Probability and Naïve Bayes

Source: S. Russell, P. Norvig and D. Jurafsky. & J. Martin

Probability

Probabilistic assertions summarize effects of

- laziness: failure to enumerate exceptions, qualifications, etc.
- ignorance: lack of relevant facts, initial conditions, etc.

Probability theory: uses uncertainty to anticipate future events

Bayesian probability:

- Probabilities relate propositions to agent's own state of knowledge

  e.g., $P(A_{25} \mid \text{no reported accidents}) = 0.06$

These are not assertions about the world

Probabilities of propositions change with new evidence:

e.g., $P(A_{25} \mid \text{no reported accidents, 5 a.m.}) = 0.15$

Prior probability

Prior or unconditional probabilities of propositions

e.g., P(Price Up = true) = 0.1 and P(*Analyst rec.* = Buy) = 0.72 correspond to belief prior to arrival of any (new) evidence

Probability distribution gives values for all possible assignments:

P(*Analyst's recommendations*) = <0.72,0.1,0.08,0.1> (normalized, i.e., sums to 1)

Joint probability distribution for a set of random variables gives the probability of every atomic event on those random variables

P(*Analyst's recommendations, Price Up*) = a 2 x 4 matrix of values:

Analysts' rec. =	Buy	Sell	Hold	Unknown
Price Up = true	0.144	0.02	0.016	0.02
Price Up = false	0.576	0.08	0.064	0.08

Every question about a domain can be answered by the joint distribution

Moving toward language

- What's the probability of drawing a 2 from a deck of 52 cards with four 2s?

$$P(drawing\ \ a\ \ two) = \frac{4}{52} = \frac{1}{13} = .077$$

- What's the probability of a random word (from a random dictionary page) being a verb?

$$P(drawing\ \ a\ \ verb) = \frac{\#of\ \ ways\ \ to\ \ get\ \ a\ \ verb}{all\ \ words}$$

Probability and part of speech tags

- What's the probability of a random word (from a random dictionary page) being a verb?

$$P(drawing\ \ a\ \ verb) = \frac{\#of\ \ ways\ \ to\ \ get\ \ a\ \ verb}{all\ \ words}$$

- How to compute each of these
- All words = just count all the words in the dictionary
- # of ways to get a verb: number of words which are verbs!
- If a dictionary has 50,000 entries, and 10,000 are verbs….
- P(V) is 10000/50000 = 1/5 = .20

Conditional probability

Conditional or posterior probabilities

e.g., P(*Price Up* | *Buy recommendation*) = 0.8

i.e., given that *Buy recommendation* is all I know

(Notation for conditional distributions:

P(*Price Up* | *Buy recommendation*) = 2-element vector of 2-element vectors)

If we know more, e.g., *Price Up* is also given, then we have

P(*Price Up* | *Buy recommendation, Address*) = 1

New evidence may be irrelevant, allowing simplification, e.g.,

P(*Price Up* | *Buy recommendation, Address*) = P(*Price Up* | *Buy recommendation*) = 0.8

This kind of inference, sanctioned by domain knowledge, is crucial

Conditional probability

Definition of conditional probability:

$P(a \mid b) = P(a \wedge b) / P(b)$ if $P(b) > 0$

Product rule gives an alternative formulation:

$P(a \wedge b) = P(a \mid b) P(b) = P(b \mid a) P(a)$

A general version holds for whole distributions, e.g.,

$\mathbf{P}$(Price up, Buy Recommendation) = $\mathbf{P}$(Price up | Buy Recommendation) $\mathbf{P}$(Buy Rec.)
(View as a set of 4 × 2 equations, not matrix mult.)

Chain rule is derived by successive application of product rule:

$$
\begin{aligned}
\mathbf{P}(X_1, \ldots, X_n) &= \mathbf{P}(X_1,\ldots,X_{n-1})\, \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1}) \\
&= \mathbf{P}(X_1,\ldots,X_{n-2})\, \mathbf{P}(X_{n-1} \mid X_1,\ldots,X_{n-2})\, \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1}) \\
&= \ldots \\
&= \pi_{i=1}\text{^}n\, \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})
\end{aligned}
$$

Inference by enumeration

Start with the joint probability distribution:

	Buy recommendation		¬ Buy recommendation	
	US	¬ US	US	¬ US
Price up	0.108	0.012	0.072	0.008
¬ Price up	0.016	0.064	0.144	0.576

Evaluate if stock price will go up when the analysts' consensus is "buy" and assets' residence (US or international)

For any proposition φ, sum the atomic events where it is true: $P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega)$

Inference by enumeration

Start with the joint probability distribution:

	Buy recommendation		¬ Buy recommendation	
	US	¬ US	US	¬ US
Price up	0.108	0.012	0.072	0.008
¬ Price up	0.016	0.064	0.144	0.576

For any proposition φ, sum the atomic events where it is true: $P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega)$

P(*Buy recommendation*) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2

Inference by enumeration

Start with the joint probability distribution:

	Buy recommendation		¬ Buy recommendation	
	US	¬ US	US	¬ US
Price up	0.108	0.012	0.072	0.008
¬ Price up	0.016	0.064	0.144	0.576

Can also compute conditional probabilities:

$$P(\text{Price up} \mid \text{Buy Rec.}) = \frac{P(\text{Price up} \wedge \text{Buy Rec.})}{P(\text{Buy Rec.})}$$

$$= \frac{0.108+0.012}{0.108 + 0.012 + 0.016 + 0.064}$$

$$= 0.6$$

Normalization

	Buy recommendation		¬ Buy recommendation	
	US	¬ US	US	¬ US
Price up	0.108	0.012	0.072	0.008
¬ Price up	0.016	0.064	0.144	0.576

- Denominator can be viewed as a normalization constant α

P (Price up | Buy Rec.) = α, **P**(Price up, Buy Rec.)
 = α, [**P**(Price up, Buy Rec.,US) + **P**(Price up, Buy Rec., ¬US)]
 = α, [<0.108,0.016> + <0.012,0.064>]
 = α, <0.12,0.08> = <0.6,0.4> (probabilities Price up or ¬Price up)

P: evaluates T and F probabilities of query variable (Price up)

General idea: compute distribution on query variable (Price up) by fixing evidence variables (Buy rec.) and summing over hidden variables (US)

Inference by enumeration, contd.

Typically, we are interested in the posterior joint distribution of the query variables **Y** given specific values **e** for the evidence variables **E**

Let the hidden variables be **H** = **X** - **Y** - **E**

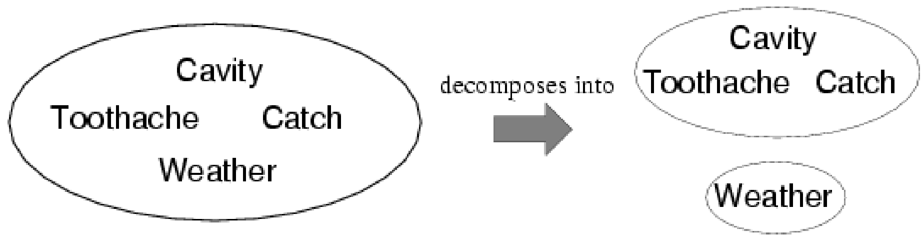
Then the required summation of joint entries is done by summing out the hidden variables:

$$\mathbf{P}(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \alpha \mathbf{P}(\mathbf{Y},\mathbf{E} = \mathbf{e}) = \alpha \Sigma_h \mathbf{P}(\mathbf{Y},\mathbf{E}= \mathbf{e}, \mathbf{H} = \mathbf{h})$$

- The terms in the summation are joint entries because **Y**, **E** and **H** together exhaust the set of random variables
- Obvious problems:
  1. Worst-case time complexity $O(d^n)$ where d is the largest arity
  2. Space complexity $O(d^n)$ to store the joint distribution
  3. How to find the numbers for $O(d^n)$ entries?

