Lecture 15a: Statistical Learning

CSCI 360

Introduction to Artificial Intelligence

USC

Here is where we are…

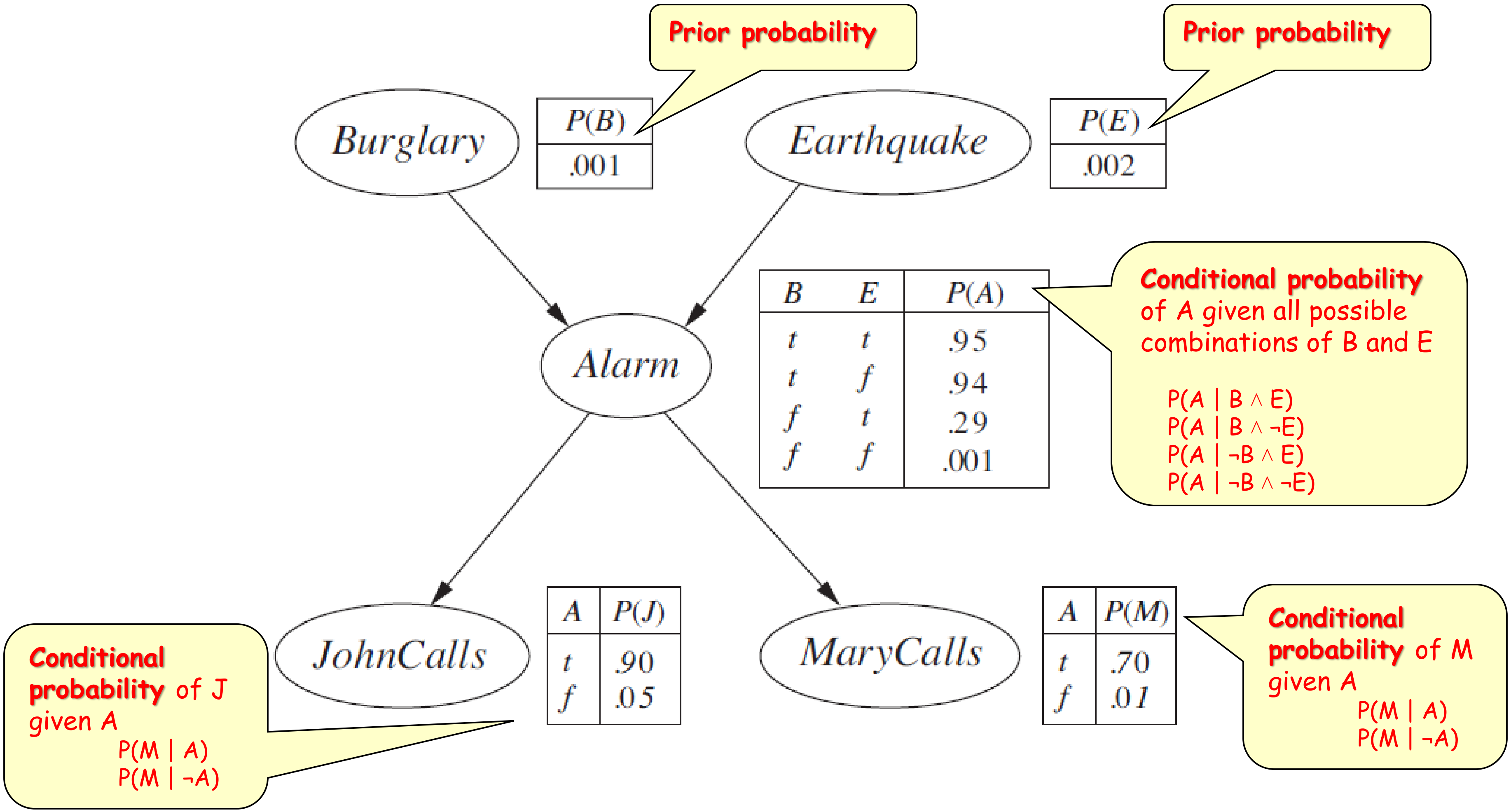
	3/1		Project 2 Out	
9	3/4	3/5	Quantifying Uncertainty	[Ch 13.1-13.6]
	3/6	3/7	Bayesian Networks	[Ch 14.1-14.2]
10	3/11	3/12	*(spring break, no class)*	
	3/13	3/14	*(spring break, no class)*	
11	3/18	3/19	Inference in Bayesian Networks	[Ch 14.3-14.4]
	3/20	3/21	Decision Theory	[Ch 16.1-16.3 and 16.5]
	3/23		Project 2 Due	
12	3/25	3/26	***Advanced topics*** *(Chao traveling to National Science Foundation)*	
	3/27	3/28	***Advanced topics*** *(Chao traveling to National Science Foundation)*	
	3/29		Homework 2 Out	
13	4/1	4/2	Markov Decision Processes	[Ch 17.1-17.2]
	4/3	4/4	Decision Tree Learning	[Ch 18.1-18.3]
	4/5		Homework 2 Due	
	4/5		Project 3 Out	
14	4/8	4/9	Perceptron Learning	[Ch 18.6]
	4/10	4/11	Neural Network Learning	[Ch 18.7]
15	4/15	4/16	Statistical Learning	[Ch 20.2.1-20.2.2]
	4/17	4/18	Reinforcement Learning	[Ch 21.1-21.2]
16	4/22	4/23	Artificial Intelligence Ethics	
	4/24	4/25	Wrap-Up and Final Review	
	4/26		Project 3 Due	
	5/3	5/2	**Final Exam** (2pm-4pm)	

Outline

- What is AI?
- Part I: Search
- Part II: Logical reasoning
- Part III: Probabilistic reasoning
- **Part IV: Machine learning**
  - Decision Tree Learning
  - Perceptron Learning
  - Neural Network Learning
  - **Statistical Learning**
  - Reinforcement Learning

Recap: *Bayesian networks (example)*

- Both the **topology** and the **conditional probability tables** (CPTs)



$P(B)$
.001

$P(E)$
.002

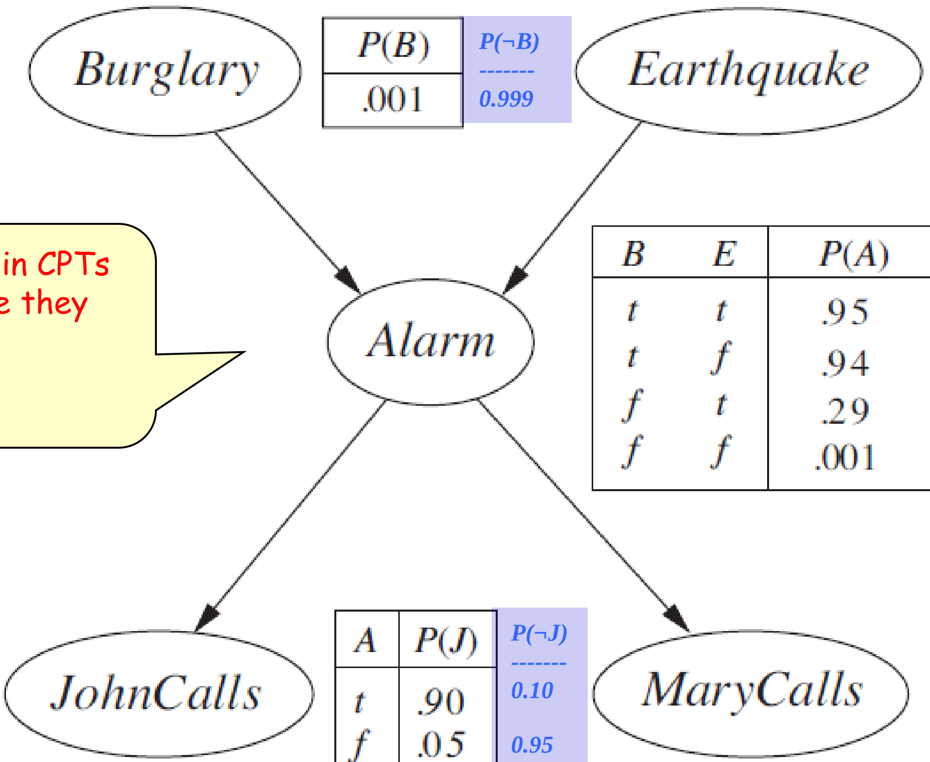
B	E	$P(A)$
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	$P(J)$
t	.90
f	.05

A	$P(M)$
t	.70
f	.01

Recap: *Bayesian networks (semantics)*

- Both the **topology** and the **conditional probability tables** (CPTs)



$P(B)$	$P(\neg B)$
.001	0.999

$P(E)$	$P(\neg E)$
.002	0.998

B	E	$P(A)$	$P(\neg A)$
t	t	.95	0.05
t	f	.94	0.06
f	t	.29	0.71
f	f	.001	0.999

A	$P(J)$	$P(\neg J)$
t	.90	0.10
f	.05	0.95

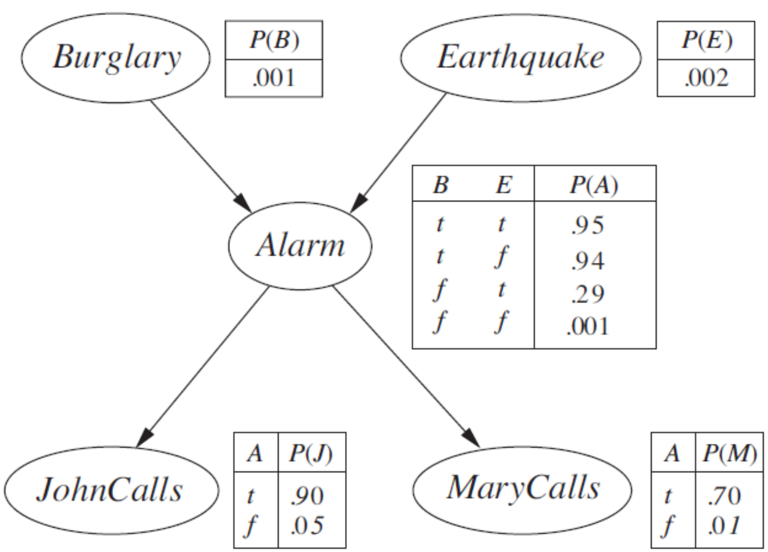
A	$P(M)$	$P(\neg M)$
t	.70	0.30
f	.01	0.99

