Introduction to Computational Intelligence

Probabilistic modeling and learning

Igor Farkaš
Centre for Cognitive Science
Comenius University Bratislava

(Russell & Norvig: Artificial Intelligence (3rd ed.), Prentice Hall, 2010)

Introduction

- Uncertainty: A state of having limited knowledge where it is impossible to exactly describe the existing state, a future outcome, or more than one possible outcome.
- Ubiquitous in the world

- To deal with uncertainty, agents must keep track of belief states.
- Probability is the measure of the likeliness that an event will occur.
- Probabilistic approach is alternative to logical approach.

Basics of probability theory

- In probability theory, the set of all possible worlds (ω) is called the sample space (Ω).
- The possible worlds are mutually exclusive and exhaustive
- E.g. if we are about to roll two (different) dice, there are 36 possible worlds to consider: (1,1), (1,2), ..., (6,6).
- A fully specified probability model associates a numerical probability *P*(ω) with each possible world. It holds that:

$$0 \leqslant P(\omega) \leqslant 1 \quad \text{for every } \omega \quad \text{and} \quad \sum_{\omega \in \Omega} P(\omega) = 1$$

- e.g. for fair dice above, the probability of each world is 1/36
- Often we are interested not in particular ω, but the sets of them

Probability theory basics (ctd)

- Events, described by propositions (in AI)
- e.g. *P*(*odd*), *P*(*doubles*), *P*(*total* = 9), etc…
- Types of probabilities (w.r.t. evidence):
  - prior (unconditional), e.g. *P*(ω<4)
  - posterior (conditional), e.g. *P*(*doubles*|die_1=5)
- Definition: for events *a*, *b*

$$P(a|b)=\frac{P(a\wedge b)}{P(b)} \qquad P(b|a)=\frac{P(b\wedge a)}{P(a)}$$

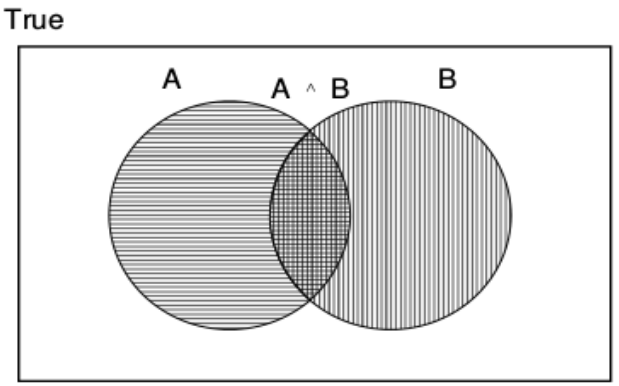
- Bayes' rule:

$$P(a|b)=\frac{P(b|a)P(a)}{P(b)}$$

Probability (ctd)

- Inclusion–exclusion principle:

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$



- Where do probabilities come from?
- Frequentist view – numbers can come only from experiments, i.e. based on empirical evidence.
- Objectivist view – probabilities are real aspects of the universe – propensities of objects to behave in certain ways, rather than being just descriptions of an observer's degree of belief.
- Subjectivist view – probabilities characterize agent's beliefs, rather than have any external physical significance.

Example: cavity-catch-toothache world

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

- Joint probability distribution (in 2×2×2 table) – provides probability of each atomic event
- Allows probabilistic inference (calculating arbitrary probs)
- Prior probability, e.g. *P*(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2
- Conditional probability, e.g. *P*(cavity | toothache) = (0.108 + 0.012) / (0.108 + 0.012 + 0.016 + 0.064) = 0.6

$$P(Effect \mid Cause) = \frac{P(Cause \mid Effect)\, P(Effect)}{P(Cause)}$$

Conditional independence

- toothache and catch are not independent…
- … but one does not imply the other
- conditioned on cavity, they are independent:

  P(toothache,catch | cavity) = P(toothache | cavity) . P(catch | cavity)
- Conditional independence (CI) allows problem simplification
- Cavity separates toothache and catch because it is a direct cause of both of them
- CI assumpion in general: $P(X,Y|Z) = P(X|Z) \cdot P(Y|Z)$
- Naive Bayes model (used also when CI does not hold):

$$P(Cause, Effect_1, \ldots, Effect_n) = P(Cause) \prod_i P(Effect_i | Cause)$$

Naive Bayes classifier

- The “class” variable *C* (which is to be predicted) is the root and the “attribute” variables x_i are the leaves.
- With observed attribute values $x_1, \ldots, x_n$, the probability of each class is given by

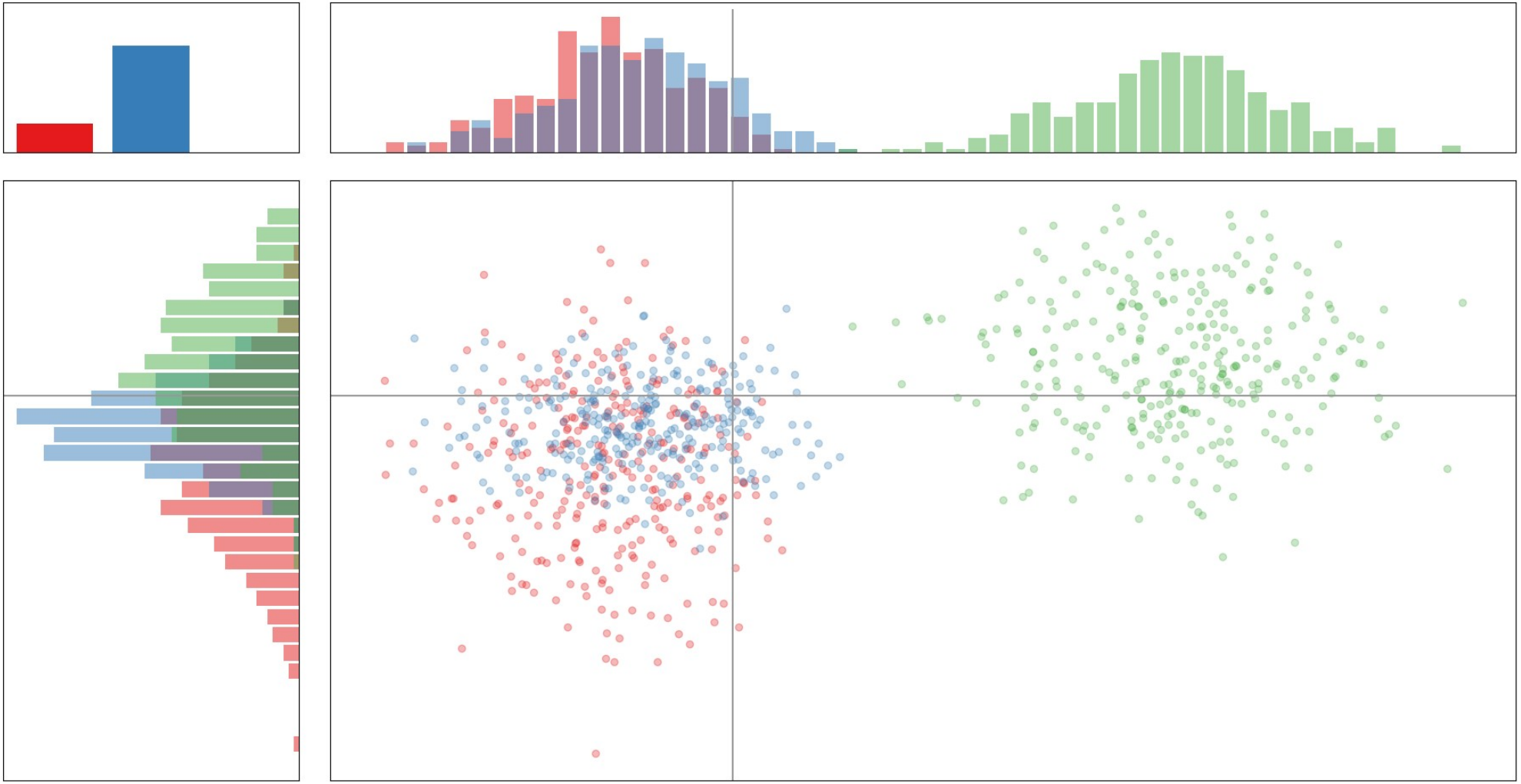
$$P(C_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|C_i)\,P(C_i)}{P(\boldsymbol{x})}$$

