Review: probability

- Monty Hall, weighted dice
- Frequentist v. Bayesian
- Independence
- Expectations, conditional expectations
  - Exp. & independence; linearity of exp.
- Estimator (RV computed from sample)
  - law of large #s, bias, variance, tradeoff

Covariance

- Suppose we want an approximate numeric measure of (in)dependence
- Let E(X) = E(Y) = 0 for simplicity
- Consider the random variable XY
  - if X, Y are typically both +ve or both -ve

    Both +, $xy > 0$ Both − $xy > 0$

    $\Rightarrow E(XY) > 0$
  - if X, Y are independent

    $E(XY) = 0$

Covariance

$E(X - E(X)) = E(X) - E(X) = 0$

- cov(X,Y) = $E((X - E(X))(Y - E(Y)))$
- Is this a good measure of dependence?
  - Suppose we scale X by 10:

$$\begin{aligned} cov(10X, Y) &= E((10X - E(10X))(Y - E(Y))) \\ &= 10E((X - E(X))(Y - E(Y))) \\ &= 10\, cov(X, Y) \end{aligned}$$

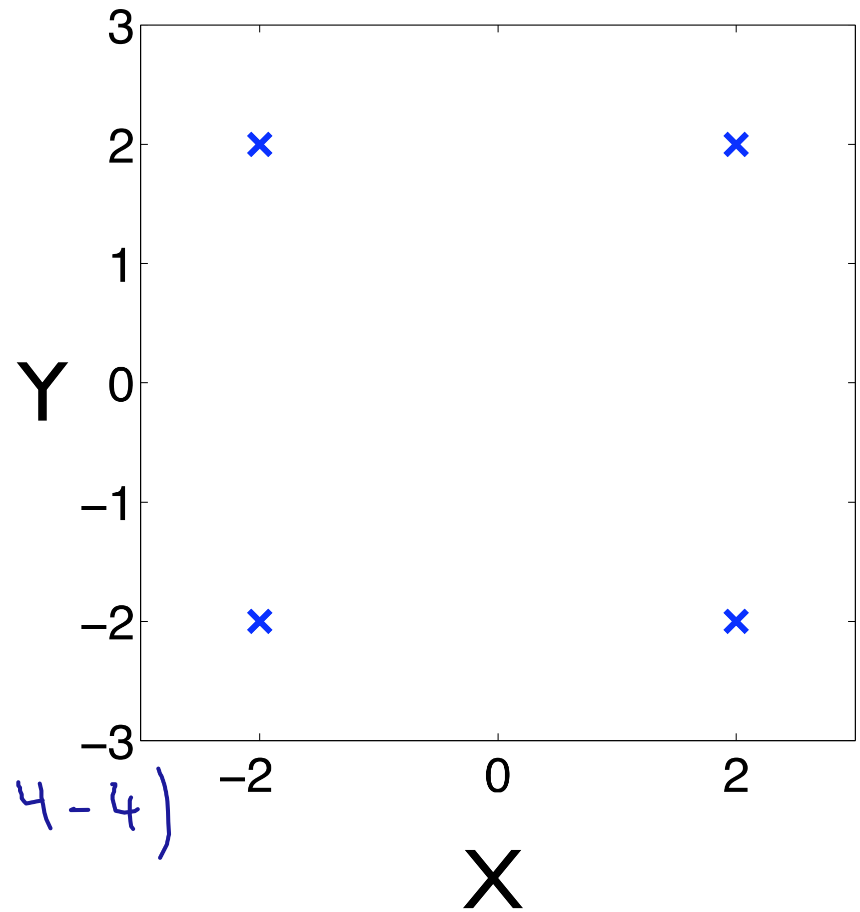
Correlation

- Like covariance, but controls for variance of individual r.v.s
- cor(X,Y) = $\mathrm{cov}(X,Y) / \sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}$
- cor(10X,Y) = $\underbrace{\mathrm{cov}(10X,Y)}_{10\,\mathrm{cov}(X,Y)} / \sqrt{\underbrace{\mathrm{var}(10X)}_{100\,\mathrm{var}(X)}\,\mathrm{var}(Y)}$