Half of the entries in CPTs are omitted because they can be inferred

Recap: *Computing the joint distribution*

- Each entry $P(x_1,\ldots,x_n)$ in the full joint distribution, which is the abbreviation of $P(X_1=x_1 \wedge \ldots \wedge X_n=x_n)$ is the product of the elements of the CPTs defined as follows:

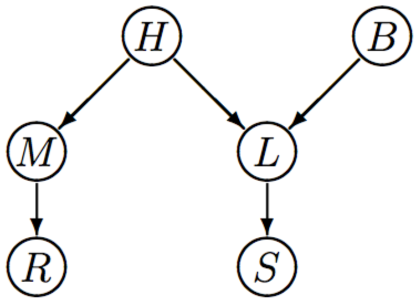
$$\boxed{P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i \mid parents(X_i))}$$



$P(j, m, a, b, e)$
$= P(j|a)\, P(m|a)\, P(a|b,e)\, \mathbf{P(b)\, P(e)}$
$= 0.90*0.70*0.95*\mathbf{0.001*0.002}$
$= 0.000001197$

$P(\neg m, j, \neg a, \neg e, b) = ?$

Recap: *Bayesian inference*



H = hardware problem
B = bug in lab
M = mailer is running
L = lab is running
R = received email
S = problem solved

Each node needs a probability table. Size of table depends on number of parents.

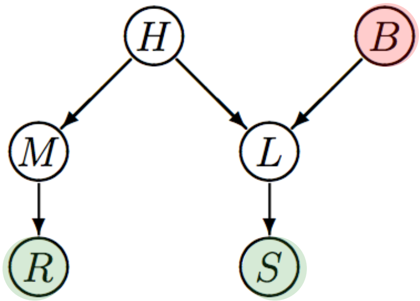
$\mathbf{P}(H)$	
$True$	$False$
0.01	0.99

H	$\mathbf{P}(M \mid H)$ $True$	$False$
$True$	0.1	0.9
$False$	0.99	0.01

H	B	$\mathbf{P}(L \mid H, B)$ $True$	$False$
$True$	$True$	0.01	0.99
$True$	$False$	0.1	0.9
$False$	$True$	0.02	0.98
$False$	$False$	1.0	0.0

$\ldots$, etc.

Recap: *Bayesian inference*



H = hardware problem
B = bug in lab
M = mailer is running
L = lab is running
R = received email
S = problem solved

- Compute $\mathbf{P}(B \mid \neg R, S)$

Each node needs a probability table. Size of table depends on number of parents.

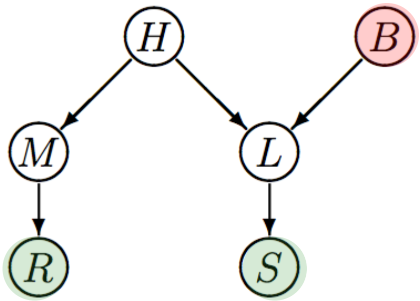
$\mathbf{P}(H)$	
True	*False*
0.01	0.99

H	$\mathbf{P}(M \mid H)$ *True*	*False*
True	0.1	0.9
False	0.99	0.01

H	B	$\mathbf{P}(L \mid H, B)$ *True*	*False*
True	*True*	0.01	0.99
True	*False*	0.1	0.9
False	*True*	0.02	0.98
False	*False*	1.0	0.0

…, etc.

Recap: *Bayesian inference*



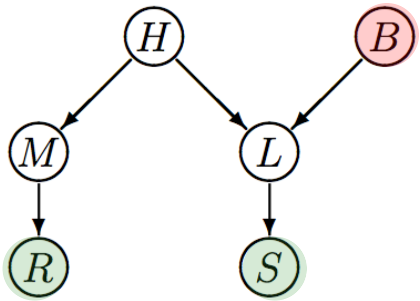
H = hardware problem
B = bug in lab
M = mailer is running
L = lab is running
R = received email
S = problem solved

- Compute **P** (B | $\neg R,S$)

1. Apply the conditional probability rule.

$$P(B \mid \neg R, S) = \frac{P(B, \neg R, S)}{P(\neg R, S)}$$

Recap: *Bayesian inference*



H = hardware problem
B = bug in lab
M = mailer is running
L = lab is running
R = received email
S = problem solved

- Compute **P** (B | $\neg R, S$)

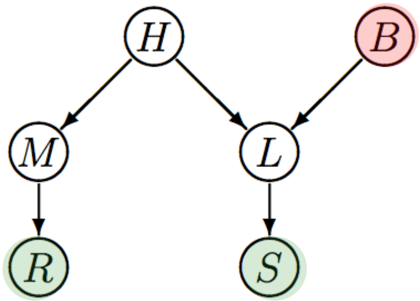
1. Apply the conditional probability rule.

$$P(B \mid \neg R, S) = \frac{P(B, \neg R, S)}{P(\neg R, S)}$$

2. Apply the marginal distribution rule to the unknown vertices. $P(B, \neg R, S)$ has 3 unknown vertices with $2^3 = 8$ possible value assignments.

$$\begin{aligned}
&P(B, \neg R, S) \\
&\quad = P(B, \neg R, S, H, M, L) \\
&\qquad + P(B, \neg R, S, H, M, \neg L) \\
&\qquad + P(B, \neg R, S, H, \neg M, L) \\
&\qquad + P(B, \neg R, S, H, \neg M, \neg L) \\
&\qquad + P(B, \neg R, S, \neg H, M, L) \\
&\qquad + P(B, \neg R, S, \neg H, M, \neg L) \\
&\qquad + P(B, \neg R, S, \neg H, \neg M, L) \\
&\qquad + P(B, \neg R, S, \neg H, \neg M, \neg L)
\end{aligned}$$

Recap: *Bayesian inference*



H = hardware problem
B = bug in lab
M = mailer is running
L = lab is running
R = received email
S = problem solved

- Compute **P** (B | $\neg R, S$)

1. Apply the conditional probability rule.

$$P(B \mid \neg R, S) = \frac{P(B, \neg R, S)}{P(\neg R, S)}$$

2. Apply the marginal distribution rule to the unknown vertices. $P(B, \neg R, S)$ has 3 unknown vertices with $2^3 = 8$ possible value assignments.

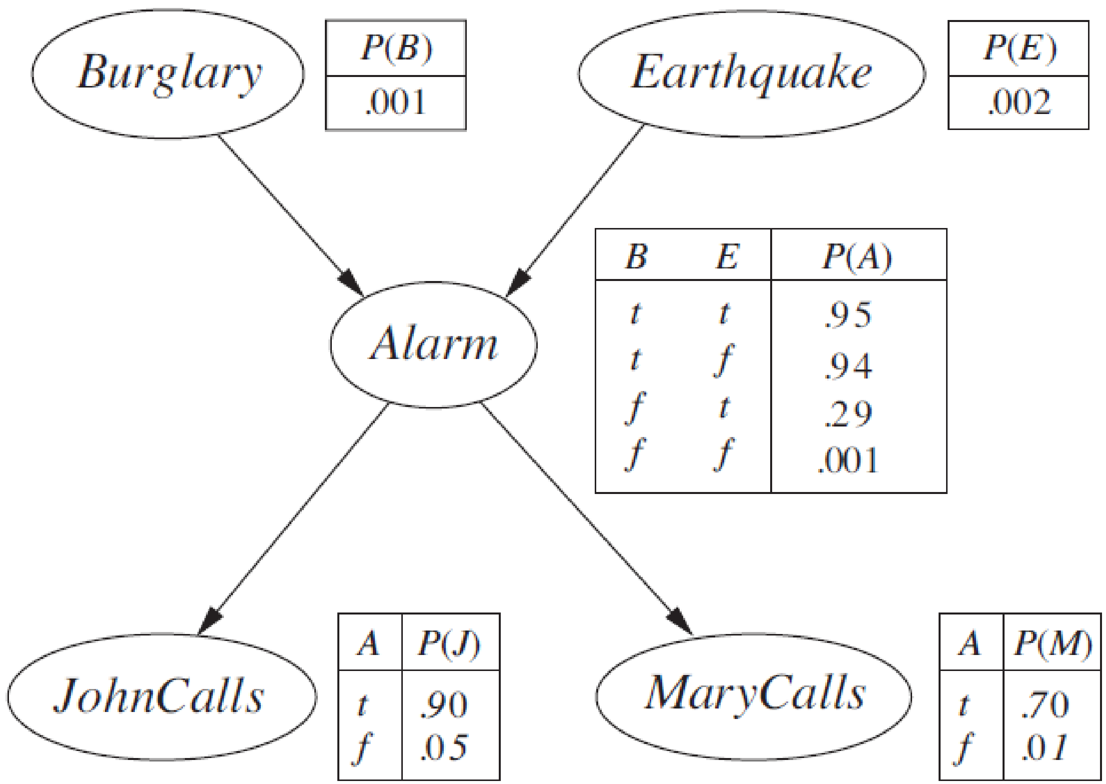
$$\begin{aligned}
&P(B, \neg R, S) \\
&= P(B, \neg R, S, H, M, L) \\
&\quad + P(B, \neg R, S, H, M, \neg L) \\
&\quad + P(B, \neg R, S, H, \neg M, L) \\
&\quad + P(B, \neg R, S, H, \neg M, \neg L) \\
&\quad + P(B, \neg R, S, \neg H, M, L) \\
&\quad + P(B, \neg R, S, \neg H, M, \neg L) \\
&\quad + P(B, \neg R, S, \neg H, \neg M, L) \\
&\quad + P(B, \neg R, S, \neg H, \neg M, \neg L)
\end{aligned}$$

