

CS 581 – ADVANCED ARTIFICIAL INTELLIGENCE

TOPIC: PROBABILITY THEORY



Mustafa Bilgic



<http://www.cs.iit.edu/~mbilgic>



<https://twitter.com/bilgicm>

MOTIVATION

- The agent needs reason in an uncertain world
- Uncertainty can be due to
 - Noisy sensors (e.g., temperature, GPS, camera, etc.)
 - Imperfect data (e.g., low resolution image)
 - Missing data (e.g., lab tests)
 - Imperfect knowledge (e.g., medical diagnosis)
 - Exceptions (e.g., all birds fly except ostriches, penguins, birds with injured wings, dead birds, ...)
 - Changing data (e.g., flu seasons, traffic conditions during rush hour, etc.)
 - ...
- The agent still must act (e.g., step on the breaks, diagnose a patient, order a lab test, ...)

TENTATIVE PLAN

- Probability background
- Classification
 - Naïve Bayes, logistic regression, neural networks
 - Maximum likelihood estimation, Bayesian estimation, gradient optimization, backpropagation
- Decision making
 - Bayesian networks, influence diagrams, Markov decision processes, multi-armed bandits
 - Variable elimination, value of information, Bellman equations, value iteration, policy iteration, UCB1, ϵ -greedy
- Reinforcement learning
 - Prediction, control, Monte-Carlo methods, temporal difference learning, Sarsa, Q-learning

SOME EXERCISES

- In a class, 70% of the grad students got an A. John got an A. What is the probability that John is a grad student?
- You design a Covid test with the following behavior
 - $P(+ | covid) = 0.95$; $P(- | covid) = 0.05$
 - $P(+ | \sim covid) = 0.10$; $P(- | \sim covid) = 0.90$
 - John takes the test, and the result is +. What is the probability that John has covid?
- In a town, 60% of the hospitalized are vaccinated. Do the vaccines provide any protection against hospitalization?
- $P(toothache | cavity) = 0.75$. What is $P(cavity | toothache)$?

RANDOM VARIABLES

- Pick variables of interest
 - Medical diagnosis
 - Age, gender, weight, temperature, LT1, LT2, ...
 - Loan application
 - Income, savings, payment history, ...
 - Exercises
 - Grad student, Grade, Covid, Test result, Toothache, Cavity
- Every variable has a domain
 - Binary (True/False)
 - Categorical
 - Real-valued
- Possible world
 - An assignment to all variables of interest

PROBABILITY MODEL

- A **probability model** associates a numerical probability $P(w)$ with each possible world w
 - $P(w)$ sums to 1 over all possible worlds
- An **event** is the set of possible worlds where a given predicate is true
 - Roll two dice
 - The possible worlds are (1,1), (1,2), ..., (6,6); 36 possible worlds
 - Predicate = two dice sum to 10
 - Event = {(4,6), (5,5), (6,4)}
 - Toothache and cavity
 - Four possible worlds: $(t, c), (t, \sim c), (\sim t, c), (\sim t, \sim c)$
 - Some worlds are more likely than others
 - Predicate can be anything about these variables: $t \wedge c, t, t \vee \sim c,$

AXIOMS OF PROBABILITY

1. The probability $P(a)$ of a proposition a is a real number between 0 and 1
2. $P(\text{true}) = 1$, $P(\text{false}) = 0$
3. $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

$P(\neg a)$

- $P(a \vee \neg a) = P(a) + P(\neg a) - P(a \wedge \neg a)$
- $P(\text{true}) = P(a) + P(\neg a) - P(\text{false})$
- $1 = P(a) + P(\neg a) - 0$
- $P(\neg a) = 1 - P(a)$
- Intuitive explanation:
 - The probability of all possible worlds is 1
 - Either a or $\neg a$ holds in one world
 - The worlds that a holds and the worlds that $\neg a$ holds are mutually exclusive and exhaustive