$= \mathrm{cor}(X,Y)$

Correlation & independence

- Equal probability on each point
- Are X and Y independent?
- Are X and Y uncorrelated?



$$\mathrm{cov}(X,Y) = E(XY) = \frac{1}{4}(-4 + 4 + 4 - 4)$$
$$= 0$$

$$\Rightarrow \mathrm{cor}(X,Y) = 0$$

Correlation & independence

- Do you think that all independent pairs of RVs are uncorrelated?

lots Y nobody N

- Do you think that all uncorrelated pairs of RVs are independent?

lots N $\frac{1}{2}$ Y

Proofs and (counter)examples

- For a question A $\stackrel{?}{\Rightarrow}$ B
  - e.g., X,Y uncorrelated $\stackrel{?}{\Rightarrow}$ X,Y independent
  - if true, usually need to provide a ***proof***
  - if false, usually only need to provide a ***counterexample***

Counterexamples

$$A \overset{?}{\Rightarrow} B$$

$$X, Y \text{ uncorrelated} \overset{?}{\Rightarrow} X, Y \text{ independent}$$

- Counterexample = example satisfying A but not B
- E.g., RVs X and Y that are ***not*** independent, but ***are*** un correlated

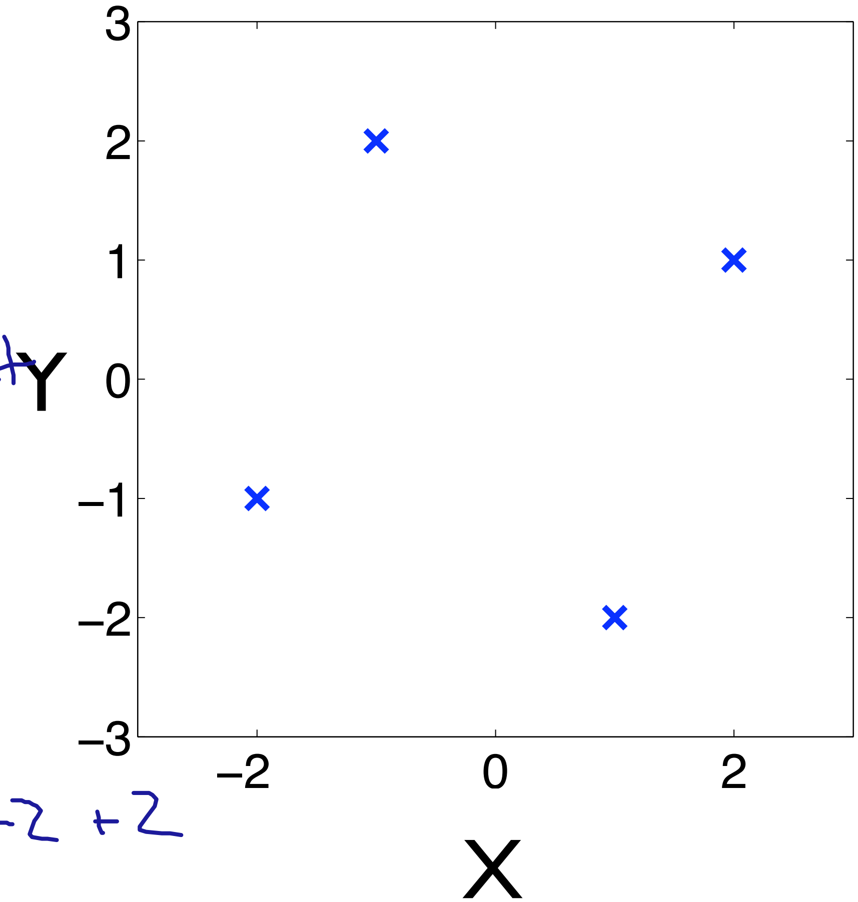
$\neg B$

A

Correlation & independence

Counterexample

- Equal probability on each point
- Are X and Y independent? dependent
- Are X and Y uncorrelated?