3. Apply joint distribution rule for Bayesian networks.

$$\begin{aligned}
&P(B, \neg R, S, H, M, L) \\
&= P(B)\, P(H) \\
&\quad P(M \mid H)\, P(\neg R \mid M) \\
&\quad P(L \mid H, B)\, P(S \mid L)
\end{aligned}$$

Question: *Where do the CPTs come from?*

- **Answer**: Learning <u>probabilistic models</u> from **data**



Outline of today's lecture

- Statistical learning
- Maximum-likelihood parameter learning
- Naïve Bayes models

Data and hypotheses

- Data (as the evidence)
  - Instantiations of random variables that describe the domain
- Hypotheses
  - Probabilistic theories of how the domain works
- Learning
  - Learn the **hypothesis** that best fit the **data**

Statistical learning example

- Candy in two flavors indistinguishable from the outside
  - cherry
  - lime
- Sold in large bags of five kinds:
  - h_1: 100% cherry
  - h_2: 75% cherry + 25% lime
  - h_3: 50% cherry + 50% cherry
  - h_4: 25% cherry + 75% lime
  - h_5: 100% lime

Learning problem: Which bag ($H = h_1$, … , or $H = h_5$) is *this* ?

Statistical learning example

- Candy in two flavors indistinguishable from the outside
  - cherry
  - lime
- Sold in large bags of five kinds:
  - h_1: 100% cherry
  - h_2: 75% cherry + 25% lime
  - h_3: 50% cherry + 50% cherry
  - h_4: 25% cherry + 75% lime
  - h_5: 100% lime

Learning problem: Which bag ($H = h_1$, … , or $H = h_5$) , *given favors of randomly drawn candies* (D_1, …, D_N) ?

Bayesian learning

- Calculate the **probability of each hypothesis**, given the **data (d)**, and make the prediction on that basis
  - $P(h_1 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_1)\ P(h_1)$
  - $P(h_2 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_2)\ P(h_2)$
  - $P(h_3 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_3)\ P(h_3)$
  - $P(h_4 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_4)\ P(h_4)$
  - $P(h_5 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_5)\ P(h_5)$

Likelihood of the data under each hypothesis $P(d \mid h_i)$

Hypothesis prior $P(h_i)$

Bayesian learning

- Calculate the **probability of each hypothesis**, given the **data (d)**, and make the prediction on that basis
  - $P(h_1 \mid \mathbf{d}) = \alpha P(\mathbf{d} \mid h_1)\ P(h_1) = \alpha P(\mathbf{d} \mid h_1) * 0.1$
  - $P(h_2 \mid \mathbf{d}) = \alpha P(\mathbf{d} \mid h_2)\ P(h_2) = \alpha P(\mathbf{d} \mid h_2) * 0.2$
  - $P(h_3 \mid \mathbf{d}) = \alpha P(\mathbf{d} \mid h_3)\ P(h_3) = \alpha P(\mathbf{d} \mid h_3) * 0.4$
  - $P(h_4 \mid \mathbf{d}) = \alpha P(\mathbf{d} \mid h_4)\ P(h_4) = \alpha P(\mathbf{d} \mid h_4) * 0.2$
  - $P(h_5 \mid \mathbf{d}) = \alpha P(\mathbf{d} \mid h_5)\ P(h_5) = \alpha P(\mathbf{d} \mid h_5) * 0.1$

Likelihood of the data under each hypothesis $P(d \mid h_i)$

Hypothesis prior $P(h_i)$

Bayesian learning

- Calculate the **probability of each hypothesis**, given the **data (d =** $\{d_1, \ldots, d_5\}$**)**, and make prediction on that basis
  - $P(h_1 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_1)\ P(h_1) = \alpha\, P(\mathbf{d} \mid h_1) * 0.1$
  - $P(h_2 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_2)\ P(h_2) = \alpha\, P(\mathbf{d} \mid h_2) * 0.2$
  - $P(h_3 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_3)\ P(h_3) = \alpha\, P(\mathbf{d} \mid h_3) * 0.4$
  - $P(h_4 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_4)\ P(h_4) = \alpha\, P(\mathbf{d} \mid h_4) * 0.2$
  - $P(h_5 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_5)\ P(h_5) = \alpha\, P(\mathbf{d} \mid h_5) * 0.1$

Likelihood of the data under each hypothesis $P(d \mid h_i)$

Hypothesis prior $P(h_i)$

$$P(\mathbf{d} \mid h_i) = \prod_j P(d_j \mid h_i)$$

Bayesian learning

- Calculate the **probability of each hypothesis**, given the **data**, and make the prediction on that basis
  - $P(h_1 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_1)\, P(h_1) = \alpha\, P(\mathbf{d} \mid h_1) * 0.1 = \alpha * 0 * 0.1$
  - $P(h_2 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_2)\, P(h_2) = \alpha\, P(\mathbf{d} \mid h_2) * 0.2 = \alpha * 0.25^{10} * 0.2$
  - $P(h_3 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_3)\, P(h_3) = \alpha\, P(\mathbf{d} \mid h_3) * 0.4 = \alpha * 0.50^{10} * 0.4$
  - $P(h_4 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_4)\, P(h_4) = \alpha\, P(\mathbf{d} \mid h_4) * 0.2 = \alpha * 0.75^{10} * 0.2$
  - $P(h_5 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_5)\, P(h_5) = \alpha\, P(\mathbf{d} \mid h_5) * 0.1 = \alpha * 1 * 0.1$

Hypothesis prior
$P(h_i)$

Likelihood of the data
under each hypothesis
$P(d \mid h_i)$

$$P(\mathbf{d} \mid h_i) = \prod_j P(d_j \mid h_i)$$

First 10 candies are
all "lime"

Bayesian learning *(in general)*

- Calculate the **probability of each hypothesis**, given the **data**, and make the prediction of **unknown quantity X**

$$\mathbf{P}(X \mid \mathbf{d}) = \sum_i \mathbf{P}(X \mid \mathbf{d}, h_i)\mathbf{P}(h_i \mid \mathbf{d}) = \sum_i \mathbf{P}(X \mid h_i)P(h_i \mid \mathbf{d})$$

Each hypothesis (h_i) determines a probability distribution over X

$$P(h_i \mid \mathbf{d}) = \alpha P(\mathbf{d} \mid h_i)P(h_i)$$

$$P(\mathbf{d} \mid h_i) = \prod_j P(d_j \mid h_i)$$

Bayesian learning *(in general)*

- Calculate the **probability of each hypothesis**, given the **data**, and make the prediction of **unknown quantity X**
  - Step 1. Computing the probability of data, given each hypothesis

$$\boxed{P(\mathbf{d} \mid h_i) = \prod_j P(d_j \mid h_i)}$$

  - Step 2. Computing the probability of each hypothesis

$$\boxed{P(h_i \mid \mathbf{d}) = \alpha P(\mathbf{d} \mid h_i) P(h_i)}$$

  - Step 3. Computing the probability of X, given the data

$$\boxed{\mathbf{P}(X \mid \mathbf{d}) = \sum_i \mathbf{P}(X \mid \mathbf{d}, h_i) \mathbf{P}(h_i \mid \mathbf{d}) = \sum_i \mathbf{P}(X \mid h_i) P(h_i \mid \mathbf{d})}$$

Bayesian learning *(before any candy is revealed)*

- Calculate the **probability of each hypothesis**, given the **data**, and make the prediction on that basis
  - $P(h_1 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_1) * 0.1 = \alpha * \quad * 0.1 = \alpha * 0.1$
  - $P(h_2 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_2) * 0.2 = \alpha * \quad * 0.2 = \alpha * 0.2$
  - $P(h_3 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_3) * 0.4 = \alpha * \quad * 0.4 = \alpha * \mathbf{0.4}$
  - $P(h_4 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_4) * 0.2 = \alpha * \quad * 0.2 = \alpha * 0.2$
  - $P(h_5 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_5) * 0.1 = \alpha * \quad * 0.1 = \alpha * 0.1$

Know nothing about the candies in this bag

Bayesian learning *(after 1 candy is revealed)*

- Calculate the **probability of each hypothesis**, given the **data**, and make the prediction on that basis
  - $P(h_1 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_1) * 0.1 = \alpha * 0 \quad * 0.1 \quad = \alpha * 0.00$
  - $P(h_2 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_2) * 0.2 = \alpha * 0.25^1 * 0.2 \quad = \alpha * 0.05$
  - $P(h_3 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_3) * 0.4 = \alpha * 0.50^1 * 0.4 \quad = \alpha * \mathbf{0.20}$
  - $P(h_4 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_4) * 0.2 = \alpha * 0.75^1 * 0.2 \quad = \alpha * 0.15$
  - $P(h_5 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_5) * 0.1 = \alpha * 1 \quad * 0.1 \quad = \alpha * 0.10$

First candy is “lime”

