

$$\frac{p(a)}{p(b)} = \alpha$$

then

$$\lim_{h \to 0} \frac{P(a-h < X < a+h)}{P(b-h < X < b+h)} = \alpha$$

A PDF of American Ages in 2000



Let X be a continuous random variable.

If p(x) is a Probability Density Function for X then…

$$P(a < X \le b) = \int_{x=a}^{b} p(x)\,dx$$

$$P(30 < \text{Age} \le 50) = \int_{\text{age}=30}^{50} p(\text{age})\,d\text{age}$$

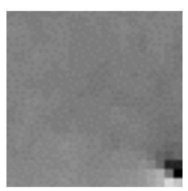
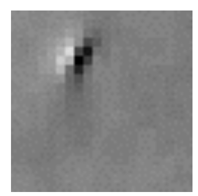
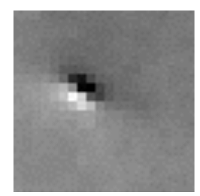
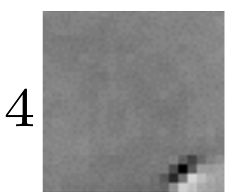
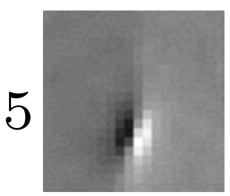
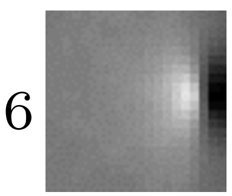
= 0.36

Linear basis decomposition (not part of EECS 391)

- A standard way to represent a pattern is by a linear superposition of vectors:

$$
\begin{aligned}
\mathbf{x} &= \vec{a}_1 s_1 + \vec{a}_2 s_2 + \cdots + \vec{a}_L s_L + \vec{\epsilon} \\
x_i &= \sum_j a_{ij} s_j + \vec{\epsilon} \\
\mathbf{x} &= \mathbf{A}\mathbf{s} + \epsilon
\end{aligned}
$$

- Images can be decomposed (or *encoded*) by a set of scaled features:

 $= s_1$  $+ s_2$  $+ s_3$  $+ s_4$  $+ s_5$  $+ s_6$  $+ \cdots$

- x = image (or pattern) vector
- A = set of features (or basis functions) $\vec{a}_i$
- the residual error is defined by ϵ

An information theoretic approach

Want algorithm to choose optimal $\mathbf{A}$ (basis matrix).

Generative model for data is:

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \boldsymbol{\epsilon}$$

Probability of pattern $\mathbf{x}$ given representation $\mathbf{s}$

$$P(\mathbf{x}|\mathbf{A}, \mathbf{s}) \quad \sim \quad f(\mathbf{x} - \mathbf{A}\mathbf{s}, \boldsymbol{\Sigma}, I)$$

Reasoning from the joint probability distribution

- The table represents the joint probability:
  - $P(J,M,D)$
- How do we fill in the table?
  - Only constraint is that the probabilities sum to 1.
- How do we reason from the joint probability? Examples:
  - $P(D) = ?$
  - $P(D|J,M) = ?$
  - $P(D|M) = ?$

<2 years at job? **“J”**	**missed payments?** **“M”**	**defaulted on loan?** **“D”**	**P(J,M,D)**
N	N	N	0.5
N	N	Y	0
N	Y	N	0.05
N	Y	Y	0.01
Y	N	N	0.3
Y	N	Y	0
Y	Y	N	0.1
Y	Y	Y	0.04

Reasoning from the joint probability distribution

- How do we compute P(D)?
- What does this say?
  - "The probability of defaulting"
- This is obtained by *marginalizing* over the joint probability:

$$P(D) = \sum_{J,M} P(J, M, D)$$

- What do we sum over?