$\mathrm{cov}(X,Y) = E(XY) = \frac{1}{4}(-2 + 2 - 2 + 2$

$= 0$

$\Rightarrow \mathrm{cor}(X,Y) = 0$

Law of iterated expectations

- For any two RVs, X and Y, we have:

$$E_Y(E_X(X|Y)) = E_{XY}(X)$$

- Convention: note in subscript the RVs that are not yet conditioned on (in this E(.)) or marginalized away (inside this E(.))

Law of iterated expectations

- $E_X(X \mid Y) =$ $\sum_x P(x \mid Y)\, x$
- $E_Y(\underbrace{E_X(X \mid Y)}) =$ $\sum_y P(y) \sum_x P(x \mid y)\, x$

$$= \sum_x \sum_y \underbrace{P(y) P(x \mid y)}_{P(x,y)}\, x$$

$$= \sum_x P(x)\, x = E(X)$$

Bayes Rule

Rev. Thomas Bayes
1702–1761

- For any X, Y, C — context
  - P(X | Y, C) P(Y | C) = P(Y | X, C) P(X | C)
- Simple version (without context)
  - P(X | Y) P(Y) = P(Y | X) P(X)

$P(x|y) = P(y|x)\,P(x) / P(y)$

- Can be taken as definition of conditioning

Exercise

- You are tested for a rare disease, emacsitis—prevalence 3 in 100,000
- Your receive a test that is 99% ***sensitive*** and 99% ***specific***
  - sensitivity = P(yes | emacsitis)
  - specificity = P(no | ~emacsitis)
- The test comes out ***positive***
- Do you have emacsitis?

E = emacsitis

Y = test true

$$P(E \mid Y) = \frac{3}{1000}$$

$$= P(Y \mid E)\,P(E) / P(Y)$$

$$= .99 \frac{3}{100000} / P(Y)$$

$$P(Y) = P(Y, E) + P(Y, \bar{E})$$

$$P(Y, E) = P(Y \mid E)\,P(E) \approx \frac{3}{100000}$$

$$P(Y, \bar{E}) = \underbrace{P(Y \mid \bar{E})}_{.01}\,\underbrace{P(\bar{E})}_{\frac{99,997}{100000}}$$

$$\approx .01$$

$$P(Y) \approx .01$$

Revisit: weighted dice

$P(wt) = 1/10$

- Fair dice: all 36 rolls equally likely $\nearrow 1/36$
- Weighted: rolls summing to 7 more likely $\nearrow$ $7 : 1/12$ $\quad \bar{7} : 1/60$
- Data: 1-6 2-5

$$P(wt \mid rolls) = P(rolls \mid wt)\,P(wt) / P(rolls)$$

$$= \frac{P(roll_1 \mid wt)\,P(roll_2 \mid wt)\,\frac{1}{10}}{P(rolls)}$$

$$= \frac{1}{12}\,\frac{1}{12}\,\frac{1}{10} \Big/ P(rolls)$$

$$= \frac{1}{12}\,\frac{1}{12}\,\frac{1}{10} \Big/ \left(\frac{1}{36}\,\frac{1}{36}\,\frac{7}{10} + \frac{1}{12}\,\frac{1}{12}\,\frac{1}{10}\right)$$

$$\approx 0.5$$

$$P(\overline{wt} \mid rolls) = P(rolls \mid \overline{wt})\,P(\overline{wt}) / P(rolls)$$

$$= \frac{1}{36}\,\frac{1}{36}\,\frac{9}{10} \Big/ P(rolls)$$

$$1 = \frac{\frac{1}{36}\,\frac{1}{36}\,\frac{9}{10} + \frac{1}{12}\,\frac{1}{12}\,\frac{1}{10}}{P(rolls)}$$

Learning from data

- Given a ***model class***
- And some data, sampled from a model in this class
- Decide which model best explains the sample