- under the assumption of independent attributes x_i

$$P(C|x_1,\ldots,x_n) = \alpha\, P(C) \prod_i P(x_i|C)$$

$$P(C_i|\boldsymbol{x}) = \frac{P(x_1|C_i)\,P(x_2|C_i)\ldots P(x_n|C_i)\,P(C_i)}{P(x_1)\,P(x_2)\ldots P(x_n)}$$

Naive Bayes classifier in 2D



Full Bayesian learning

View learning as Bayesian updating of a probability distribution over the hypothesis space

H is the hypothesis variable, values $h_1, h_2, \ldots$, prior $\mathbf{P}(H)$

jth observation d_j gives the outcome of random variable D_j
training data $\mathbf{d} = d_1, \ldots, d_N$

Given the data so far, each hypothesis has a posterior probability:

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

where $P(\mathbf{d}|h_i)$ is called the likelihood

Predictions use a likelihood-weighted average over the hypotheses:

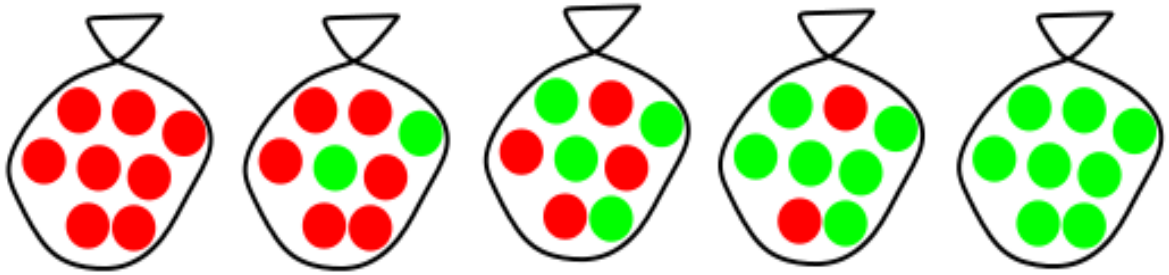
$$\mathbf{P}(X|\mathbf{d}) = \Sigma_i \, \mathbf{P}(X|\mathbf{d}, h_i)P(h_i|\mathbf{d}) = \Sigma_i \, \mathbf{P}(X|h_i)P(h_i|\mathbf{d})$$

No need to pick one best-guess hypothesis!

Example of probabilistic learning

Suppose there are five kinds of bags of candies:

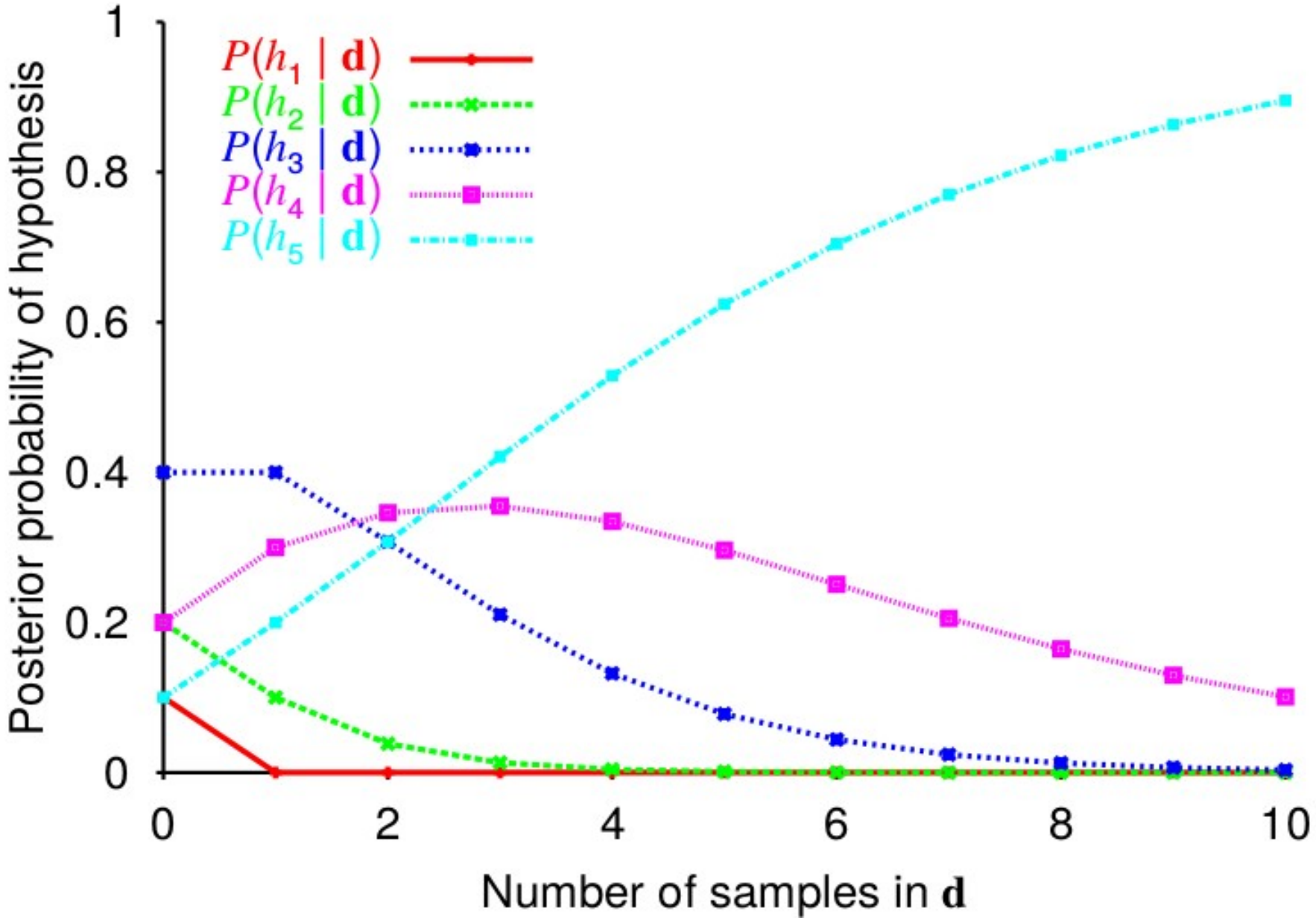
10% are h_1: 100% cherry candies
20% are h_2: 75% cherry candies + 25% lime candies
40% are h_3: 50% cherry candies + 50% lime candies
20% are h_4: 25% cherry candies + 75% lime candies
10% are h_5: 100% lime candies