Independence

- *A* and *B* are independent iff
  $\mathbf{P}(A|B) = \mathbf{P}(A)$ or $\mathbf{P}(B|A) = \mathbf{P}(B)$ or $\mathbf{P}(\mathrm{A, B}) = \mathbf{P}(A)\,\mathbf{P}(B)$



$\mathbf{P}(\textit{Toothache, Catch, Cavity, Weather}) = \mathbf{P}(\textit{Toothache, Catch, Cavity})\mathbf{P}(\textit{Weather})$

- 32 entries reduced to 12; for *n* independent biased coins, $O(2^n) \rightarrow O(n)$
- Absolute independence powerful but rare
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Conditional independence

$\mathbf{P}(Toothache, Cavity, Catch)$ has $2^3 - 1 = 7$ independent entries

If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:

(1) $\mathbf{P}(catch \mid toothache, cavity) = \mathbf{P}(catch \mid cavity)$

The same independence holds if I haven't got a cavity:

(2) $\mathbf{P}(catch \mid toothache, \neg cavity) = \mathbf{P}(catch \mid \neg cavity)$

Catch is conditionally independent of *Toothache* given *Cavity*:

$\mathbf{P}(Catch \mid Toothache, Cavity) = \mathbf{P}(Catch \mid Cavity)$

Equivalent statements:

$\mathbf{P}(Toothache \mid Catch, Cavity) = \mathbf{P}(Toothache \mid Cavity)$

$\mathbf{P}(Toothache, Catch \mid Cavity) = \mathbf{P}(Toothache \mid Cavity)\,\mathbf{P}(Catch \mid Cavity)$

Conditional independence contd.

- Write out full joint distribution using chain rule:

  P(*Toothache, Catch, Cavity*)
  = **P**(*Toothache* | *Catch, Cavity*) **P**(*Catch, Cavity*)

  = **P**(*Toothache* | *Catch, Cavity*) **P**(*Catch* | *Cavity*) **P**(*Cavity*)

  = **P**(*Toothache* | *Cavity*) **P**(*Catch* | *Cavity*) **P**(Cavity)

  I.e., 2 + 2 + 1 = 5 independent numbers

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in *n* to linear in *n*.

- Conditional independence is our most basic and robust form of knowledge about uncertain environments.

Bayes' Rule

Product rule $P(a \wedge b) = P(a \mid b) P(b) = P(b \mid a) P(a)$

$\Rightarrow$ Bayes' rule: $P(a \mid b) = P(b \mid a) P(a) / P(b)$

or in distribution form

$$\mathbf{P}(Y|X) = \mathbf{P}(X|Y)\,\mathbf{P}(Y) / \mathbf{P}(X) = \alpha\mathbf{P}(X|Y)\,\mathbf{P}(Y)$$

Useful for assessing diagnostic probability from causal probability:

P(Cause|Effect) = P(Effect|Cause) P(Cause) / P(Effect)

E.g., let *M* be money supply increase, *S* be silver price increase:

$P(m|s) = P(s|m)\,P(m) / P(s) = 0.8 \times 0.0001 / 0.1 = 0.0008$

Note: posterior probability of money supply increase still very small!

Naïve Bayes

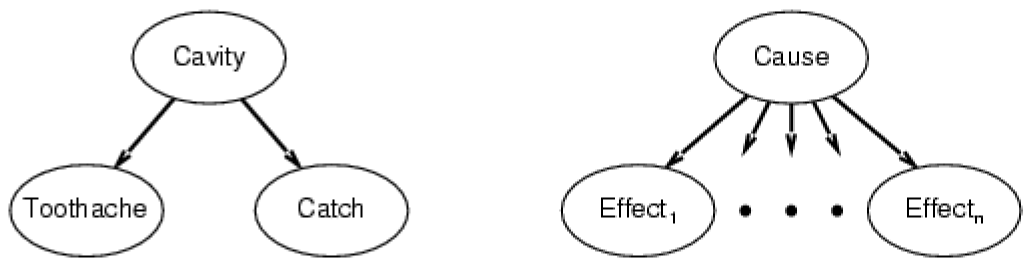
- What happens if we have more than one piece of evidence?
- If we can assume conditional independence, it is easier to solve:
  - Overslept and traffic jam are independent, given late

P(late | overslept Λ traffic jam) =
αP(overslept Λ traffic jam | late)P(late) = αP(overslept | late)P(traffic jam | late)P(late)

Naïve Bayes where a single cause directly influences a number of effects, all conditionally independent

- Independence often assumed even when not so

- This is an example of a naïve Bayes model:
  $\mathbf{P}(\text{Cause}, \text{Effect}_1, \ldots, \text{Effect}_n) = \mathbf{P}(\text{Cause})\, \pi_i \mathbf{P}(\text{Effect}_i|\text{Cause})$



- Total number of parameters is linear in *n*

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 independence provide the tools