$P(H) = \{0, .01, .02 \ldots 1\}$

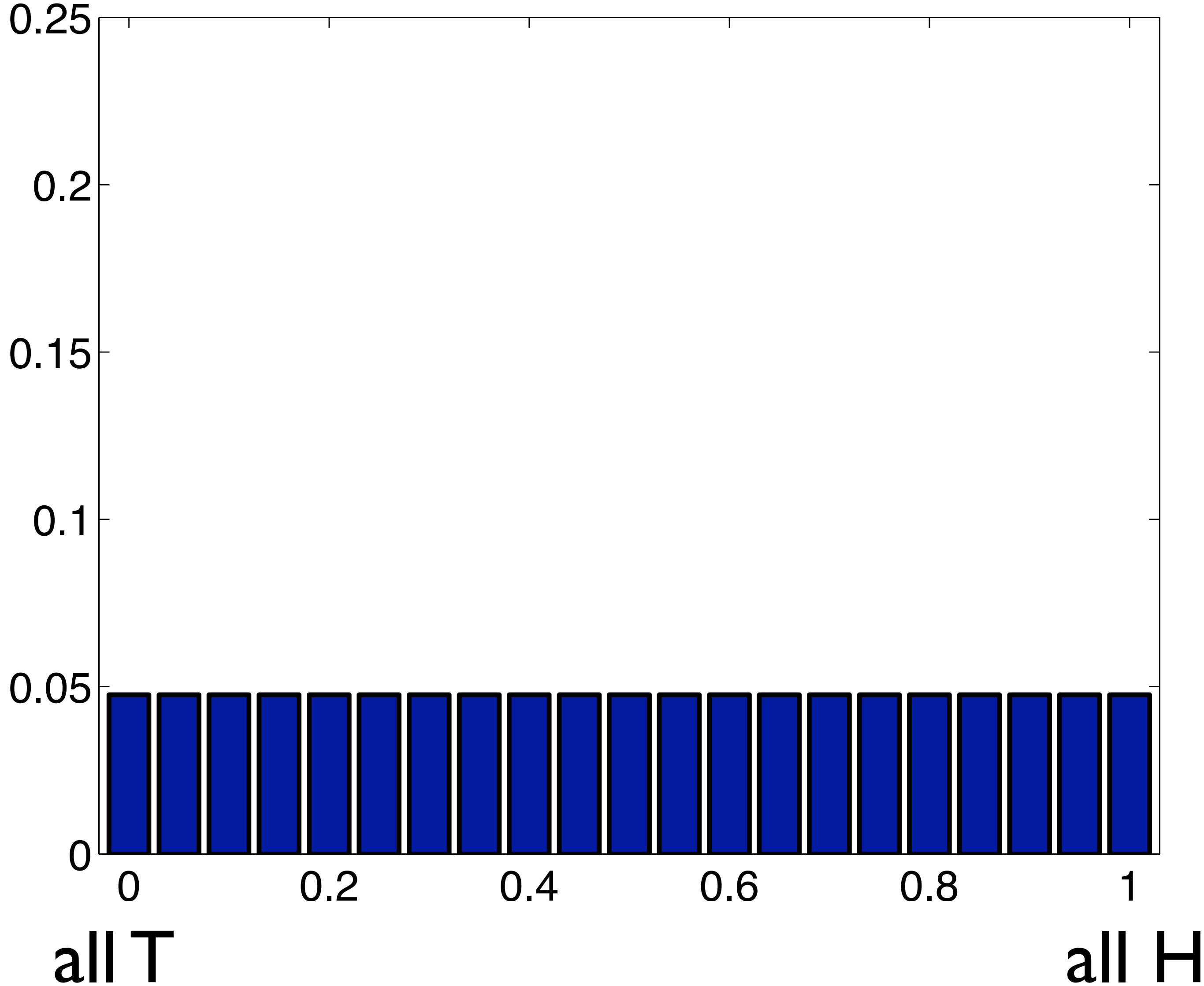
face recogn params

$= \{(2, .7, -.1), (2, .7, +.1), \ldots\}$

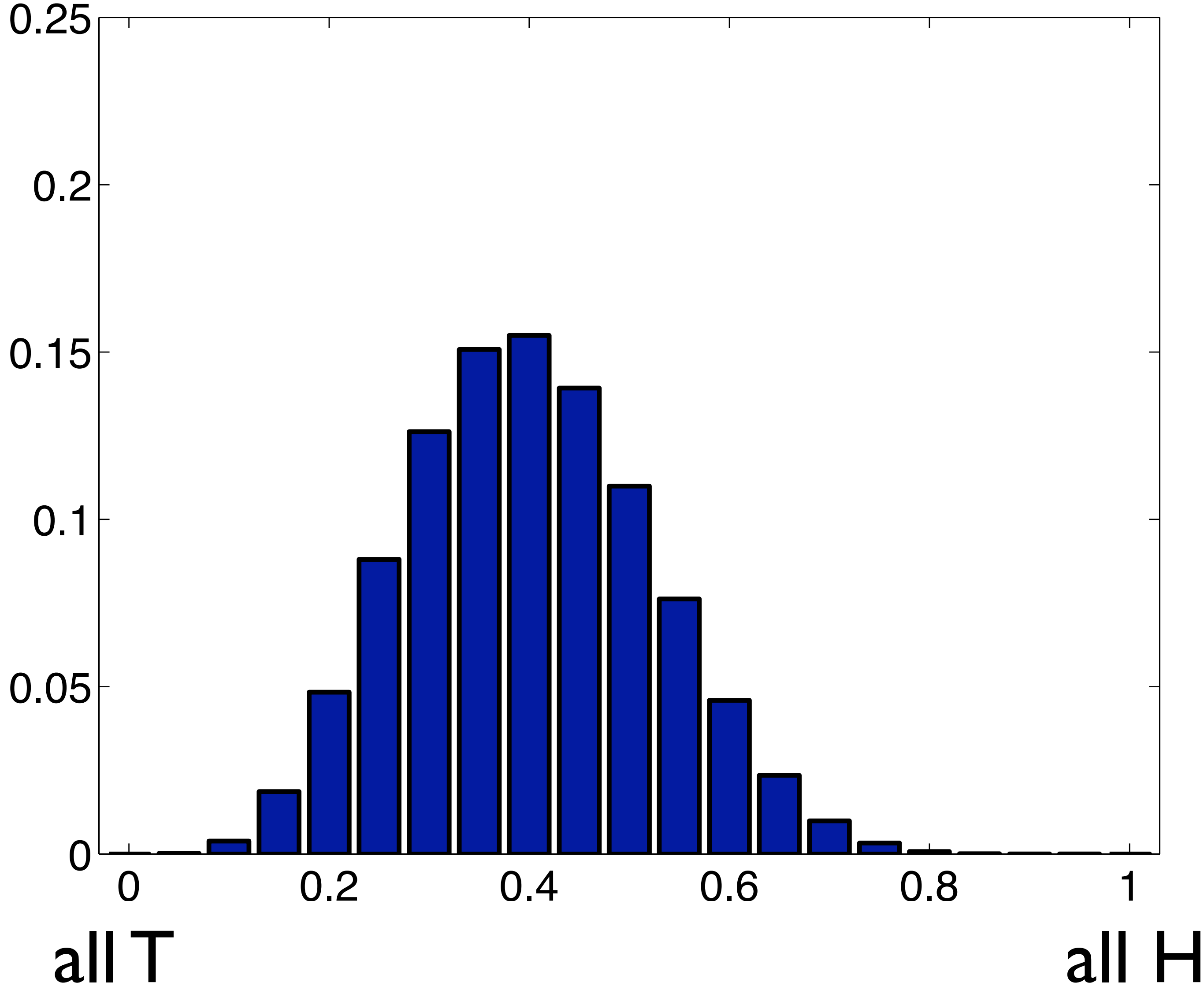
Bayesian model learning

- P(model | data) = P(data | model) P(model) / P(data)
  P(data | model) P(model) / Z
- Z = P(data)
- So, for each model, compute: P(data | model) P(model)
- Then: normalize → sum over models & divide
  ↳ hard!! if lots of models

