CS 581 – Advanced Artificial Intelligence

Topic: Probability Theory

Mustafa Bilgic

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

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

Random Variables

- Pick variables of interest
  - Medical diagnosis
    - Age, gender, weight, temperature, LT1, LT2, …
  - Loan application
    - Income, savings, payment history, …
  - Earlier examples
    - Grad student, Grade, Covid, Test result, Ache, X-Ray
- Every variable has a domain
  - Binary (e.g., True/False)
  - Categorical (e.g., Red/Green/Blue)
  - Real-valued (e.g., 97.8)
- **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: $t \land c, t, t \lor \sim c$,

Axioms of Probability

1. The probability P(a) of a proposition a is a real number between 0 and 1

2. P(true) = 1, P(false) = 0

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

$P(\neg a)$

- $P(a \vee \neg a) = P(a) + P(\neg a) - P(a \wedge \neg a)$
- $P(true) = P(a) + P(\neg a) - P(false)$
- $1 = P(a) + P(\neg a) - 0$
- $P(\neg a) = 1 - P(a)$
- 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: a variable
- Lowercase: x: a particular value of X
- Val(X): the set of values X can take
- Bold Capital: **X**: a set of variables
- Bold lowercase: **x**: an assignment to all variables in **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, $V_1, V_2, \ldots, V_n$
- We are interested in the probability of a possible world, where
  - $V_1 = v_1, V_2 = v_2, \ldots, V_n = v_n$
- $P(V_1, V_2, \ldots, V_n)$ associates a probability for each possible world $\equiv$ the **joint distribution**
- How many entries are there, if we assume the variables are all binary?

Toothache example

Ache	X-Ray	P(A, X)
toothache	cavity	0.15
toothache	¬cavity	0.10
¬toothache	cavity	0.05
¬toothache	¬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 a distribution over n variables, you can calculate the distribution over any subset of the variables by summing out the irrelevant ones
- For example
  - Given P(A, B, C, D)
  - Calculate
    - P(A)
    - P(A, C)
    - … (any subset)

LET'S ANSWER A FEW QUERIES

Ache	X-Ray	P(A, X)
toothache	cavity	0.15
toothache	¬cavity	0.10
¬toothache	cavity	0.05
¬toothache	¬cavity	0.70

- P(cavity) = ? 0.20
- P(¬cavity) = ? 0.80
- P(toothache) = ? 0.25
- P(¬toothache) = ? 0.75

Conditional Distribution

- $P(A, B, C \mid D, E, F, G) = \frac{P(A,B,C,D,E,F,G)}{P(D,E,F,G)}$

Let's Answer a few queries

Ache	X-Ray	P(A, X)
toothache	cavity	0.15
toothache	¬cavity	0.10
¬toothache	cavity	0.05
¬toothache	¬cavity	0.70

- P(cavity | toothache) = ?
- P(cavity | ¬toothache) = ?
- P(¬cavity | toothache) = ?
- P(¬cavity | ¬toothache) = ?
- P(toothache | cavity) = ?
- P(¬toothache | cavity) = ?
- P(toothache | ¬cavity) = ?
- P(¬toothache | ¬cavity) = ?

$$\frac{P(c,t)}{P(t)} = \frac{0.15}{0.25} = .6$$

Bayes' Rule

- $P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$

Bayes' Rule

- $P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$

$$P(B|A) = \frac{P(B,A)}{P(A)}$$

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

$$P(A,B) = P(A|B)P(B)$$

Bayes' Rule

- $P(B|A) = \frac{P(A|B)*P(B)}{P(A)}$
- Example use
  - P(cause | effect) = P(effect | cause)*P(cause) / P(effect)
- Why is this useful?
  - Because in practice it is easier to get probabilities for P(effect|cause) and P(cause) than for P(cause|effect)
    - E.g., P(disease|symptoms) = P(symptoms|disease)*P(disease) / P(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 $P(\beta|\alpha)$?