Bayesian learning *(after 2 candies are revealed)*

- Calculate the **probability of each hypothesis**, given the **data**, and make the prediction on that basis
  - $P(h_1 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_1) * 0.1 = \alpha * 0 * 0.1 = \alpha * 0.0000$
  - $P(h_2 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_2) * 0.2 = \alpha * 0.25^2 * 0.2 = \alpha * 0.0125$
  - $P(h_3 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_3) * 0.4 = \alpha * 0.50^2 * 0.4 = \alpha * \mathbf{0.1000}$
  - $P(h_4 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_4) * 0.2 = \alpha * 0.75^2 * 0.2 = \alpha * \mathbf{0.1125}$
  - $P(h_5 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_5) * 0.1 = \alpha * 1 * 0.1 = \alpha * 0.1000$

First 2 candies are "lime"

Bayesian learning *(after 3 candies are revealed)*

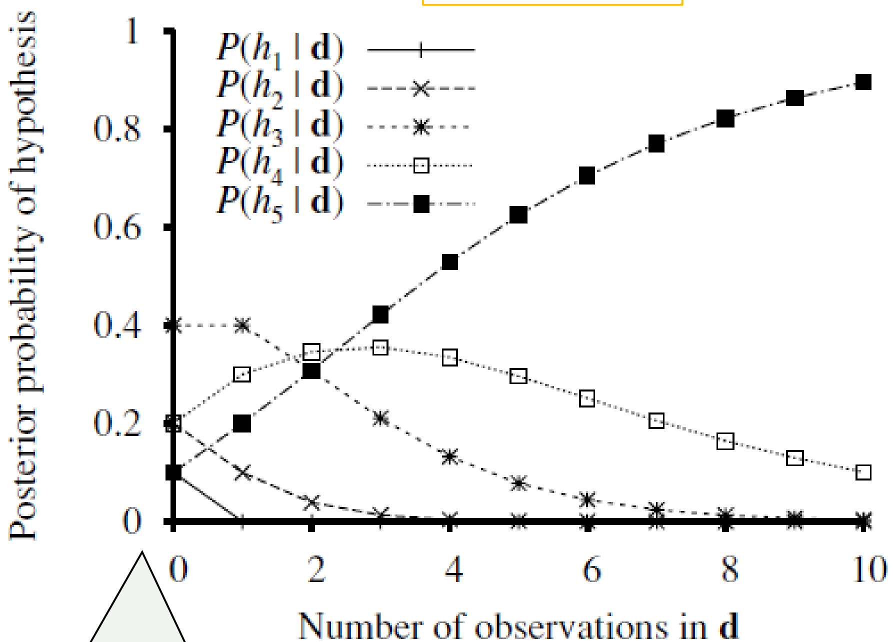
- Calculate the **probability of each hypothesis**, given the **data**, and make the prediction on that basis
  - $P(h_1 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_1) * 0.1 = \alpha * 0 \quad * 0.1 = \alpha * 0.000000$
  - $P(h_2 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_2) * 0.2 = \alpha * 0.25^2 * 0.2 = \alpha * 0.003125$
  - $P(h_3 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_3) * 0.4 = \alpha * 0.50^3 * 0.4 = \alpha * \mathbf{0.050000}$
  - $P(h_4 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_4) * 0.2 = \alpha * 0.75^3 * 0.2 = \alpha * \mathbf{0.084375}$
  - $P(h_5 \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_5) * 0.1 = \alpha * 1 \quad * 0.1 = \alpha * \mathbf{0.100000}$

First 3 candies are “lime”

Bayesian learning (results)

- Bayesian prediction **eventually** agrees with the **true hypothesis**

$P(h_i \mid d_1, \ldots, d_N)$



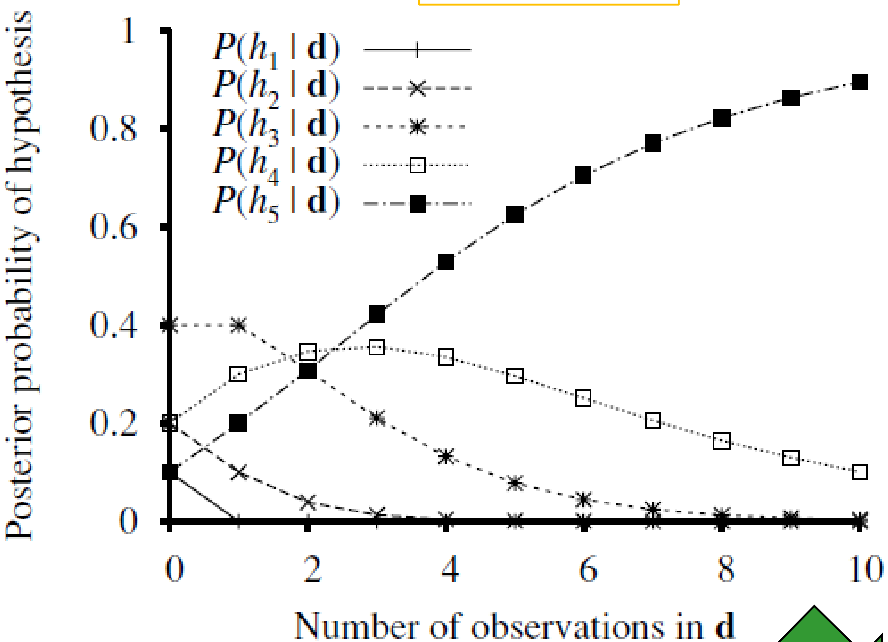
Eventually, the probability of the **true hypothesis** rises to the top

Initially, it is equal to the ***prior*** probability of the hypothesis

Bayesian learning (results)

- Bayesian prediction **eventually** agrees with the **true hypothesis**

$P(h_i \mid d_1, \ldots, d_N)$



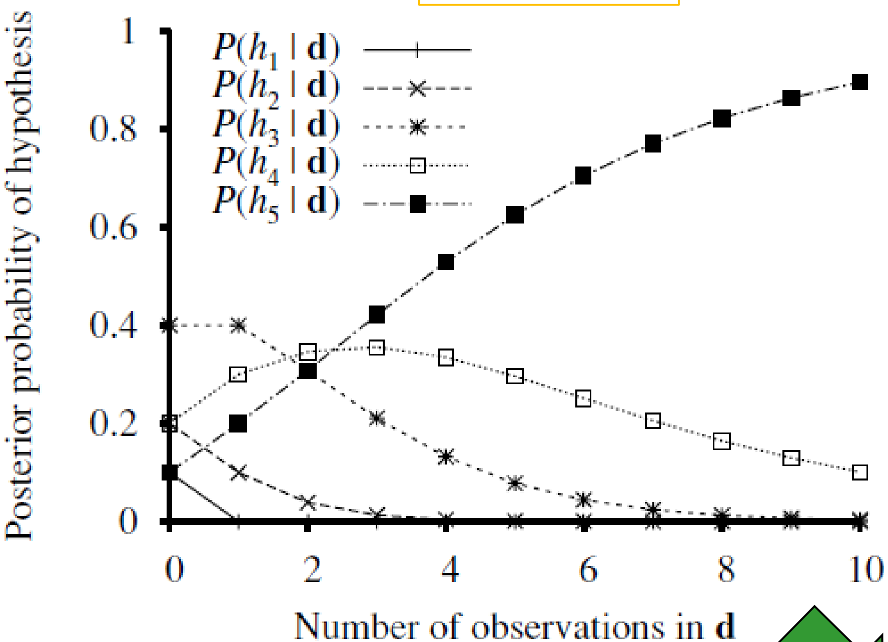
X is the quantity to be predicted

$$\mathbf{P}(X \mid \mathbf{d}) = \sum_i \mathbf{P}(X \mid h_i) P(h_i \mid \mathbf{d})$$

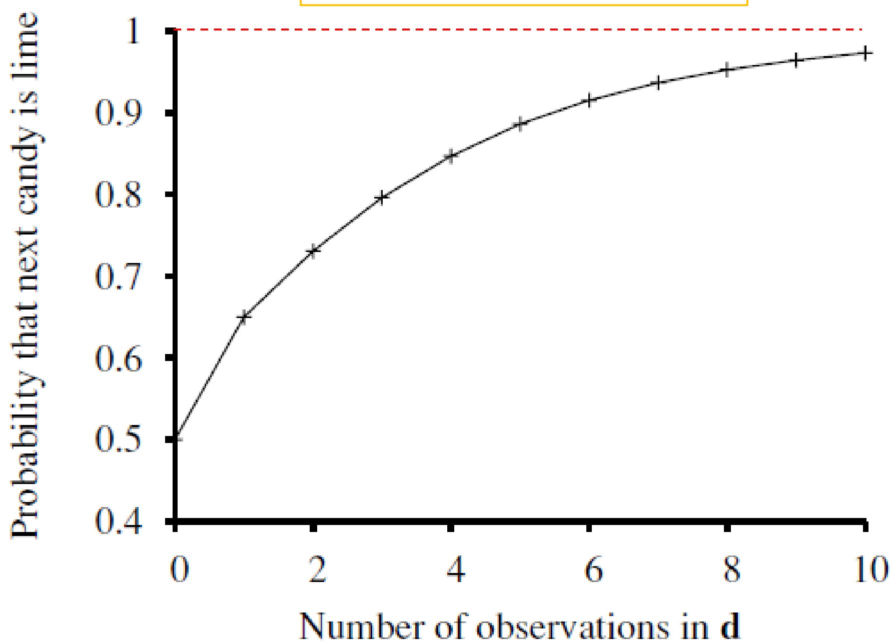
Bayesian learning (results)