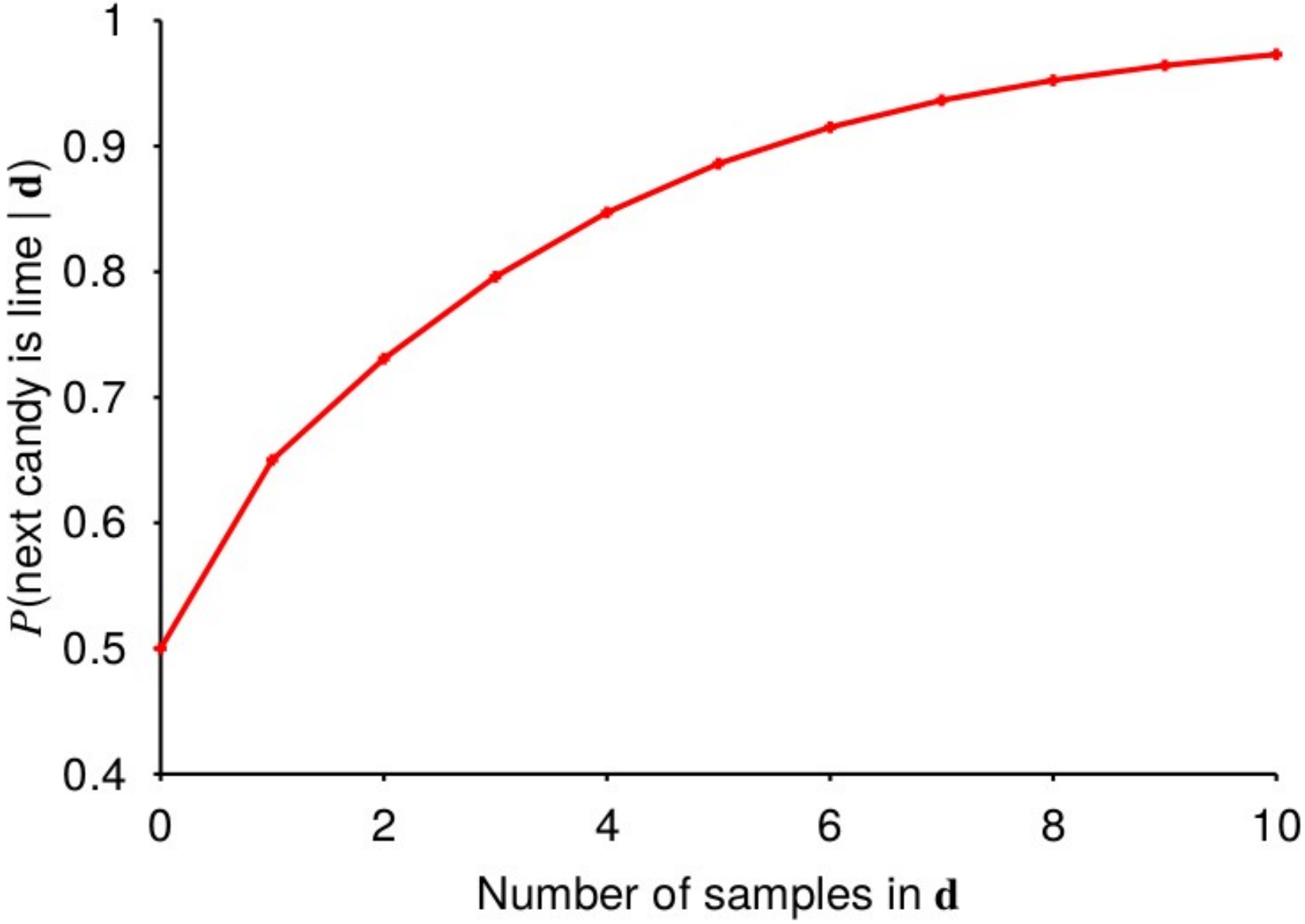
Then we observe candies drawn from some bag: 

What kind of bag is it? What flavour will the next candy be?

Posterior probability of hypotheses



Example of prediction probability

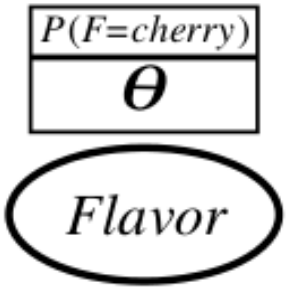


MAP and ML approximation

- summing over the hypothesis space is often intractable
- Maximum a posteriori (MAP) learning: choose h_{MAP} maximizing $P(h_i|\boldsymbol{d})$
- i.e. maximize $P(\boldsymbol{d}|h_i)$. $P(h_i)$ or $\log P(\boldsymbol{d}|h_i) + \log P(h_i)$
- Log terms can be viewed as (negatives of)
- bits to encode data given hypothesis + bits to encode hypothesis
- This is the basic idea of minimum description length learning
- For large data sets, we can ignore $P(h_i)$ => empiricist
- Maximum likelihood (ML) learning: choose h_{ML} maximizing $P(\boldsymbol{d}|h_i)$
- ML is the standard (non-Bayesian) statistical learning method

ML parameter learning in Bayes nets

Bag from a new manufacturer; fraction θ of cherry candies?
Any θ is possible: continuum of hypotheses h_θ
θ is a parameter for this simple (binomial) family of models



Suppose we unwrap N candies, c cherries and $\ell = N - c$ limes
These are i.i.d. (independent, identically distributed) observations, so