RANDOM VARIABLES – NOTATION

- Capital: X : variable
- Lowercase: x : a particular value of X
- $\text{Val}(X)$: the set of values X can take
- Bold Capital: \mathbf{X} : a set of variables
- Bold lowercase: \mathbf{x} : an assignment to all variables in \mathbf{X}
- $P(X=x)$ will be shortened as $P(x)$
- $P(X=x \cap Y=y)$ will be shortened as $P(x,y)$

JOINT DISTRIBUTION

- We have n random variables, V_1, V_2, \dots, V_n
- We are interested in the probability of a possible world, where
 - $V_1=v_1, V_2=v_2, \dots, V_n = v_n$
- $P(V_1, V_2, \dots, V_n)$ associates a probability for each possible world \equiv the **joint distribution**
- How many entries are there, if we assume the variables are all binary?

2^n

TOOTHACHE EXAMPLE

Joint $P(T, C)$

Toothache	Cavity	$P(T, C)$
toothache	cavity	0.15
toothache	\neg cavity	0.10
\neg toothache	cavity	0.05
\neg toothache	\neg cavity	0.70

PRIOR AND POSTERIOR

- Prior probability
 - Probability of a proposition in the absence of any other information
 - E.g., $P(V_1, V_3, V_5)$
- Conditional/posterior probability
 - Probability of a proposition given another piece of information
 - E.g., $P(V_2, V_3 \mid V_5 = T, V_7 = F)$
 - $P(A \mid B) = P(A \wedge B) / P(B)$

MARGINALIZATION

- Given $P(V_1, V_2, \dots, V_n \mid V_{n+1}, V_{n+2}, \dots, V_{n+m})$, where $n > 0$ and $m \geq 0$, we can find, for example
 - $P(V_i, V_j, V_k \mid V_{n+1}, V_{n+2}, \dots, V_{n+m})$ where $i, j, k < n$ by summing out all the irrelevant variables
- Examples

LET'S ANSWER A FEW QUERIES

Toothache	Cavity	P(T,C)
toothache	cavity	0.15
toothache	\neg cavity	0.10
\neg toothache	cavity	0.05
\neg toothache	\neg cavity	0.70

- $P(\text{cavity}) = ?$ 0.20
- $P(\neg \text{cavity}) = ?$ 0.80
- $P(\text{toothache}) = ?$ 0.25
- $P(\neg \text{toothache}) = ?$ 0.75

LET'S ANSWER A FEW QUERIES

Toothache	Cavity	P(T,C)
toothache	cavity	0.15
toothache	¬cavity	0.10
¬toothache	cavity	0.05
¬toothache	¬cavity	0.70

- $P(\text{cavity} \mid \text{toothache}) = ?$
- $P(\text{cavity} \mid \neg\text{toothache}) = ?$
- $P(\neg\text{cavity} \mid \text{toothache}) = ?$
- $P(\neg\text{cavity} \mid \neg\text{toothache}) = ?$
- $P(\text{toothache} \mid \text{cavity}) = ?$ ←
- $P(\neg\text{toothache} \mid \text{cavity}) = ?$
- $P(\text{toothache} \mid \neg\text{cavity}) = ?$
- $P(\neg\text{toothache} \mid \neg\text{cavity}) = ?$

$$t|c = \frac{t, c}{c}$$
$$= \frac{0.15}{0.20} - 0.75$$

BAYES' RULE

- $P(B | A) = P(A | B) * P(B) / P(A)$
- Example use
 - $P(\text{cause} | \text{effect}) = P(\text{effect} | \text{cause}) * P(\text{cause}) / P(\text{effect})$
- Why is this useful?
 - Because in practice it is easier to get probabilities for $P(\text{effect} | \text{cause})$ and $P(\text{cause})$ than for $P(\text{cause} | \text{effect})$
 - E.g., $P(\text{disease} | \text{symptoms}) = P(\text{symptoms} | \text{disease}) * P(\text{disease}) / P(\text{symptoms})$
 - It is easier to know what symptoms diseases cause. It is harder to diagnose a disease given symptoms