<2 years at job? "J"	**missed payments?** "M"	**defaulted on loan?** "D"	**P(J,M,D)**
N	N	N	0.5
N	N	Y	0
N	Y	N	0.05
N	Y	Y	0.01
Y	N	N	0.3
Y	N	Y	0
Y	Y	N	0.1
Y	Y	Y	0.04

Reasoning from the joint probability distribution

- How do we compute P(D)?
- What does this say?
  - "The probability of defaulting"
- This is obtained by *marginalizing* over the joint probability:

$$P(D) = \sum_{J,M} P(J, M, D)$$

- What do we sum over?

$$P(D = \mathrm{Y}) = 0.00 + 0.01 + 0.00 + 0.04$$
$$= 0.05$$

<2 years at job? "J"	missed payments? "M"	defaulted on loan? "D"	P(J,M,D)
N	N	N	0.5
N	N	Y	0
N	Y	N	0.05
N	Y	Y	0.01
Y	N	N	0.3
Y	N	Y	0
Y	Y	N	0.1
Y	Y	Y	0.04

Reasoning from the joint probability distribution

- How do we compute P(D)?
- What does this say?
  - "The probability of defaulting"
- This is obtained by *marginalizing* over the joint probability:

$$P(D) = \sum_{J,M} P(J, M, D)$$

- What do we sum over?

$$\begin{aligned} P(D = \mathrm{Y}) &= 0.00 + 0.01 + 0.00 + 0.04 \\ &= 0.05 \end{aligned}$$

$$\begin{aligned} P(D = \mathrm{N}) &= 0.50 + 0.05 + 0.30 + 0.10 \\ &= 0.95 \end{aligned}$$

<2 years at job? "J"	missed payments? "M"	defaulted on loan? "D"	P(J,M,D)
N	N	N	0.5
N	N	Y	0
N	Y	N	0.05
N	Y	Y	0.01
Y	N	N	0.3
Y	N	Y	0
Y	Y	N	0.1
Y	Y	Y	0.04

Reasoning from the joint probability distribution

- How do we compute P(D)?
- What does this say?
  - "The probability of defaulting"
- This is obtained by *marginalizing* over the joint probability:

$$P(D) = \sum_{J,M} P(J, M, D)$$

- What do we sum over?

$$P(D = \mathrm{Y}) = 0.00 + 0.01 + 0.00 + 0.04$$
$$= 0.05$$

$$P(D = \mathrm{N}) = 0.50 + 0.05 + 0.30 + 0.10$$
$$= 0.95$$

P(J,M,D)

<2 years at job? "J"	missed payments? "M"	defaulted on loan? "D"	Probability
N	N	N	0.50
N	N	Y	0.00
N	Y	N	0.05
N	Y	Y	0.01
Y	N	N	0.30
Y	N	Y	0.00
Y	Y	N	0.10
Y	Y	Y	0.04

D	P(D)
N	0.95
Y	0.05

Reasoning from the joint probability distribution

- How do we compute P(D | J,M)?
- What does this say?
  - "The probability of defaulting given <2 years at job & missed payments"
- How do we calculate this from the joint probability P(J, M, D)?

$$P(D|J,M) = \frac{P(J,M,D)}{P(J,M)}$$

- What is P(J,M)?

$$P(J,M) = \sum_D P(J,M,D)$$

<2 years at job? "J"	**missed payments?** "M"	**defaulted on loan?** "D"	**P(J,M,D)**
N	N	N	0.5
N	N	Y	0
N	Y	N	0.05
N	Y	Y	0.01
Y	N	N	0.3
Y	N	Y	0
Y	Y	N	0.1
Y	Y	Y	0.04

Reasoning from the joint probability distribution

- How do we compute P(D | J,M)?
- What does this say?
  - "The probability of defaulting given <2 years at job & missed payments"
- How do we calculate this from the joint probability P(J, M, D)?

$$P(D|J,M) = \frac{P(J,M,D)}{P(J,M)}$$

- What is P(J,M)?

$$P(J,M) = \sum_D P(J,M,D)$$

$$\begin{aligned} P(\bar{J},\bar{M}) &= P(\bar{J},\bar{M},\bar{D}) \\ &+ P(\bar{J},\bar{M},D) \\ &= 0.50 + 0.00 \end{aligned}$$

<2 years at job? "J"	**missed payments?** "M"	**defaulted on loan?** "D"	**P(J,M,D)**
N	N	N	0.5
N	N	Y	0
N	Y	N	0.05
N	Y	Y	0.01
Y	N	N	0.3
Y	N	Y	0
Y	Y	N	0.1
Y	Y	Y	0.04