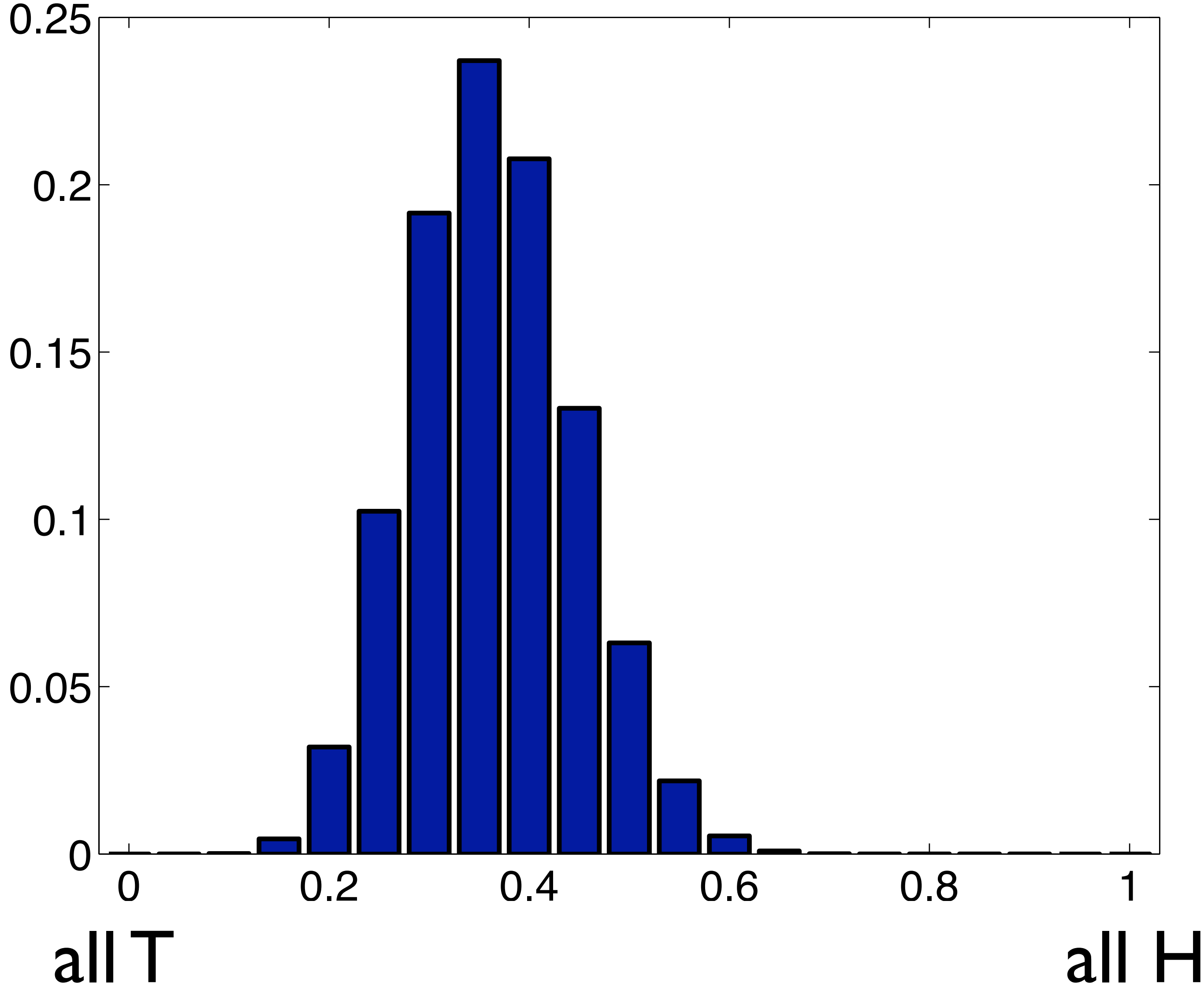
Prior: uniform



Posterior: after 5H, 8T



Posterior: 11H, 20T



Graphical models

Why do we need graphical models?