BAYES RULE

- Can we compute $P(\alpha|\beta)$ from $P(\beta|\alpha)$? *No*

CHAIN RULE

- $P(X_1, X_2, X_3, \dots, X_k) =$
 - $P(X_1) P(X_2 | X_1) P(X_3 | X_1, X_2) \dots P(X_k | X_1, X_2, X_3, \dots, X_{k-1})$
 - or
 - $P(X_2) P(X_1 | X_2) P(X_3 | X_1, X_2) \dots P(X_k | X_1, X_2, X_3, \dots, X_{k-1})$
 - or
 - $P(X_2) P(X_3 | X_2) P(X_1 | X_3, X_2) \dots P(X_k | X_1, X_2, X_3, \dots, X_{k-1})$
 - or
 - Pick an order, then
 - $P(\text{first})P(\text{second} | \text{first})P(\text{third} | \text{first}, \text{second}) \dots P(\text{last} | \text{all_previous})$

MARGINAL INDEPENDENCE

- An event α is **independent** of event β in P , denoted as $P \models \alpha \perp \beta$, if
 - $P(\alpha \mid \beta) = P(\alpha)$, or
 - $P(\beta) = 0$
- Proposition: A distribution P satisfies $\alpha \perp \beta$ if and only if
 - $P(\alpha, \beta) = P(\alpha) P(\beta)$
 - *Can you prove it?*
- Corollary: $\alpha \perp \beta$ implies $\beta \perp \alpha$

MARGINAL INDEPENDENCE

X	Y	P(X, Y)
t	t	0.18
t	f	0.42
f	t	0.12
f	f	0.28

Is $X \perp Y$?

CONDITIONAL INDEPENDENCE

- Two events are independent given another event
- An event α is **independent** of event β given event γ in P , denoted as $P \models (\alpha \perp \beta \mid \gamma)$, if
 - $P(\alpha \mid \beta, \gamma) = P(\alpha \mid \gamma)$, or
 - $P(\beta, \gamma) = 0$
- Proposition: A distribution P satisfies $\alpha \perp \beta \mid \gamma$ if and only if
 - $P(\alpha, \beta \mid \gamma) = P(\alpha \mid \gamma) P(\beta \mid \gamma)$

QUERYING A DISTRIBUTION

- **Evidence ($E=e$):** what is known, **Query (Y):** variables of interest, **X** is the set of all variables that include **E** , **Y** , and potentially others
- 1. **Probability query**
 - $P(Y | e) = ?$
- 2. **MAP query**
 - $W = X \setminus E$ (i.e., all the non-evidence variables)
 - $MAP(W | e) = \operatorname{argmax}_w P(w, e)$
 - Important: We cannot find w by finding the maximum likely value for each variable individually
- 3. **Marginal MAP query**
 - $MAP(Y | e) = \operatorname{argmax}_y P(y | e)$
 - Let $Z = X \setminus E \cup Y$
 - $MAP(Y | e) = \operatorname{argmax}_y \sum_z P(z, y | e)$

MAP EXAMPLE

A	B	P(A, B)
t	t	0.10
t	f	0.25
f	t	0.35
f	f	0.30

Maximum likely assignment for A = f

Maximum likely assignment for B = f

$$\text{MAP}(A, B) = \langle A=f, B=t \rangle$$

NUMBER OF PARAMETERS

- Assuming everything is binary
- $P(V_1)$ requires
 - 1 independent parameter
- $P(V_1, V_2, \dots, V_n)$ requires
 - $2^n - 1$ independent parameters
- $P(V_1 | V_2)$ requires
 - 2 independent parameters
- $P(V_1, V_2, \dots, V_n | V_{n+1}, V_{n+2}, \dots, V_{n+m})$ requires
 - $2^m \times (2^n - 1)$ independent parameters

CONTINUOUS SPACES

- Assume X is continuous and $\text{Val}(X) = [0,1]$
- If you would like to assign the same probability to all real numbers in $[0, 1]$, what is, for e.g., $P(X=0.5) = ?$
- Answer: $P(X=0.5) = 0$.