Class Example

- In a class, 70% of the hardworking students got an A. John got an A. What is the probability that John is a hardworking student?
- Possible worlds: 4
  - <h, a>, <h, ~a>, <~h, a>, <~h, ~a>
- Let's say there are 100 students in a class
- Let's say 10 of them work hard (h), 90 do not (~h)
- Probability of a randomly picked student being hardworking
  - $P(h) = 0.1$
- We are told that 70% of the hardworking students got an A.
  - $P(a \mid h) = 0.7$
  - 7 hardworking students got an A; 3 did not get an A.
- What is $P(h|a) = ?$

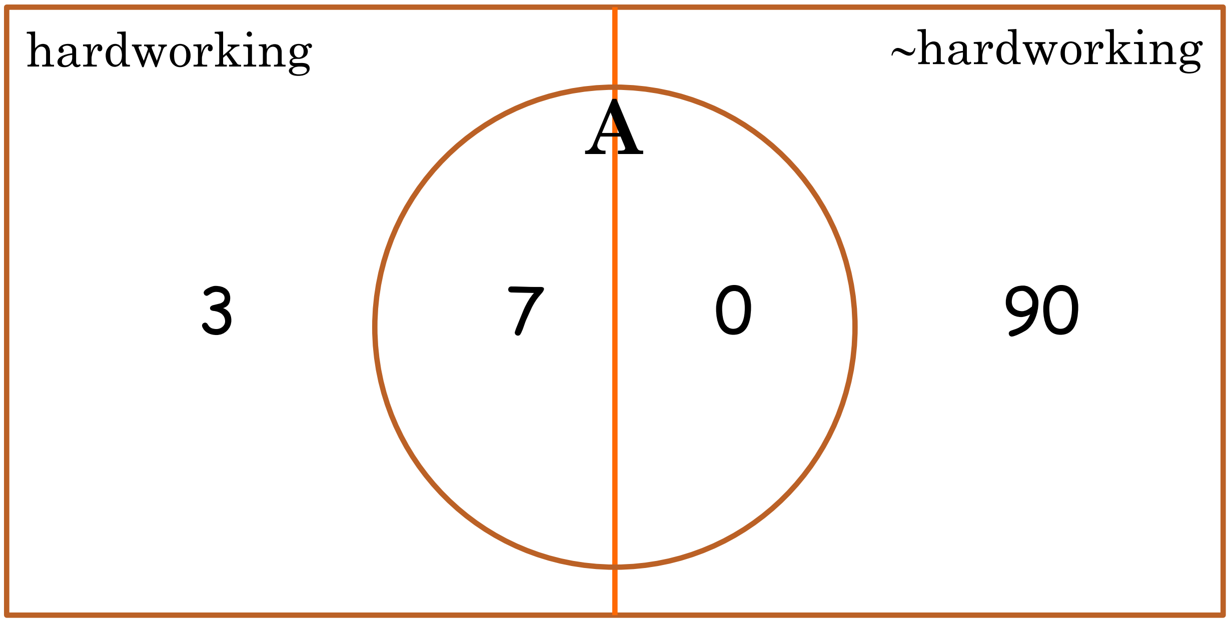
H , A

Class Example

$P(a|h) = 0.7$