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

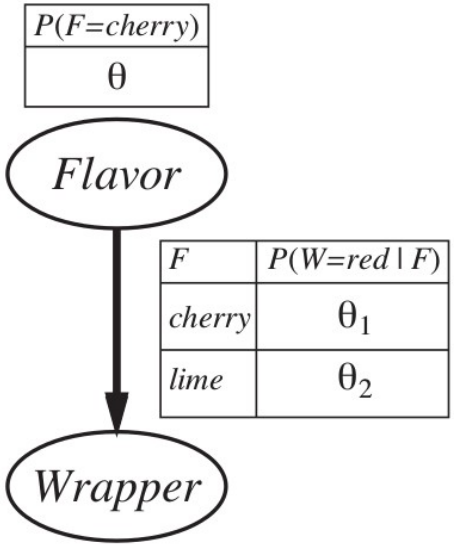
Maximize this w.r.t. θ—which is easier for the log-likelihood:

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

$$\frac{dL(\mathbf{d}|h_\theta)}{d\theta} = \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 \qquad \Rightarrow \qquad \theta = \frac{c}{c+\ell} = \frac{c}{N}$$

A more complicated case

- The wrapper color depends (probabilistically) on the candy flavor
- Let unwrap N candies, of which c are cherries and l are limes. Let r_c (g_c) of the cherries have red (green) wrappers, while r_l (g_l) of the limes have red (green). Then



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

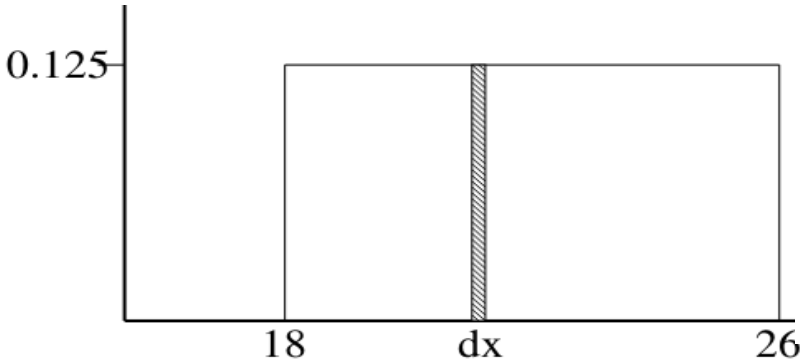
- ML parameter learning problem for a Bayesian network decomposes into separate learning problems, one for each parameter.

$$\theta = \frac{c}{c+\ell}$$
$$\theta_1 = \frac{r_c}{r_c+g_c}$$
$$\theta_2 = \frac{r_\ell}{r_\ell+g_\ell}$$

Probability for continuous variables

Express distribution as a parameterized function of value:

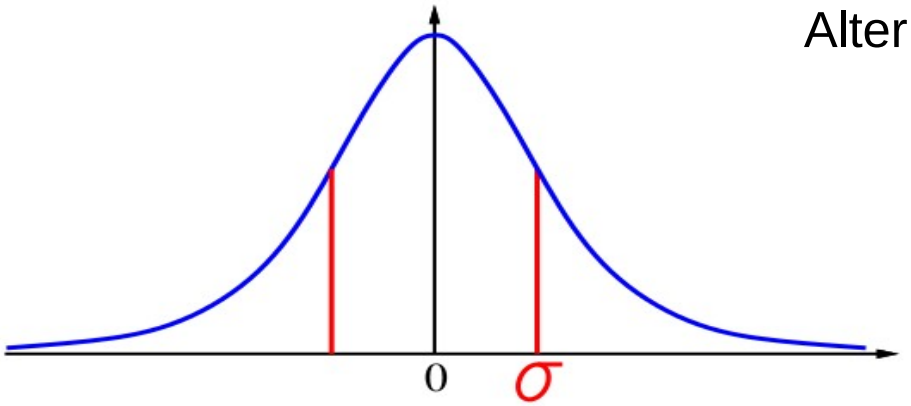
$P(X = x) = U[18, 26](x)$ = uniform density between 18 and 26



Here P is a density; integrates to 1.