- So far, only way we've seen to write down a distribution is as a big table
- Gets unwieldy fast!
  - E.g., 10 RVs, each w/ 10 settings
  - Table size = 10^{10}
- Graphical model: way to write distribution compactly using diagrams & numbers

Example ML problem

- US gov't inspects food packing plants
  - 27 tests of contamination of surfaces
  - 12-point ISO 9000 compliance checklist
  - are there food-borne illness incidents in 30 days after inspection? (15 types)
- Q: after an inspection, do we have to worry about illness
- A: write down dist'n of all 54 RV
  P(E. coli | data from inspection)

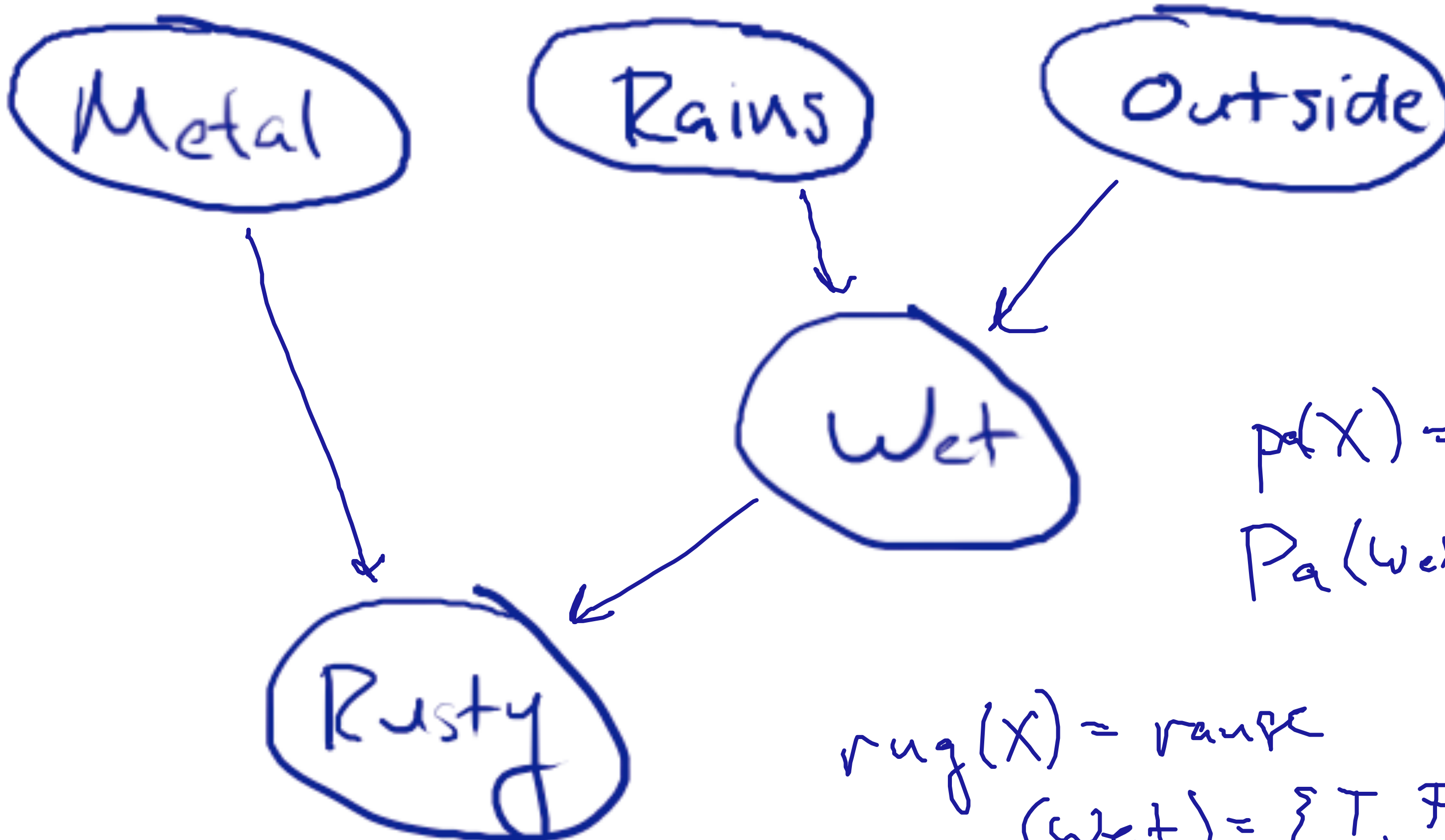
Big graphical models

- Later in course, we'll use graphical models to express various ML algorithms
  - e.g., the one from the last slide
- These graphical models will be big!
- Please bear with some smaller examples for now so we can fit them on the slides and do the math in our heads…