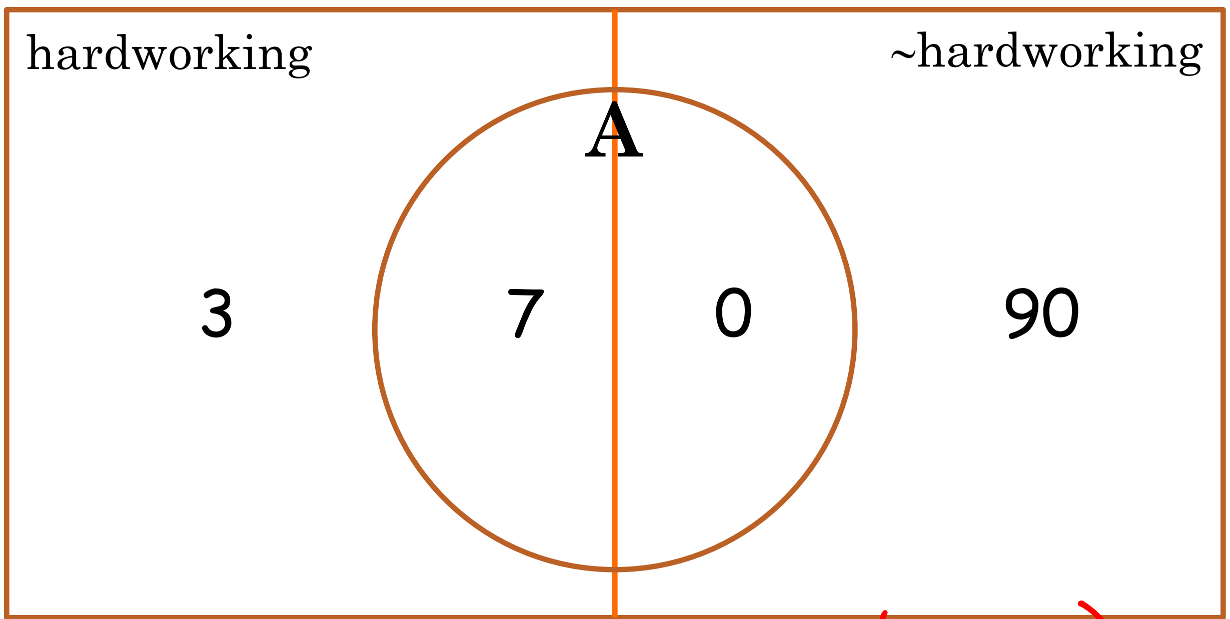
- In a class, 70% of the hardworking students got an A. John got an A. What is the probability that John is a hardworking student? $P(h|a)$
- Possible worlds: 4
  - <h, a>, <h, ~a>, <~h, a>, <~h, ~a>
- Let's say there are 100 students in a class
- Let's say 10 of them work hard (h), 90 do not (~h)
- Probability of a randomly picked student being hardworking
  - P(h) = 0.1
- We are told that 70% of the hardworking students got an A.
  - P(a | h) = 0.7
  - 7 hardworking students got an A; 3 did not get an A.
- What is P(h|a) = ?

Very Difficult Class



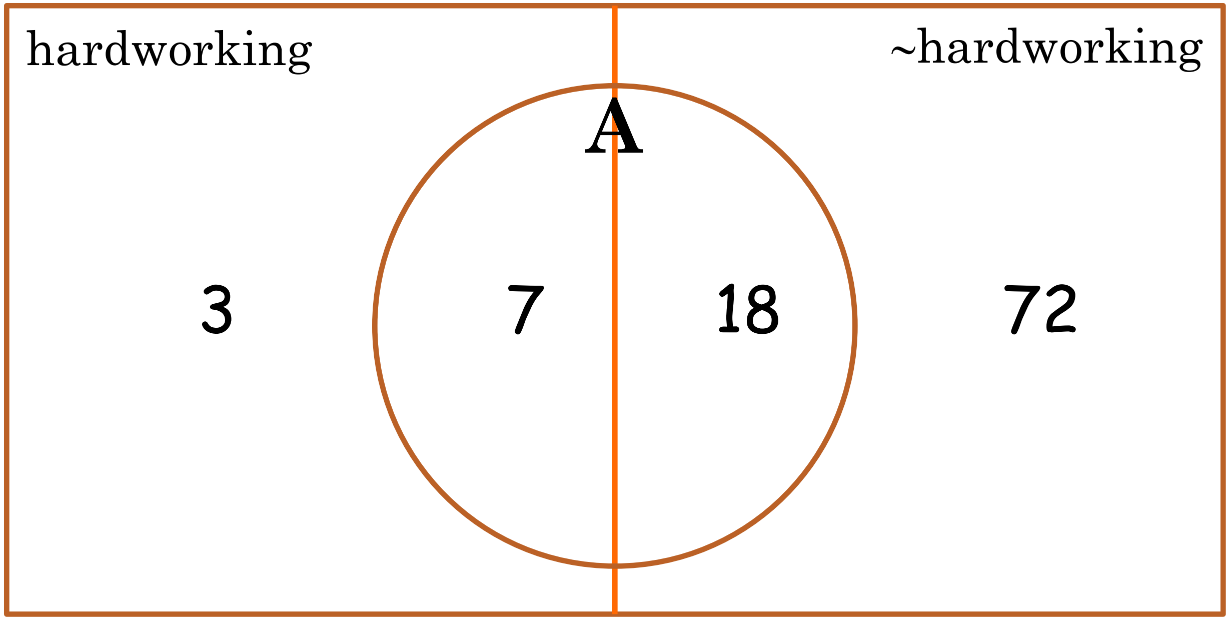
P(h | a) = ?