PROBABILITY DENSITY FUNCTION

- We define **probability density function**, $p(x)$, a non-negative integrable function, such that $\int_{\text{Val}(X)} p(x)dx = 1$

$$P(X \leq a) = \int_{-\infty}^a p(x)dx$$

$$P(a \leq X \leq b) = \int_a^b p(x)dx$$

UNIFORM DISTRIBUTION

- A variable X has a uniform distribution over $[a,b]$ if it has the PDF

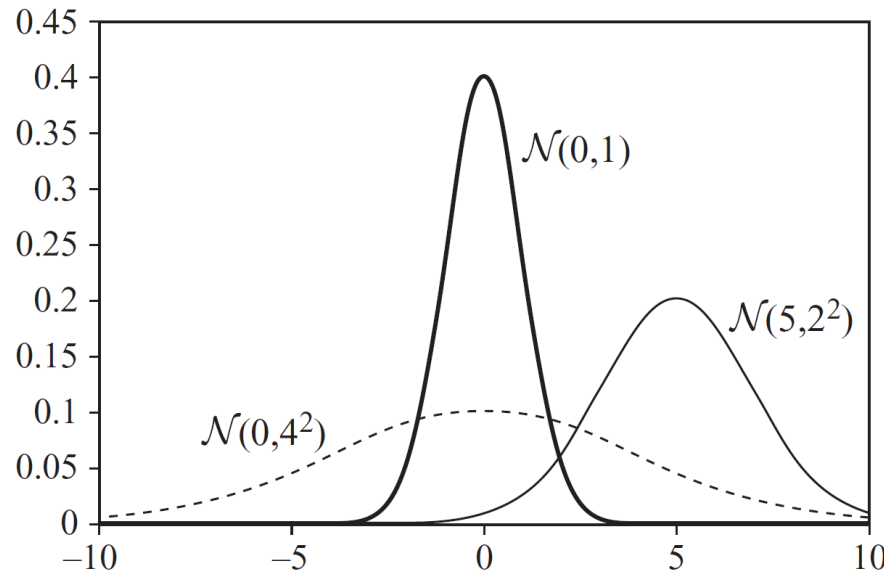
$$p(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Check and make sure that $p(x)$ integrates to 1.

GAUSSIAN DISTRIBUTION

- A variable X has a Gaussian distribution with mean μ and variance σ^2 , if it has the PDF

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



Can $p(x)$ be ever greater than 1?

CONDITIONAL PROBABILITY

- We want $P(Y | X=x)$ where X is continuous, Y is discrete
- $P(Y | X=x) = P(Y, X=x) / P(X=x)$
 - What's wrong with this expression?
- Instead, we use the following expression

$$P(Y | X = x) = \lim_{\varepsilon \rightarrow 0} P(Y | x - \varepsilon \leq X \leq x + \varepsilon)$$

CONDITIONAL PROBABILITY

- We want $p(Y | X)$ where X is discrete, Y is continuous
- How would you represent it?

EXPECTATION

$$E_P[X] = \sum_x xP(x)$$

$$E_P[X] = \int_x xp(x)dx$$

$$E_P[aX + b] = aE_P[X] + b$$

$$E_P[X + Y] = E_P[X] + E_P[Y]$$

$$E_P[X | y] = \sum_x xP(x | y)$$

What about $E[X*Y]$?

VARIANCE

$$\text{Var}_P[X] = E_P \left[\left(X - E_P[X] \right)^2 \right]$$

$$\text{Var}_P[X] = E_P[X^2] - \left(E_P[X] \right)^2$$

Can you derive the second expression using the first expression?

$$\text{Var}_P[aX + b] = a^2 \text{Var}_P[X]$$

What is $\text{Var}[X+Y]$?

UNIFORM AND GAUSSIAN DISTRIBUTION

- If $X \sim N(\mu, \sigma^2)$, then $E[X] = \mu$, $\text{Var}[X] = \sigma^2$
- What about the expectation and variance of a uniform distribution?