- Bayesian prediction **eventually** agrees with the **true hypothesis**

$P(h_i \mid d_1, \ldots, d_N)$



$P(d_{N+1} = lime \mid d_1, \ldots, d_N)$



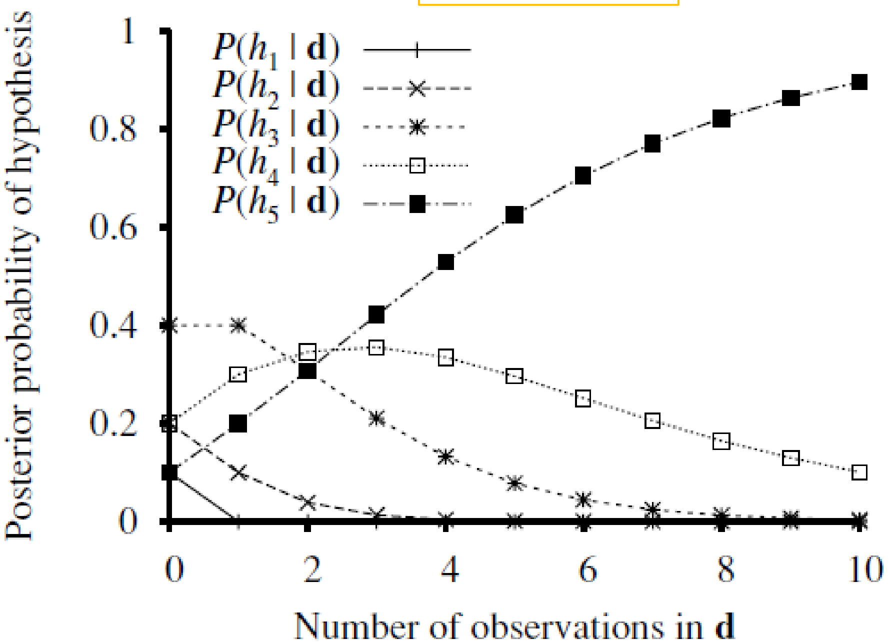
$$\mathbf{P}(X \mid \mathbf{d}) = \sum_i \mathbf{P}(X \mid h_i) P(h_i \mid \mathbf{d})$$

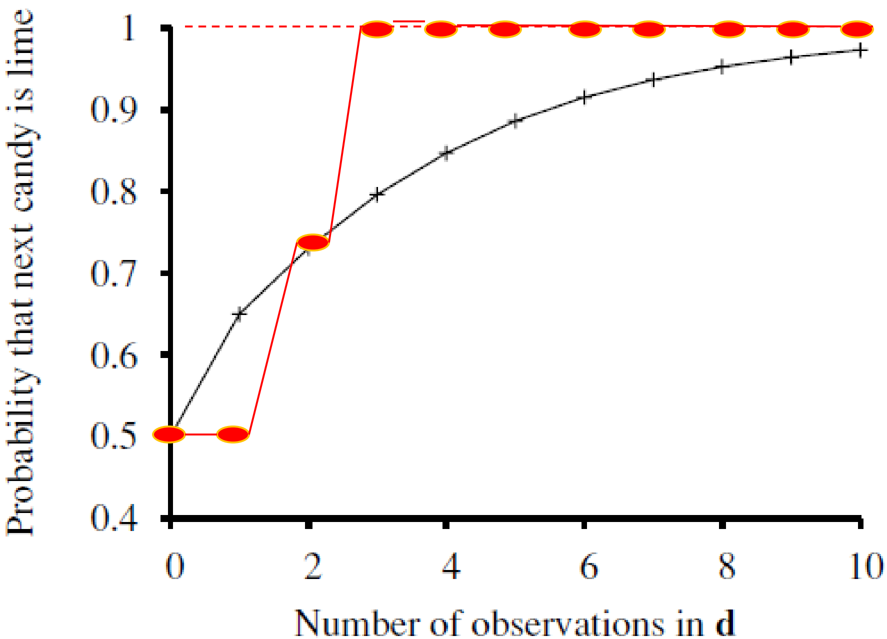
Maximum a posteriori (MAP)

- Make prediction based on a single, **most probable**, hypothesis

$$\boxed{\mathbf{P}(X \mid \mathbf{d}) \approx \mathbf{P}(X \mid h_{\mathrm{MAP}})}$$

$\boxed{P(h_i \mid d_1, \ldots, d_N)}$





$$\cancel{\mathbf{P}(X \mid \mathbf{d}) = \sum_i \mathbf{P}(X \mid h_i) P(h_i \mid \mathbf{d})}$$

Bayesian learning *(MAP)*

- Approximating Bayesian prediction (**weighted sum**) by making decisions using the **most probable hypothesis**

$$\mathbf{P}(X \mid \mathbf{d}) = \sum_i \mathbf{P}(X \mid \mathbf{d}, h_i)\mathbf{P}(h_i \mid \mathbf{d}) = \sum_i \mathbf{P}(X \mid h_i)P(h_i \mid \mathbf{d})$$

$$\boxed{\mathbf{P}(X \mid \mathbf{d}) \approx \mathbf{P}(X \mid h_{\mathrm{MAP}})}$$

h_{MAP} is the hypothesis with largest $P(h_i|\mathbf{d})$

$$\boxed{P(h_i \mid \mathbf{d}) = \alpha P(\mathbf{d} \mid h_i)P(h_i)}$$

$$\boxed{P(\mathbf{d} \mid h_i) = \prod_j P(d_j \mid h_i)}$$

The role of prior – $P(h_i)$

- Bayesian (and MAP) learning uses **the prior** to penalize complexity
  - Complex hypothesis has a lower prior probability

$$\boxed{\mathbf{P}(X \mid \mathbf{d}) \approx \mathbf{P}(X \mid h_{\mathrm{MAP}})}$$

h_{MAP} is the hypothesis with largest $P(h_i|d)$

$$\boxed{P(h_i \mid \mathbf{d}) = \alpha P(\mathbf{d} \mid h_i) P(h_i)}$$

$$\boxed{P(\mathbf{d} \mid h_i) = \prod_j P(d_j \mid h_i)}$$

Ockham's razor: find the simplest hypothesis that is consistent with the data

The role of prior – $P(h_i)$

- **Maximum-likelihood** hypothesis can be learned by using a **uniform prior**
  - Equivalent to maximizing ***P(d | h_i)***

$$\boxed{\mathbf{P}(X \mid \mathbf{d}) \approx \mathbf{P}(X \mid h_{\mathrm{MAP}})}$$

h_{MAP} is the hypothesis with largest ***$P(h_i|d)$***

$$P(h_i \mid \mathbf{d}) = \alpha P(\mathbf{d} \mid h_i) \cancel{P(h_i)}$$

$$P(\mathbf{d} \mid h_i) = \prod_j P(d_j \mid h_i)$$

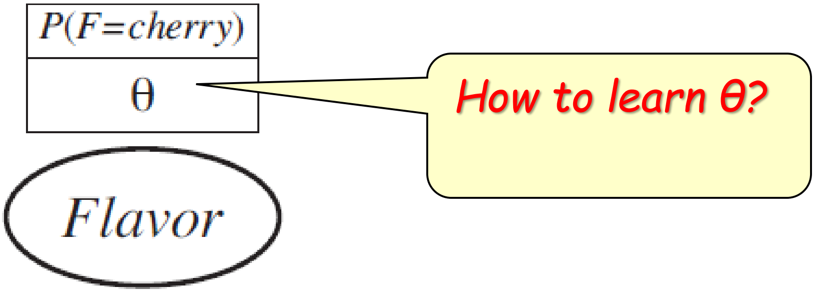
It provides a good approximation to Bayesian and MAP learning when the data set is large… but do not work well with a small data set

Outline of today's lecture

- Statistical learning
- Maximum-likelihood parameter learning
- Naïve Bayes models

Density estimation

- Learning a **probability model**, given **data** generated from that model
  - E.g., finding the **conditional probabilities** in a **Bayesian network** whose structure is fixed

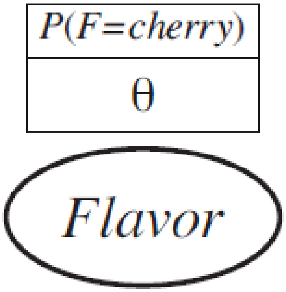


Likelihood

- Let the proportions of "**cherry**" and "**lime**" candies be (θ) and ($1-\theta$), respectively.
- If we unwrap N candies, and find that there are
  - (c) **cherry** candies
  - ($l = N - c$) **lime** candies

- The likelihood of this data set (**d**) is

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



Log likelihood

- Let the proportions of "**cherry**" and "**lime**" candies be (θ) and (1-θ), respectively.
- If we unwrap N candies, and find that there are
  - (c) **cherry** candies
  - ($l = N - c$) **lime** candies

- The likelihood of this data set is

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

$$\boxed{L(\mathbf{d} \mid h_\theta) = \log P(\mathbf{d} \mid h_\theta) = \sum_{j=1}^{N} \log P(d_j \mid h_\theta) = c \log \theta + \ell \log(1-\theta)}$$

Maximizing the log likelihood

- To find the maximum, compute the derivative of *$L(d \mid h_\theta)$* and set it to zero.

$$L(\mathbf{d} \mid h_\theta) = \log P(\mathbf{d} \mid h_\theta) = \sum_{j=1}^{N} \log P(d_j \mid h_\theta) = c \log \theta + \ell \log(1-\theta)$$

$$\frac{dL(\mathbf{d} \mid h_\theta)}{d\theta}$$

$$= \frac{c}{\theta} - \frac{\ell}{1-\theta}$$

$$= 0 \qquad \Rightarrow \quad \theta = \frac{c}{c+\ell} = \frac{c}{N}$$

Maximizing the log likelihood

- To find the maximum, compute the derivative of $L(d \mid h_\theta)$ and set it to zero.