Very Difficult Class



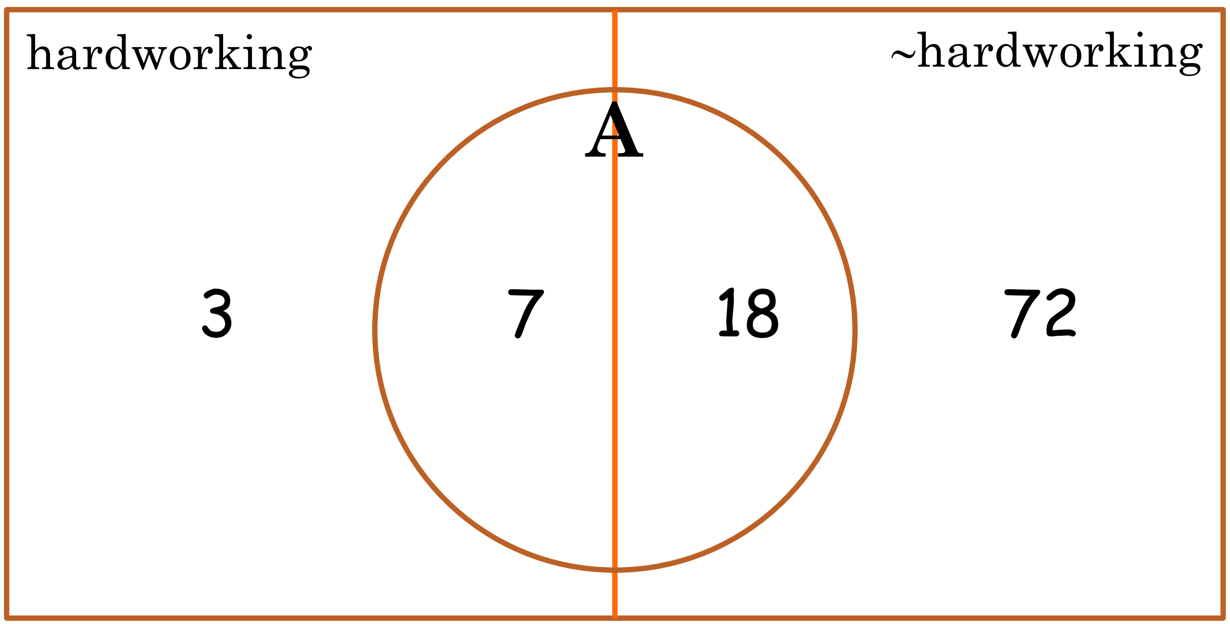
P(h | a) = ? $\frac{p(h,a)}{p(a)} = \frac{7}{7}$

MEDIUM DIFFICULT CLASS



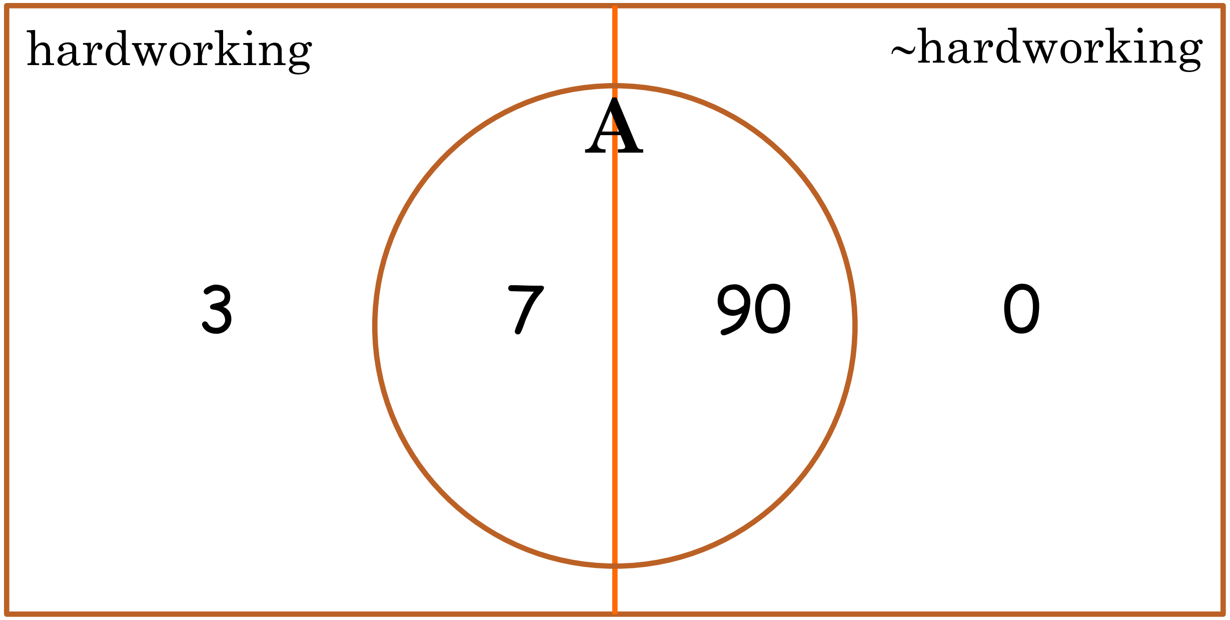
P(h | a) = ?