Bayes nets

- Best-known type of graphical model
- Two parts: DAG and CPTs

Rusty robot: the DAG



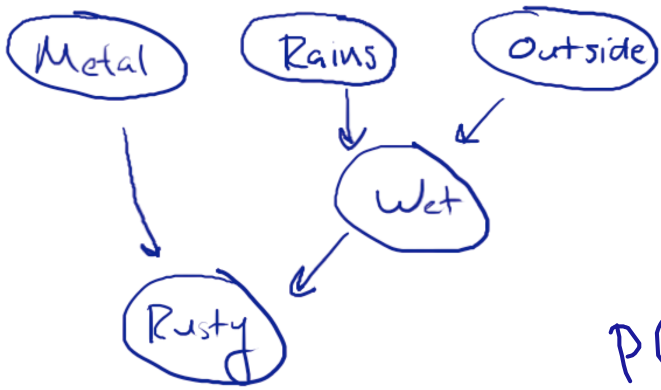
pa(X) = parent set

Pa(Wet) = {Rains, Outside}

rng(X) = range

rng(Wet) = {T, F}

Rusty robot: the CPTs



$P(M) = 0.9$

$P(Ra) = 0.7$

$P(O) = 0.2$

$P(W \mid Ra, O) =$

TT	0.9
TF	0.1
FT	0.1
FF	0.1

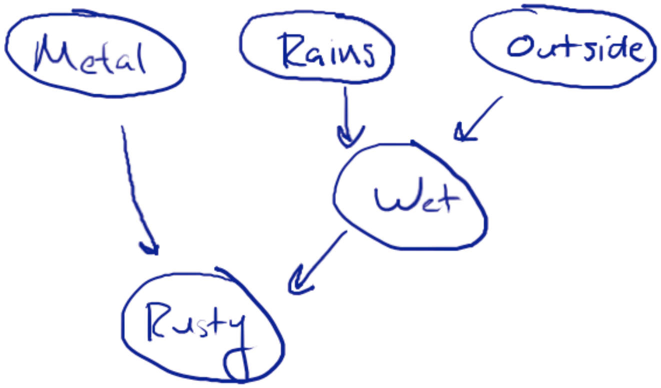
$P(Ru \mid M, W) =$

TT	0.8
TF	0.1
FT	0
FF	0

11 #s

- For each RV (say X), there is one CPT specifying P(X | pa(X))

Interpreting it



$$P(RV_s) = \prod_{X \in RV_s} P(X \mid Pa(X))$$

$$P(M, Ra, O, W, Ru) = P(M)\,P(Ra)\,P(O)\,P(W \mid Ra, O)\,P(Ru \mid W, M)$$

M	Ra	O	W	Ru	P
F	F	F	F	F	0.1 0.3 0.8 0.9 1 = 0.0216
F	F	F	F	T	" " " " 0 = 0

⋮

31 #s

Benefits

- 11 v. 31 numbers
- Fewer parameters to learn
- Efficient ***inference*** = computation of marginals, conditionals $\Rightarrow$ posteriors

Inference example

- P(M, Ra, O, W, Ru) =

  P(M) P(Ra) P(O) P(W|Ra,O) P(Ru|M,W)

- Find marginal of M, O

$$\sum_{Ra \in \{T,F\}} \sum_{W \in \{T,F\}} \sum_{Ru \in \{T,F\}} P(M) P(Ra) P(O) P(W|Ra,O) P(Ru|M,W)$$

$$= \sum_{Ra} \sum_{W} P(M) P(Ra) P(O) P(W|Ra,O) \cancel{\sum_{Ru} P(Ru|W,M)}$$

$$= \sum_{Ra} P(M) P(Ra) P(O) \cancel{\sum_{W} P(W|Ra,O)}$$

$$= P(M) P(O) \cancel{\sum_{Ra} P(Ra)}$$

		O	
		T	F
M	T	.92	.9.8
	F		