$$L(\mathbf{d} \mid h_\theta) = \log P(\mathbf{d} \mid h_\theta) = \sum_{j=1}^{N} \log P(d_j \mid h_\theta) = c \log \theta + \ell \log(1-\theta)$$

$$\frac{dL(\mathbf{d} \mid h_\theta)}{d\theta}$$

$$= \frac{c}{\theta} - \frac{\ell}{1-\theta}$$

$$= 0$$

The maximum-likelihood hypothesis (hML) says that the **proportion (θ) of cherry candies** is equal to the **observed proportion**

$$\Rightarrow \quad \theta = \frac{c}{c+\ell} = \frac{c}{N}$$

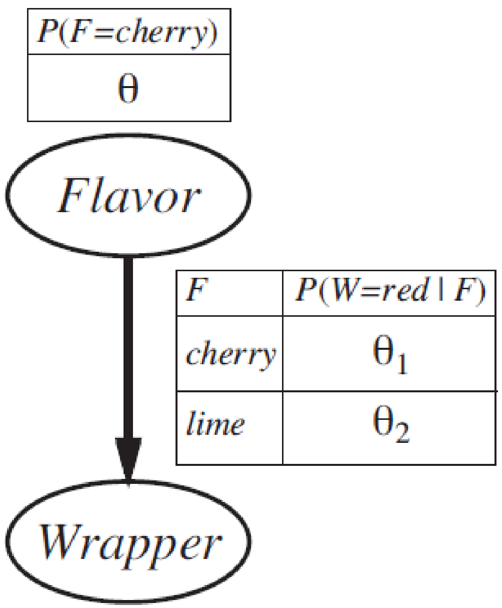
Obvious, but also comforting

General method *for maximum-likelihood parameter learning*

- Write down the **likelihood of the data** as a function of the parameters
- Compute the **derivative** of the log **log likelihood** w.r.t. each parameter
- Find the parameter values such that the **derivatives are zero**

Another example

- Wrapper for each candy is chosen probabilistically based the flavor
  - But the conditional distributions are unknown



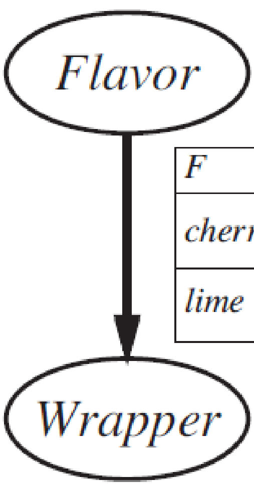
Question: What's the likelihood of seeing a **cherry** candy in a **green** wrapper?

Another example

- Wrapper for each candy is chosen probabilistically based the flavor
  - But the conditional distributions are unknown

$$P(Flavor = cherry, Wrapper = green \mid h_{\theta,\theta_1,\theta_2})$$

$P(F{=}cherry)$
θ



F	$P(W{=}red \mid F)$	$P(W{=}green \mid F)$
cherry	θ_1	$(1-\theta_1)$
lime	θ_2	$(1-\theta_2)$

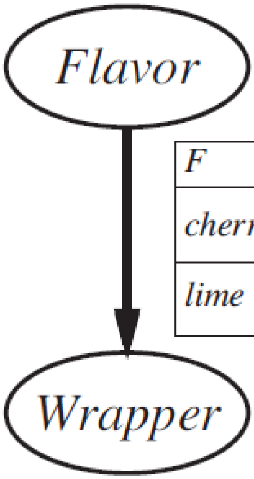
Question: What's the likelihood of seeing a **cherry** candy in a **green** wrapper?

Another example

- Wrapper for each candy is chosen probabilistically based the flavor
  - But the conditional distributions are unknown

$$
\begin{aligned}
& P(Flavor = cherry, Wrapper = green \mid h_{\theta,\theta_1,\theta_2}) \\
&= P(Flavor = cherry \mid h_{\theta,\theta_1,\theta_2}) P(Wrapper = green \mid Flavor = cherry, h_{\theta,\theta_1,\theta_2}) \\
&= \theta \cdot (1 - \theta_1)\ .
\end{aligned}
$$

$P(F{=}cherry)$
θ



F	$P(W{=}red \mid F)$	$P(W{=}green \mid F)$
cherry	θ_1	$(1 - \theta_1)$
lime	θ_2	$(1 - \theta_2)$

Question: What’s the likelihood of seeing a **cherry** candy in a **green** wrapper?

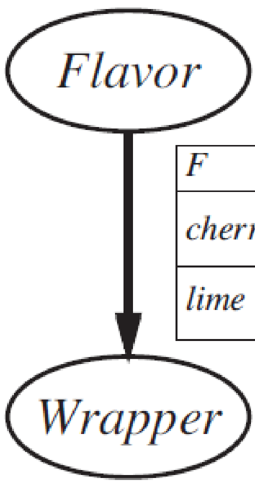
Another example

- Data: Unwrapping **N** candies, of which **c** are cherries

$$\boxed{P(\mathbf{d} \mid h_{\theta,\theta_1,\theta_2}) = \theta^c (1-\theta)^{\ell} \cdot \theta_1^{r_c}(1-\theta_1)^{g_c} \cdot \theta_2^{r_\ell}(1-\theta_2)^{g_\ell}}$$

$$\boxed{L = [c \log \theta + \ell \log(1-\theta)] + [r_c \log \theta_1 + g_c \log(1-\theta_1)] + [r_\ell \log \theta_2 + g_\ell \log(1-\theta_2)]}$$

P(F=cherry)
θ



F	*P(W=red* | *F)*	*P(W=green*|*F)*
cherry	θ_1	$(1-\theta_1)$
lime	θ_2	$(1-\theta_2)$

Another example

- Data: Unwrapping **N** candies, of which **c** are cherries

$$\boxed{P(\mathbf{d} \mid h_{\theta,\theta_1,\theta_2}) = \theta^c(1-\theta)^\ell \cdot \theta_1^{r_c}(1-\theta_1)^{g_c} \cdot \theta_2^{r_\ell}(1-\theta_2)^{g_\ell}}$$

$$\boxed{L = [c\log\theta + \ell\log(1-\theta)] + [r_c\log\theta_1 + g_c\log(1-\theta_1)] + [r_\ell\log\theta_2 + g_\ell\log(1-\theta_2)]}$$

- Derivatives

$$\begin{aligned}
\frac{\partial L}{\partial \theta} &= \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 \\
\frac{\partial L}{\partial \theta_1} &= \frac{r_c}{\theta_1} - \frac{g_c}{1-\theta_1} = 0 \\
\frac{\partial L}{\partial \theta_2} &= \frac{r_\ell}{\theta_2} - \frac{g_\ell}{1-\theta_2} = 0
\end{aligned}$$

Another example

- Data: Unwrapping **N** candies, of which **c** are cherries

$$\boxed{P(\mathbf{d} \mid h_{\theta,\theta_1,\theta_2}) = \theta^c(1-\theta)^\ell \cdot \theta_1^{r_c}(1-\theta_1)^{g_c} \cdot \theta_2^{r_\ell}(1-\theta_2)^{g_\ell}}$$

$$\boxed{L = [c\log\theta + \ell\log(1-\theta)] + [r_c\log\theta_1 + g_c\log(1-\theta_1)] + [r_\ell\log\theta_2 + g_\ell\log(1-\theta_2)]}$$

- Derivatives

$$\begin{aligned}
\frac{\partial L}{\partial \theta} &= \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 & \Rightarrow \quad \theta &= \frac{c}{c+\ell} \\
\frac{\partial L}{\partial \theta_1} &= \frac{r_c}{\theta_1} - \frac{g_c}{1-\theta_1} = 0 & \Rightarrow \quad \theta_1 &= \frac{r_c}{r_c+g_c} \\
\frac{\partial L}{\partial \theta_2} &= \frac{r_\ell}{\theta_2} - \frac{g_\ell}{1-\theta_2} = 0 & \Rightarrow \quad \theta_2 &= \frac{r_\ell}{r_\ell+g_\ell}
\end{aligned}$$

Another example

- Data: Unwrapping **N** candies, of which **c** are cherries

$$\boxed{P(\mathbf{d} \mid h_{\theta,\theta_1,\theta_2}) = \theta^c (1-\theta)^\ell \cdot \theta_1^{r_c}(1-\theta_1)^{g_c} \cdot \theta_2^{r_\ell}(1-\theta_2)^{g_\ell}}$$

$$\boxed{L = [c \log\theta + \ell \log(1-\theta)] + [r_c \log\theta_1 + g_c \log(1-\theta_1)] + [r_\ell \log\theta_2 + g_\ell \log(1-\theta_2)]}$$

- Derivatives

Same as before – obvious, but also comforting

$$\begin{aligned}
\frac{\partial L}{\partial \theta} &= \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 && \Rightarrow && \theta = \frac{c}{c+\ell} \\
\frac{\partial L}{\partial \theta_1} &= \frac{r_c}{\theta_1} - \frac{g_c}{1-\theta_1} = 0 && \Rightarrow && \theta_1 = \frac{r_c}{r_c+g_c} \\
\frac{\partial L}{\partial \theta_2} &= \frac{r_\ell}{\theta_2} - \frac{g_\ell}{1-\theta_2} = 0 && \Rightarrow && \theta_2 = \frac{r_\ell}{r_\ell+g_\ell}
\end{aligned}$$

