What's the probability that the next candy is lime?

What is $P(d_{i+1} \mid d_1, \ldots, d_i) = P(X \mid \boldsymbol{D})$?

We don't know which bag of candy we got, so we have to assume it could be any one of them:

$$P(X \mid \boldsymbol{D}) = \sum_i P(X \mid \boldsymbol{D}, h_i) P(h_i \mid \boldsymbol{D})$$
$$= \sum_i P(X \mid h_i) P(h_i \mid \boldsymbol{D})$$

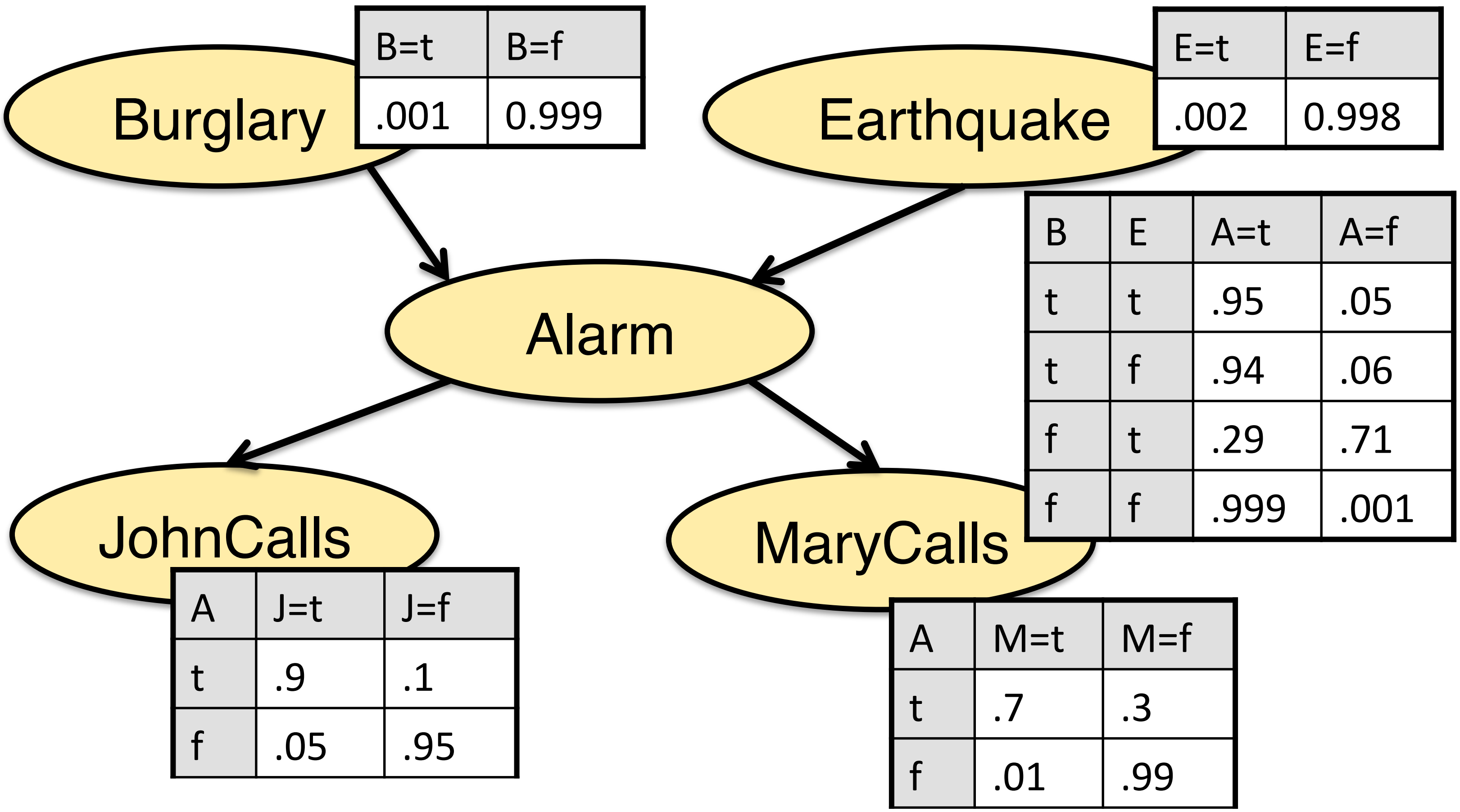
CS440/ECE448: Intro to Artificial Intelligence