Reasoning from the joint probability distribution

- How do we compute P(D | J,M)?
- What does this say?
  - "The probability of defaulting given <2 years at job & missed payments"
- How do we calculate this from the joint probability P(J, M, D)?

$$P(D|J,M) = \frac{P(J,M,D)}{P(J,M)}$$

- What is P(J,M)?

$$P(J,M) = \sum_D P(J,M,D)$$

$$\begin{aligned} P(\bar{J},M) &= P(\bar{J},M,\bar{D}) \\ &\quad + P(\bar{J},M,D) \\ &= 0.05 + 0.01 \end{aligned}$$

<2 years at job? "J"	**missed payments?** "M"	**defaulted on loan?** "D"	**P(J,M,D)**
N	N	N	0.5
N	N	Y	0
N	Y	N	0.05
N	Y	Y	0.01
Y	N	N	0.3
Y	N	Y	0
Y	Y	N	0.1
Y	Y	Y	0.04

Reasoning from the joint probability distribution

- How do we compute P(D | J,M)?
- What does this say?
  - "The probability of defaulting given <2 years at job & missed payments"
- How do we calculate this from the joint probability P(J, M, D)?

$$P(D|J,M) = \frac{P(J,M,D)}{P(J,M)}$$

- What is P(J,M)?

$$P(J,M) = \sum_D P(J,M,D)$$

$$\begin{aligned} P(J,\bar{M}) &= P(J,\bar{M},\bar{D}) \\ &+ P(J,\bar{M},D) \\ &= 0.30 + 0.00 \end{aligned}$$

<2 years at job? "J"	**missed payments?** "M"	**defaulted on loan?** "D"	**P(J,M,D)**
N	N	N	0.5
N	N	Y	0
N	Y	N	0.05
N	Y	Y	0.01
Y	N	N	0.3
Y	N	Y	0
Y	Y	N	0.1
Y	Y	Y	0.04

Reasoning from the joint probability distribution

- How do we compute P(D | J,M)?
- What does this say?
  - "The probability of defaulting given <2 years at job & missed payments"
- How do we calculate this from the joint probability P(J, M, D)?

$$P(D|J,M) = \frac{P(J,M,D)}{P(J,M)}$$

- What is P(J,M)?

$$P(J,M) = \sum_D P(J,M,D)$$

$$\begin{aligned} P(J,M) &= P(J,M,\bar{D}) \\ &+ P(J,M,D) \\ &= 0.10 + 0.04 \end{aligned}$$

<2 years at job? "J"	**missed payments?** "M"	**defaulted on loan?** "D"	**P(J,M,D)**
N	N	N	0.5
N	N	Y	0
N	Y	N	0.05
N	Y	Y	0.01
Y	N	N	0.3
Y	N	Y	0
Y	Y	N	0.1
Y	Y	Y	0.04

Reasoning from the joint probability distribution

- Now we know $P(J,M)$, but how do we calculate $P(D \mid J,M)$?

$$P(D|J,M) = \frac{P(J,M,D)}{P(J,M)}$$

J	**M**	**P(J,M)**
N	N	0.5
N	Y	0.06
Y	N	0.3
Y	Y	0.14

Reasoning from the joint probability distribution

$$P(J, M)$$

J	M	P(J,M)
N	N	0.5
N	Y	0.06
Y	N	0.3
Y	Y	0.14

$$P(D|J, M) = \frac{P(J, M, D)}{P(J, M)}$$

J	M	D	P(J,M,D)		P(D | J,M)
N	N	N	0.5	0.50/0.50	1
N	N	Y	0	0.00/0.50	0
N	Y	N	0.05	0.05/0.06	0.83
N	Y	Y	0.01	0.01/0.06	0.17
Y	N	N	0.3	0.30/0.30	1
Y	N	Y	0	0.00/0.30	0
Y	Y	N	0.1	0.10/0.14	0.71
Y	Y	Y	0.04	0.04/0.14	0.29

Reasoning from the joint probability distribution

- How do we compute $P(D \mid M)$?