Proportion is equal to the "observed" proportion

Another example

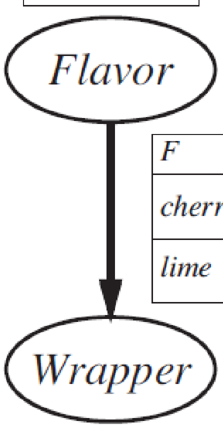
- Data: Unwrapping **N** candies, of which **c** are cherries

$$\boxed{P(\mathbf{d} \mid h_{\theta,\theta_1,\theta_2}) = \theta^c (1-\theta)^\ell \cdot \theta_1^{r_c}(1-\theta_1)^{g_c} \cdot \theta_2^{r_\ell}(1-\theta_2)^{g_\ell}}$$

$$\boxed{L = [c \log \theta + \ell \log(1-\theta)] + [r_c \log \theta_1 + g_c \log(1-\theta_1)] + [r_\ell \log \theta_2 + g_\ell \log(1-\theta_2)]}$$

Learning is **compositional**, one for each parameter

Same as before – obvious, but also comforting



P(F=cherry)
θ

F	P(W=red | F)
cherry	θ_1
lime	θ_2

$$\theta = \frac{c}{c+\ell}$$

$$\theta_1 = \frac{r_c}{r_c+g_c}$$

$$\theta_2 = \frac{r_\ell}{r_\ell+g_\ell}$$

Proportion is equal to the "observed" proportion

Outline of today’s lecture

- Statistical learning
- Maximum-likelihood parameter learning
- Naïve Bayes models

Recap: *Naïve Bayes model*

- A single **cause** directly influence a number of **effects**, all of which are conditionally independent, given the cause

$$\boxed{\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i \mid Cause)}$$

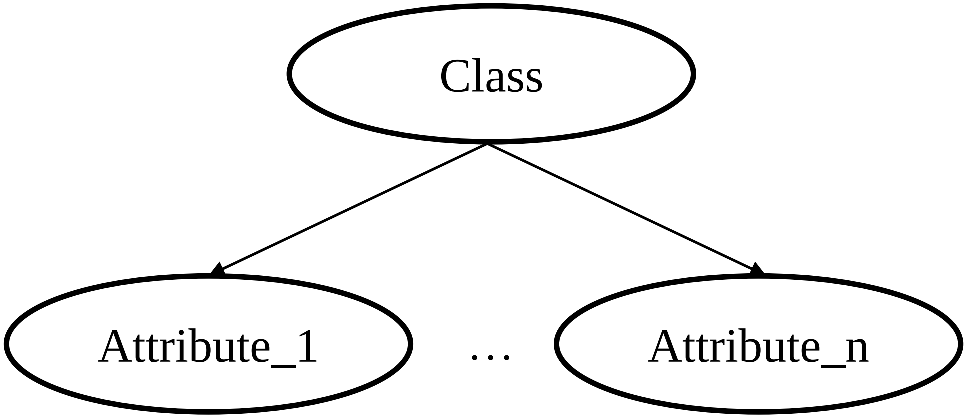
$$\begin{aligned}
&\mathbf{P}(Toothache, Catch, Cavity) \\
&\quad = \mathbf{P}(Toothache, Catch \mid Cavity)\mathbf{P}(Cavity) \\
&\quad = \mathbf{P}(Toothache \mid Cavity)\mathbf{P}(Catch \mid Cavity)\mathbf{P}(Cavity)
\end{aligned}$$

Naïve Bayesian learning

- **Assumption:** Attributes ($X_1,\ldots,X_n$) are conditionally independent of each other, given the class (C)

$$\boxed{\mathbf{P}(C \mid x_1, \ldots, x_n) = \alpha\, \mathbf{P}(C) \prod_i \mathbf{P}(x_i \mid C)}$$

- Using **maximum-likelihood estimates** to learn CDTs; that is, using “frequencies” to compute the “probabilities”



Naïve Bayesian learning

- Properties
  - Tolerant of noise in **attribute** and **class** values of examples
  - Learn quickly, even for large problems

- Early application
  - Email spam detector, where
    - *Attribute_i = "how often does the i-th word in a dictionary appear in the email?"*
    - *Class = "is this email spam?"*

How to use the learned model?

- **Assumption:** Attributes ($X_1,\dots,X_n$) are conditionally independent of each other, given the class (C)

$$\boxed{\mathbf{P}(C \mid x_1,\ldots,x_n) = \alpha\, \mathbf{P}(C) \prod_i \mathbf{P}(x_i \mid C)}$$

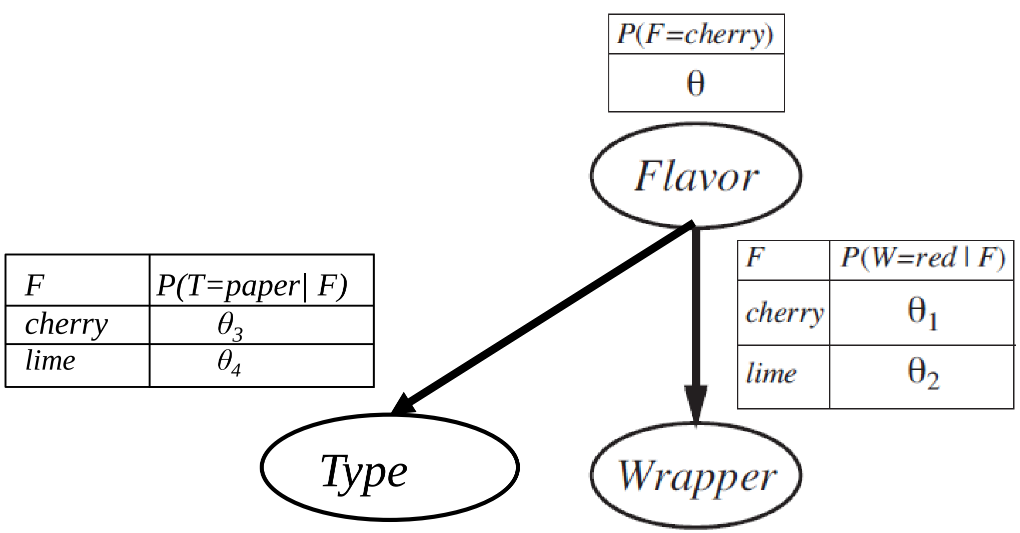
With observed attribute values $x_1,\dots,x_n$, what's the probability of each class C?

How to use the learned model?

- **Assumption:** Attributes ($X_1,\ldots,X_n$) are conditionally independent of each other, given the class (C)

$$\boxed{\mathbf{P}(C \mid x_1, \ldots, x_n) = \alpha\, \mathbf{P}(C) \prod_i \mathbf{P}(x_i \mid C)}$$

With observed attribute values $x_1,\ldots,x_n$, what's the probability of each class C?



The **wrapper** color may be "red" or "green"

The material **type** may be "paper" or "plastic"

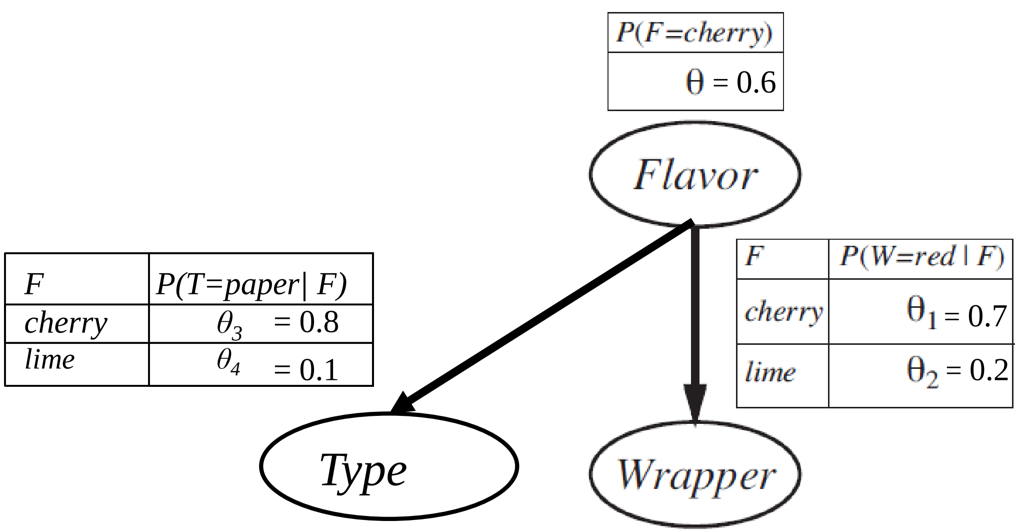
Example

- **Assumption:** Attributes ($X_1,\ldots,X_n$) are conditionally independent of each other, given the class (C)

$$\mathbf{P}(C \mid x_1, \ldots, x_n) = \alpha\, \mathbf{P}(C) \prod_i \mathbf{P}(x_i \mid C)$$

P(F = cherry | W= red, T= paper) = ?

P(F = lime | W= red, T= paper) = ?