Lecture 19
Learning graphical models

Prof. Julia Hockenmaier
juliahmr@illinois.edu

http://cs.illinois.edu/fa11/cs440

The *Burglary* example



B=t	B=f
.001	0.999

E=t	E=f
.002	0.998

B	E	A=t	A=f
t	t	.95	.05
t	f	.94	.06
f	t	.29	.71
f	f	.999	.001

A	J=t	J=f
t	.9	.1
f	.05	.95

A	M=t	M=f
t	.7	.3
f	.01	.99

What is the probability of a burglary if John and Mary call?

Learning Bayes Nets

How do we know the parameters of a Bayes Net?

We want to estimate the parameters based on data *D.*

Data = instantiations of some or all random variables in the Bayes Net.

The data are our evidence.

Surprise Candy

There are two flavors of Surprise Candy: cherry and lime. Both have the same wrapper.

There are five different types of bags (which all look the same) that Surprise Candy is sold in:

- h1: 100% cherry
- h2: 75% cherry + 25% lime
- h3: 50% cherry + 50% lime
- h4: 25% cherry + 75% lime
- h5: 100% lime

Surprise Candy

You just bought a bag of Surprise Candy.
Which kind of bag did you get?

There are five different hypotheses: h1-h5

You start eating your candy. This is your data
D1 = cherry, D2 = lime, …., DN= ….

What is *the most likely hypothesis given your data* (evidence)?

Conditional probability refresher

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

$$P(X \mid Y)P(Y) = P(X,Y)$$

$$P(X \mid Y)P(Y) = P(Y \mid X)P(X)$$

Bayes Rule

$$P(cause \mid effect) = \frac{P(effect \mid cause)P(cause)}{P(effect)}$$

$P(cause)$: prior probability of cause
$P(cause \mid effect)$: posterior probability of cause.
$P(effect \mid cause)$: likelihood of effect

Prior ∝ posterior × likelihood

$$P(cause \mid effect) \propto P(effect \mid cause)P(cause)$$

Bayes Rule

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

$P(h)$: prior probability of hypothesis
$P(h \mid D)$: posterior probability of hypothesis.
$P(D \mid h)$: likelihood of data, given hypothesis

Prior ∝ posterior × likelihood

$$P(h \mid D) \propto P(D \mid h)P(h)$$

Bayes Rule

$$\text{argmax}_h P(h \mid D)$$

$$= \text{argmax}_h \frac{P(D \mid h)P(h)}{P(D)}$$

$$= \arg\max_h P(D \mid h)P(h)$$

$P(h)$: prior probability of hypothesis

$P(h \mid D)$: posterior probability of hypothesis.

$P(D \mid h)$: likelihood of data, given hypothesis

Bayesian learning

Use Bayes rule to calculate the probability of each hypothesis given the data.

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

How do we know the prior and the likelihood?

The prior P(h)

Sometimes we know P(h) in advance.

– Surprise Candy: (0.1, 0.2, 0.4, 0.2, 0.1)

Sometimes we have to make an assumption (e.g. a uniform prior, when we don't know anything)