Medium Difficult Class



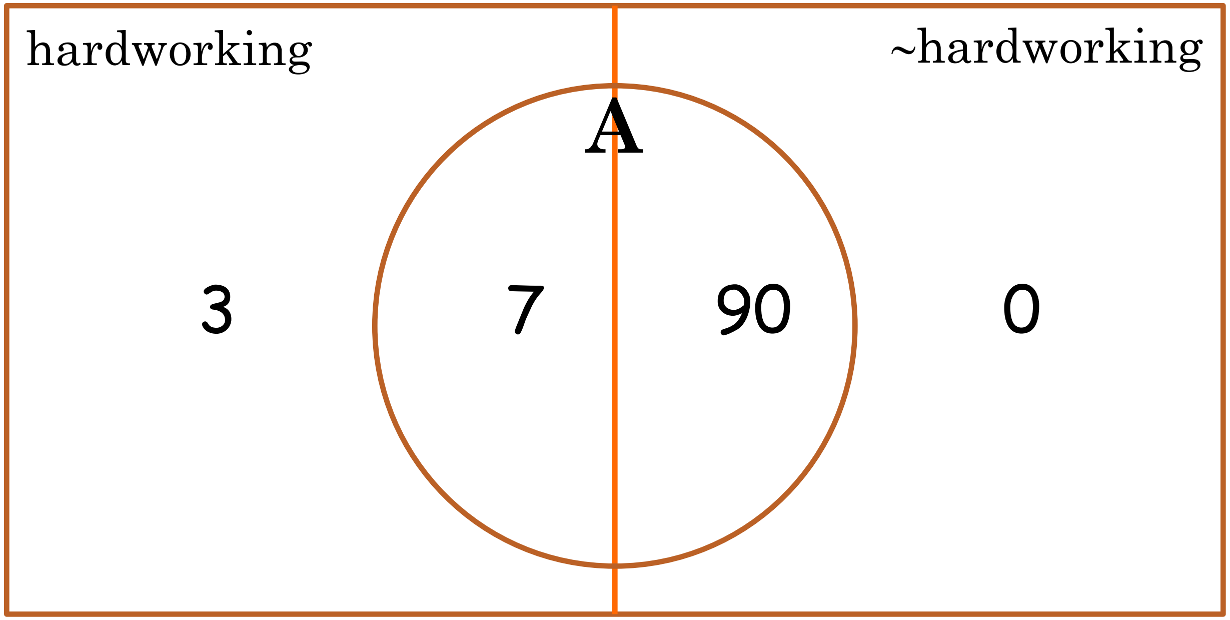
$P(h \mid a) = ?$ $\frac{7}{25} = 0.28$

Weird Class



$P(h \mid a) = ?$

Weird Class



P(h | a) = ?