$P(X = 20.5) = 0.125$ really means

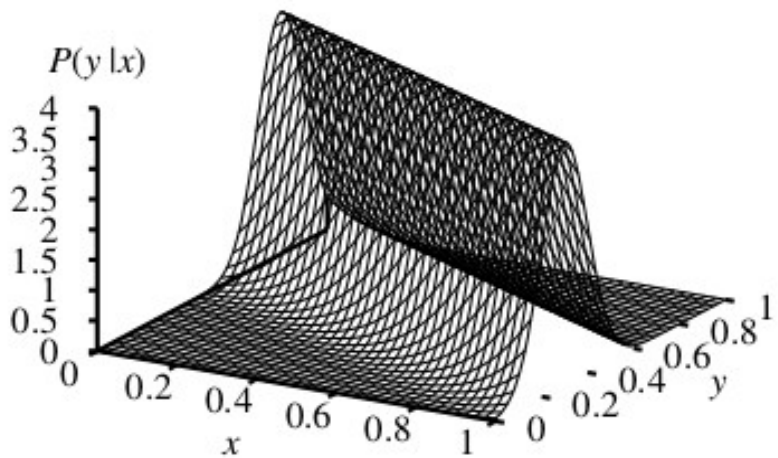
$$\lim_{dx \to 0} P(20.5 \leq X \leq 20.5 + dx)/dx = 0.125$$

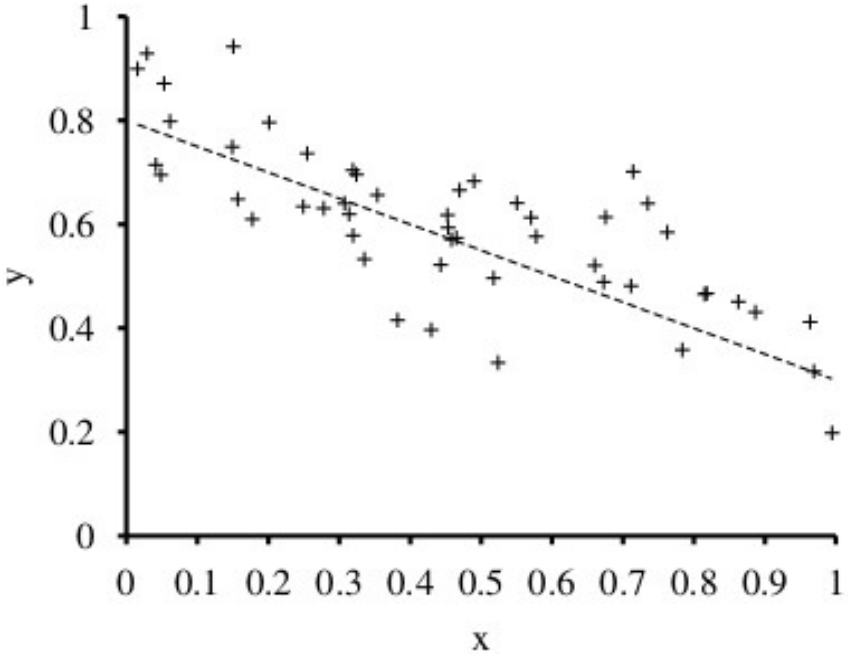


Alternative: Gaussian density

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

Example: Linear Gaussian model





Maximizing $P(y|x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-(\theta_1 x+\theta_2))^2}{2\sigma^2}}$ w.r.t. θ_1, θ_2

$=$ minimizing $E = \sum_{j=1}^{N} (y_j - (\theta_1 x_j + \theta_2))^2$

That is, minimizing the sum of squared errors gives the ML solution for a linear fit **assuming Gaussian noise of fixed variance**

Summary

- Probability is a rigorous formalism for uncertain knowledge
- Joint probability distribution specifies probability of every atomic event
- Queries can be answered by summing over atomic events
- For nontrivial domains, we must find a way to reduce the joint size
- Independence and conditional in dependence provide the tools (naive Bayes model).
- Probabilistic models can be learned from evidence
- Approximations of Bayesian learning useful (MAP, ML)