$$P(D|M) = \frac{P(D,M)}{P(M)}$$

- How do we compute $P(D,M)$?

$$P(D,M) = \sum_{J} P(J,M,D)$$

Summary of inference with the joint probability distribution

- The complete (probabilistic) relationship between variables is specified by the joint probability:

$$P(X_1, X_2, \ldots, X_n)$$
$$= P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

- All conditional and marginal distributions can be derived from this using the basic rules of probability, the sum rule and the product rule

$$P(X) = \sum_Y P(X, Y)$$ sum rule, "marginalization"

$$P(X, Y) = P(Y|X)P(X) = P(X|Y)P(Y)$$ product rule

$$P(Y|X) = \frac{P(X, Y)}{P(Y)}$$ corollary, conditional probability

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$ corollary, Bayes rule

Simple example: medical test results

- Test report for rare disease is positive, 90% accurate
- What's the probability that you have the disease?
- What if the test is repeated?

- This is the simplest example of reasoning by combining sources of information.

How do we model the problem?

- Which is the correct description of "Test is 90% accurate" ?

$$\begin{aligned} P(T = \text{true}) &= 0.9 \\ P(T = \text{true}|D = \text{true}) &= 0.9 \\ P(D = \text{true}|T = \text{true}) &= 0.9 \end{aligned}$$

- What do we want to know?

$$P(T = \text{true})$$
$$P(T = \text{true}|D = \text{true})$$
$$P(D = \text{true}|T = \text{true})$$

- More compact notation:

$$\begin{aligned} P(T = \text{true}|D = \text{true}) &\rightarrow P(T|D) \\ P(T = \text{false}|D = \text{false}) &\rightarrow P(\bar{T}|\bar{D}) \end{aligned}$$

Evaluating the posterior probability through Bayesian inference

- We want P(D|T) = "The probability of the having the disease given a positive test"
- Use Bayes rule to relate it to what we know: P(T|D)

$$\text{posterior} \quad P(D|T) = \frac{\overbrace{P(T|D)}^{\text{likelihood}}\overbrace{P(D)}^{\text{prior}}}{\underbrace{P(T)}_{\text{normalizing constant}}}$$

- What's the prior P(D)?
- Disease is rare, so let's assume

$$P(D) = 0.001$$

- What about P(T)?
- What's the interpretation of that?

Evaluating the normalizing constant

likelihood *prior*

posterior $$P(D|T) = \frac{P(T|D)P(D)}{P(T)}$$

normalizing constant

- P(T) is the marginal probability of P(T,D) = P(T|D) P(D)
- So, compute with summation

$$P(T) = \sum_{\text{all values of D}} P(T|D)P(D)$$

- For true or false propositions:

$$P(T) = P(T|D)P(D) + P(T|\bar{D})P(\bar{D})$$

What are these?

Refining our model of the test

- We also have to consider the negative case to incorporate all information:

$$
\begin{aligned}
P(T|D) &= 0.9 \\
P(T|\bar{D}) &= ?
\end{aligned}
$$

- What should it be?
  - It can be any value between 0 and 1. It does not depend on P(T|D).

- What about this:

$$P(T|D) = 0.9 \quad \Rightarrow \quad P(\bar{T}|D) = 0.1$$

- This is because P(T|D) must be a valid probability distribution

$$\sum_{T=\text{True}, T=\text{False}} P(T|D) = 1$$

False positives and false negatives

- What is the expression for the probability of false positives?
  - The probability that the test is true given the disease is false:

$$P(T = \text{True}|D = \text{False})$$

- What is the expression for the probability of false negatives?
  - The probability that the test is false given the disease is true.

$$P(T = \text{False}|D = \text{True})$$

- What would you call $P(T = \text{True}|D = \text{True})$ and $P(T = \text{False}|D = \text{False})$?
- The probability of a true positive and a true negative.
- This is closer to what is meant when we say the test is "90% accurate".

Plugging in the numbers

$$P(D|T) \quad = \quad \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})}$$

$$P(D|T) = \frac{0.9 \times 0.001}{0.9 \times 0.001 + 0.1 \times 0.999} = 0.0089$$

Note: here we assume the false positive rate and the false negative rate are the same.