$\frac{7}{97} \approx 0.07$

Chain Rule

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

$P(A,B,C,D) = P(A)\,P(B|A)\,P(C|A,B)\,P(D|A,B,C)$

A B C D

B D A C

$= P(B)\,P(D|B)$

$P(A|B,D)$

$P(C|B,D,A)$

MARGINAL INDEPENDENCE

- An event α is **independent** of event β in P, denoted as P ⊨ α ⊥ β, if
  - P(α | β) = P(α), or
  - P(β) = 0
- Proposition: A distribution P satisfies α ⊥ β if and only if
  - P(α, β) = P(α) P(β)
  - Can you prove it?
- Corollary: α ⊥ β implies β ⊥ α

$$P(\alpha,\beta) = P(\beta)\,\underbrace{P(\alpha|\beta)}$$

$$P(\beta)\,P(\alpha)$$

Marginal Independence

$P(x) \cdot P(y)$

X	Y	P(X, Y)
t	t	0.18 $0.6 \times 0.3 = 0.18$
t	f	0.42 $0.6 \times 0.7 = 0.42$
f	t	0.12 $0.4 \times 0.3 = 0.12$
f	f	0.28 $0.4 \times 0.7 = 0.28$

Is $X \perp Y$?

$$P(x,y) = P(x)P(y)$$

$$\langle 0.6, 0.4 \rangle \cdot \langle 0.3, 0.7 \rangle$$

$P(x|y) \overset{?}{=} P(x)$ $\quad$ $P(x|y) = \frac{P(x,y)}{P(y)}$

Marginal Independence

$\frac{0.18}{0.3} \overset{?}{=} 0.6$ $\quad$ $\frac{0.42}{0.7} \overset{?}{=} 0.6$ $\quad$ $P(x) \cdot P(y)$

X	Y	P(X, Y)	
t	t	0.18	$0.6 \times 0.3 = 0.18$
t	f	0.42	$0.6 \times 0.7 = 0.42$
f	t	0.12	$0.4 \times 0.3 = 0.12$
f	f	0.28	$0.4 \times 0.7 = 0.28$

Is X ⊥ Y?

$$P(x,y) = P(x)P(y)$$

$$\langle 0.6, 0.4 \rangle \cdot \langle 0.3, 0.7 \rangle$$

J M T

$\cancel{J \perp M}$

$J \perp M \mid T$

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
  - $P(\alpha, \beta \mid \gamma) = P(\alpha \mid \gamma)\, P(\beta \mid \gamma)$

Height Knowledge

$H \not\perp K$

$H \perp K \mid A$

Number of Parameters

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

Number of Parameters

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

Continuous Spaces

- Assume X is continuous and 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) = ?

Probability Density Function

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

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

$$P(a \le X \le b) = \int_{a}^{b} p(x)dx$$

Uniform Distribution

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

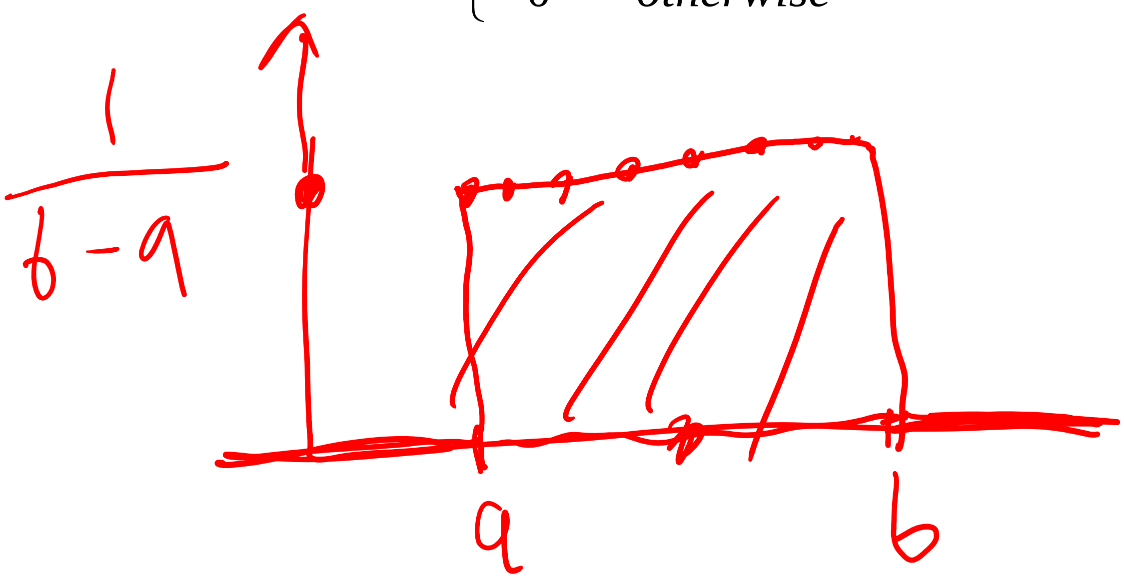
$$p(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & otherwise \end{cases}$$