The likelihood *P(D|h)*

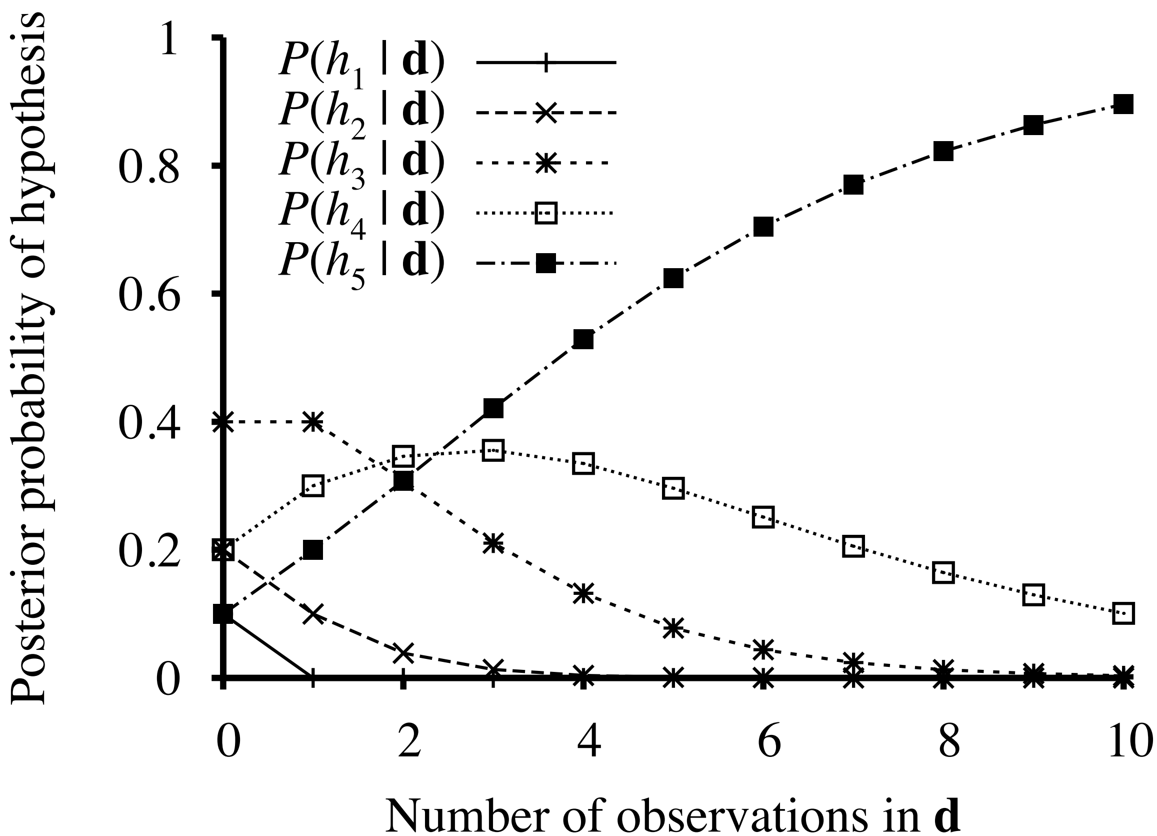
We typically assume that each observation d_i is drawn "i.i.d." - independently from the same (identical) distribution.

Therefore:

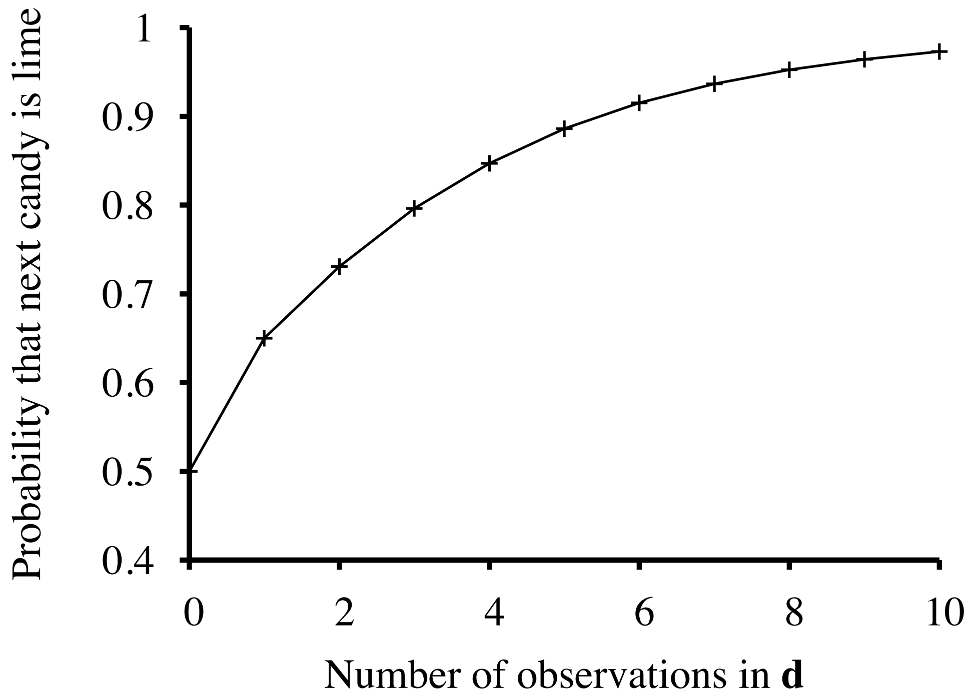
$$P(D \mid h) = \prod_i P(d_i \mid h)$$

The posterior P(h|D)

Assume we've seen 10 lime candies:



What's the probability that the next candy is lime?



This probability will eventually (if we had an infinite amount of data) agree with the true hypothesis.

Bayes optimal prediction

We don't know which hypothesis is true, so we marginalize them out:

$$P(X \mid \boldsymbol{D}) = \sum_i P(X \mid h_i)P(h_i \mid \boldsymbol{D})$$

This is guaranteed to converge to the true hypothesis.

Maximum a-posteriori (MAP)

We assume the hypothesis with the maximum posterior probability

$$h_{MAP} = argmax_h \, P(h|D)$$

is true:

$$P(X \mid \boldsymbol{D}) = P(X \mid h_{MAP})$$

Maximum likelihood (ML)

We assume a uniform prior $P(h)$.
We then choose the hypothesis that assigns the highest likelihood to the data

$$h_{ML} = argmax_{h} P(D|h)$$

$$P(X \mid \mathbf{D}) = P(X \mid h_{ML})$$

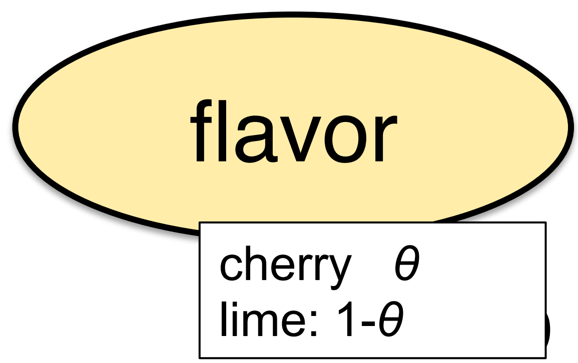
This is commonly used in machine learning.

Surprise candy again

Now the manufacturer has been bought up by another company.

Now we don't know the lime-cherry proportions θ *(= P(cherry))* anymore.

Can we estimate θ from data?



Maximum likelihood learning

Given data $\boldsymbol{D}$, we want to find the parameters that maximize $P(\boldsymbol{D} \mid \theta)$.

We have a data set with N candies.
c candies are cherry.
$l = (N\text{-}c)$ candies are lime.

Maximum likelihood learning

Out of *N* candies, *c* are cherry, *(N-c)* lime.

The likelihood of our data set:

$$P(\mathbf{d} \mid \theta) = \prod_{j=1}^{N} P(d_j \mid \theta) = \theta^c (1-\theta)^l$$

Log likelihood

It's actually easier to work with the log-likelihood:

$$L(\mathbf{d} \mid \theta) = \log \mathrm{P}(\mathbf{d}|\theta)$$

$$= \sum_{j=1}^{N} \log P(d_i \;\theta)$$

$$= \mathrm{c} \log\theta \;\; + l \log(1-\theta)$$

Maximizing Log-likelihood

$$\frac{dL(\mathbf{D} \mid \theta)}{d\theta} = \frac{c}{\theta} - \frac{l}{1-\theta} = 0$$

$$\Rightarrow \theta = \frac{c}{c+l} = \frac{c}{N}$$

Maximum likelihood estimation

We can simply count how many cherry candies we see.

This is also called the relative frequency estimate.

It is appropriate when we have complete data (i.e. we know the flavor of each candy).

Today's reading

Chapter 13.5, Chapter 20.1 and 20.2.1