The **wrapper** color may be "red" or "green"

The material **type** may be "paper" or "plastic"

Example

- **Assumption:** Attributes ($X_1,\ldots,X_n$) are conditionally independent of each other, given the class (C)

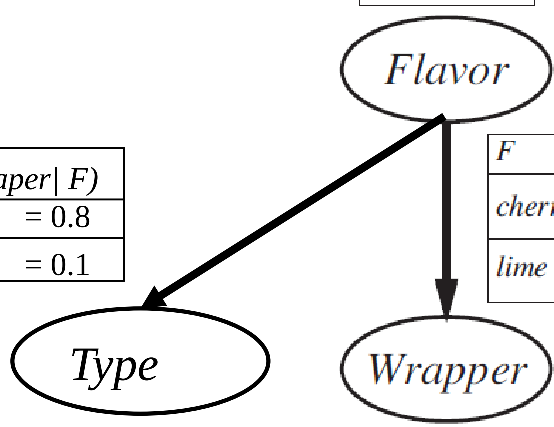
$$\boxed{\mathbf{P}(C \mid x_1,\ldots,x_n) = \alpha\, \mathbf{P}(C) \prod_i \mathbf{P}(x_i \mid C)}$$

P(F = cherry | W= red, T= paper) $= \alpha * P(cherry) * P(red|\ cherry) * P(paper|\ cherry)$

$= \alpha * 0.6 * 0.7 * 0.8 = \alpha * 0.336$

P(F = lime | W= red, T= paper) = ?

$P(F{=}cherry)$
$\theta = 0.6$



F	$P(T{=}paper \mid F)$
cherry	$\theta_3 = 0.8$
lime	$\theta_4 = 0.1$

F	$P(W{=}red \mid F)$
cherry	$\theta_1 = 0.7$
lime	$\theta_2 = 0.2$

The **wrapper** color may be "red" or "green"

The material **type** may be "paper" or "plastic"

Example

- **Assumption:** Attributes ($X_1,\ldots,X_n$) are conditionally independent of each other, given the class (C)

$$\mathbf{P}(C \mid x_1, \ldots, x_n) = \alpha\, \mathbf{P}(C) \prod_i \mathbf{P}(x_i \mid C)$$

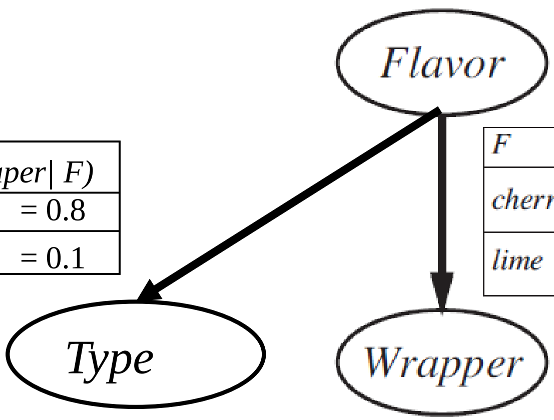
P(F = cherry | W= red, T= paper) $= \alpha * P(cherry) * P(red| cherry) * P(paper| cherry)$

$= \alpha * 0.6 * 0.7 * 0.8 = \alpha * 0.336$

P(F = lime | W= red, T= paper) $= \alpha * P(lime) * P(red| lime) * P(paper| lime)$

$= \alpha * 0.4 * 0.2 * 0.1 = \alpha * 0.008$

P(F=cherry)
$\theta = 0.6$



F	P(T=paper| F)
cherry	$\theta_3 = 0.8$
lime	$\theta_4 = 0.1$

F	P(W=red | F)
cherry	$\theta_1 = 0.7$
lime	$\theta_2 = 0.2$

The **wrapper** color may be "red" or "green"

The material **type** may be "paper" or "plastic"

Example

- **Assumption:** Attributes ($X_1,\ldots,X_n$) are conditionally independent of each other, given the class (C)

$$\boxed{\mathbf{P}(C \mid x_1,\ldots,x_n) = \alpha\, \mathbf{P}(C) \prod_i \mathbf{P}(x_i \mid C)}$$

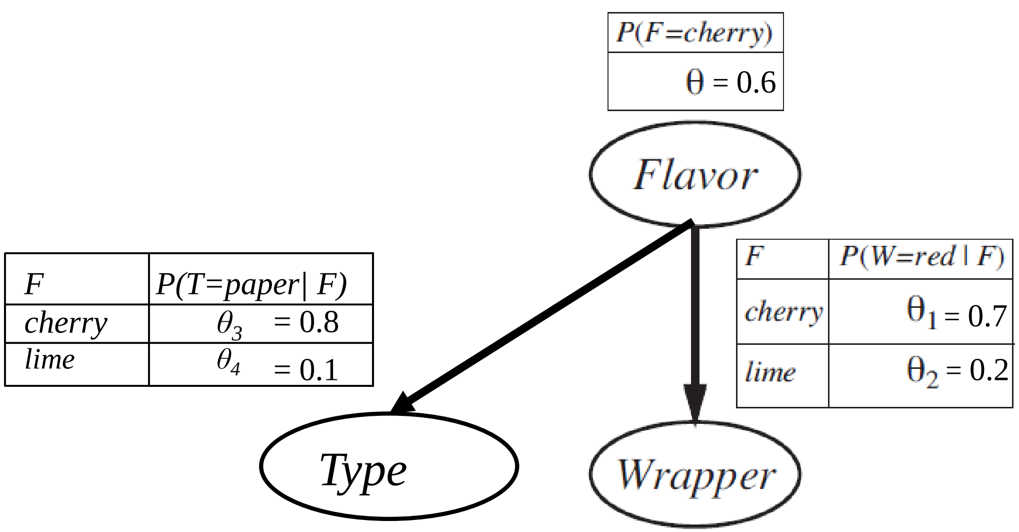
P(F = cherry | W= red, T= paper) = α * P(cherry) * P(red| cherry) * P(paper| cherry)

= α* 0.6 * 0.7 * 0.8 = α* 0.336

P(F = lime | W= red, T= paper) = α * P(lime) * P(red| lime) * P(paper| lime)

= α* 0.4 * 0.2 *0.1 = α* 0.008

α = 1 / (0.336+0.008) = 1/0.344



P(F=cherry)
$\theta = 0.6$

F	P(T=paper| F)
cherry	$\theta_3 = 0.8$
lime	$\theta_4 = 0.1$

F	P(W=red | F)
cherry	$\theta_1 = 0.7$
lime	$\theta_2 = 0.2$

The **wrapper** color may be "red" or "green"

The material **type** may be "paper" or "plastic"

Example

- **Assumption:** Attributes ($X_1,\ldots,X_n$) are conditionally independent of each other, given the class (C)

$$\boxed{\mathbf{P}(C \mid x_1, \ldots, x_n) = \alpha\, \mathbf{P}(C) \prod_i \mathbf{P}(x_i \mid C)}$$

P(F = cherry | W= red, T= paper) $= \alpha * P(cherry) * P(red \mid cherry) * P(paper \mid cherry)$

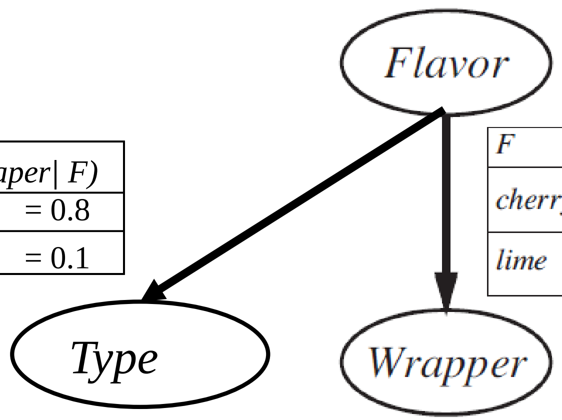
$= \alpha * 0.6 * 0.7 * 0.8 = \alpha * 0.336$ = 0.977 (**97.7%**)

P(F = lime | W= red, T= paper) $= \alpha * P(lime) * P(red \mid lime) * P(paper \mid lime)$

$= \alpha * 0.4 * 0.2 * 0.1 = \alpha * 0.008$ = 0.023 (**2.3%**)

$\alpha = 1 / (0.336+0.008) = 1/0.344$

P(F=cherry)
$\theta = 0.6$



F	P(T=paper| F)
cherry	$\theta_3 = 0.8$
lime	$\theta_4 = 0.1$

F	P(W=red | F)
cherry	$\theta_1 = 0.7$
lime	$\theta_2 = 0.2$

The **wrapper** color may be "red" or "green"

The material **type** may be "paper" or "plastic"

Quiz 13

- **Assumption:** Attributes ($X_1,\ldots,X_n$) are conditionally independent of each other, given the class (C)

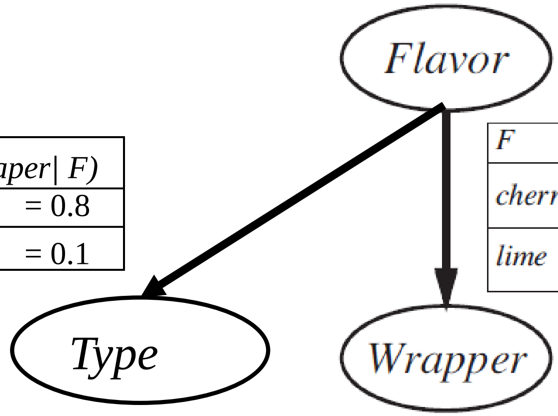
$$\boxed{\mathbf{P}(C \mid x_1, \ldots, x_n) = \alpha\, \mathbf{P}(C) \prod_i \mathbf{P}(x_i \mid C)}$$

With the observed wrapper being “green” and “plastic”, what is the probability of the candy being “cherry” and “lime”, respectively?

P(F = cherry | W= green, T= plastic) = ?

P(F = lime | W= green, T= plastic) = ?

P(F=cherry)
$\theta = 0.6$



F	P(T=paper| F)
cherry	θ_3 = 0.8
lime	θ_4 = 0.1

F	P(W=red | F)
cherry	$\theta_1 = 0.7$
lime	$\theta_2 = 0.2$

The **wrapper** color may be “red” or “green”

The material **type** may be “paper” or “plastic”