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

Uniform Distribution

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

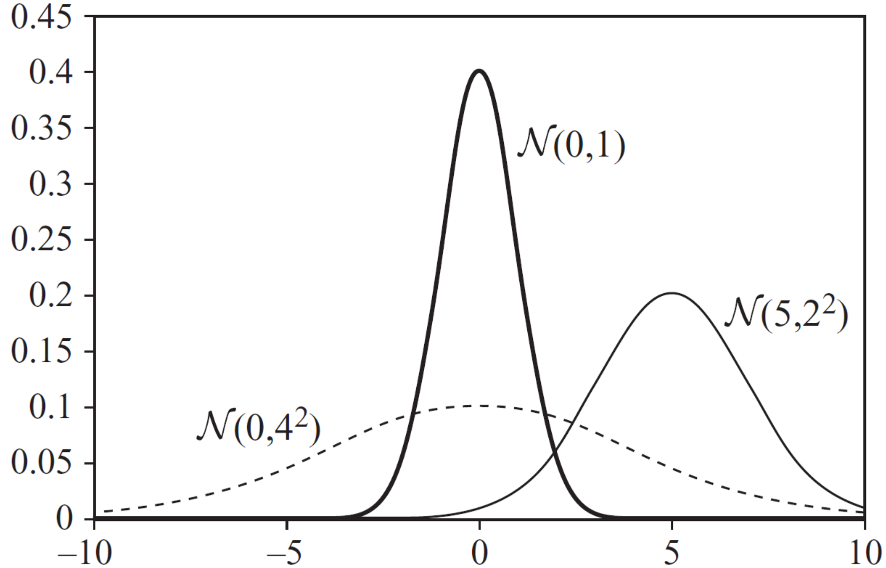
$$p(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & otherwise \end{cases}$$



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?

Yes!

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

$$P(Y \mid X = x) = \lim_{\varepsilon \to 0} P(Y \mid x - \varepsilon \le X \le x + \varepsilon)$$

Conditional Probability

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

$p(y \mid X=a)$ $\quad \mathcal{N}(\mu_a, \sigma_a^2)$

$p(y \mid X=b)$ $\quad \mathcal{N}(\mu_b, \sigma_b^2)$

$\vdots$ $\quad \vdots$ $\quad \vdots$

$p(y \mid X=z)$

Expectation

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

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

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

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

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

What about E[X*Y]?

Expectation

3 5 10

0.1 0.2 0.7

0.3 + 1 + 7 = 8.3

10 14 24

0.1 0.2 0.7

×2+4

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

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

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

2 15 16

0.3 0.4 0.3

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

$$E_P[X \mid y] = \sum_x x P(x \mid y)$$

What about E[X*Y]? X E[X] x E[Y]

Variance

$$Var_P[X] = E_P\left[\left(X - E_P[X]\right)^2\right]$$

$$Var_P[X] = E_P\left[X^2\right] - \left(E_P[X]\right)^2$$

Can you derive the second expression using the first expression?

$$Var_P[aX + b] = a^2 Var_P[X]$$

What is Var[X+Y]?

3 5 10

0.1 0.2 0.7

VARIANCE

$$Var_P[X] = E_P\left[\left(X - E_P[X]\right)^2\right] = E\left[x^2 - 2 \cdot x \cdot \bar{x} + \bar{x}^2\right]$$

$$Var_P[X] = E_P\left[X^2\right] - \left(E_P[X]\right)^2$$

Can you derive the second expression using the first expression?

$$Var_P[aX + b] = a^2 Var_P[X]$$

What is Var[X+Y]?

Uniform and Gaussian Distribution

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