CS 584 – Machine Learning

Topic: Probability Theory

Mustafa Bilgic

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

https://twitter.com/bilgicm

MOTIVATION

- Learning
  - Statistics, expectations, etc.
  - E.g., decision trees, naïve Bayes, logistic regression, …
- Evaluation
  - Statistics, expectations, variance, etc.
  - E.g., expected error using a sample

Some Questions

- Given a domain with n variables, $X_1, X_2, \dots, X_n$, each of which has $v_1, v_2, \dots, v_n$ possible values, what is the size of the instance space?
- Given a sample dataset and a hypothesis h, calculate the mean, variance, and 95% confidence interval for the error rate of h
- Type I error has cost c_1 and type II error has cost c_2. Correct decisions have no cost. The probability of the object belonging to Positive class is p. Should it be classified as Positive or Negative?
- Given *P(Symptoms|Diseases)*, *P(Diseases)* and *P(Symptoms)*, calculate *P(Diseases|Symptoms)*
- Given a probability distribution $p_1, p_2, \dots, p_k$, calculate its entropy

RANDOM VARIABLES

- Pick variables of interest
  - Medical diagnosis
    - Age, gender, weight, temperature, LT1, LT2, …
  - Loan application
    - Income, wealth, payment history, …
- Every variable has a domain
  - Binary (True/False)
  - Categorical
  - Real-valued
- Possible world
  - An assignment to all variables of interest

MP	I	W
Y	L	L
Y	L	H

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

  1..6 × 1..6 : 36
  - 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(true) = 1, P(false) = 0

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

$P(\neg a)$

- $P(a \vee \neg a) = P(a) + P(\neg a) - P(a \wedge \neg a)$ ← Axiom #3
- $P(\text{true}) = P(a) + P(\neg a) - P(\text{false})$
- $1 = P(a) + P(\neg a) - 0$ ← Axiom #2
- $P(\neg a) = 1 - P(a)$
- 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

  I

- Lowercase: x: a particular value of X

  l h

- Val(X): the set of values X can take

  {l, h}

- Bold Capital: **X**: a set of variables

  X : {I, MP}

- Bold lowercase: **x**: an assignment to all variables in **X**

  {l, n}

- P(X=x) will be shortened as P(x)
- P(X=x $\cap$ Y=y) will be shortened as P(x,y)

P(l, n)

JOINT DISTRIBUTION

- We have n random variables, $X_1, X_2, \ldots, X_n$
- We are interested in the probability of a possible world, where
  - X_1=low, X_2=red, …, X_n=circle
- $P(X_1, X_2, \ldots, X_n)$ associates a probability for each possible world ≡ the **joint distribution**
- How many entries are there, if we assume the variables are all binary?

$$\left.\begin{matrix} T & \cdots & T \\ T & \cdots & F \\ & \vdots & \\ F & \cdots & F \end{matrix}\right\} 2^n$$

TOOTHACHE EXAMPLE

Feeling	X-Ray	P(F,X)
toothache	cavity	0.15
toothache	¬cavity	0.10
¬toothache	cavity	0.05
¬toothache	¬cavity	0.70

P(I)

MP	I	W	
y	l	e	0.03
y	l	h	

1

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) $= \sum_B \sum_C \sum_D P(A,B,C,D)$
    - P(A, C) $= \sum_B \sum_D P(A,B,C,D)$
    - … (any subset)

LET'S ANSWER A FEW QUERIES

Feeling	X-Ray	P(F,X)
toothache	cavity	0.15
toothache	¬cavity	0.10
¬toothache	cavity	0.05
¬toothache	¬cavity	0.70

- P(cavity) = ? 0.15 + 0.05 = 0.20
- P(¬cavity) = ? 0.80
- P(toothache) = ? 0.15 + 0.10 = 0.25
- P(¬toothache) = ? 0.75

Conditional Distribution

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

$P(MP \mid I = l)$

$y \quad n$

$\langle 0.6, 0.4 \rangle$

$P(MP \mid I = h)$

$y \quad n$

$\langle 0.05, 0.95 \rangle$

LET'S ANSWER A FEW QUERIES

Feeling	X-Ray	P(F,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) = ?

$$\frac{P(\neg t, c)}{P(c)} = \frac{0.05}{0.20} = \frac{1}{4}$$

$$\frac{P(\neg t, \neg c)}{P(\neg c)} = \frac{0.70}{0.80} = 7/8$$

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)

d s

- 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 $P(\beta|\alpha)$?

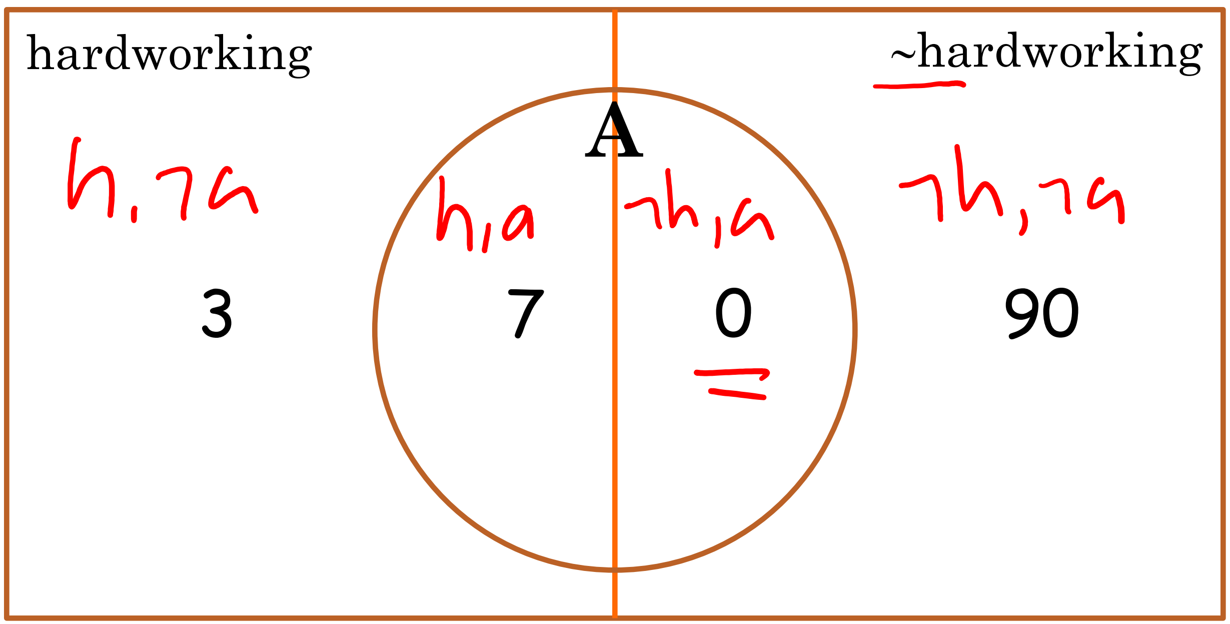
Class Example

- Let's say there are 100 students in the 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) = ?

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

$p(h) = 0.10$

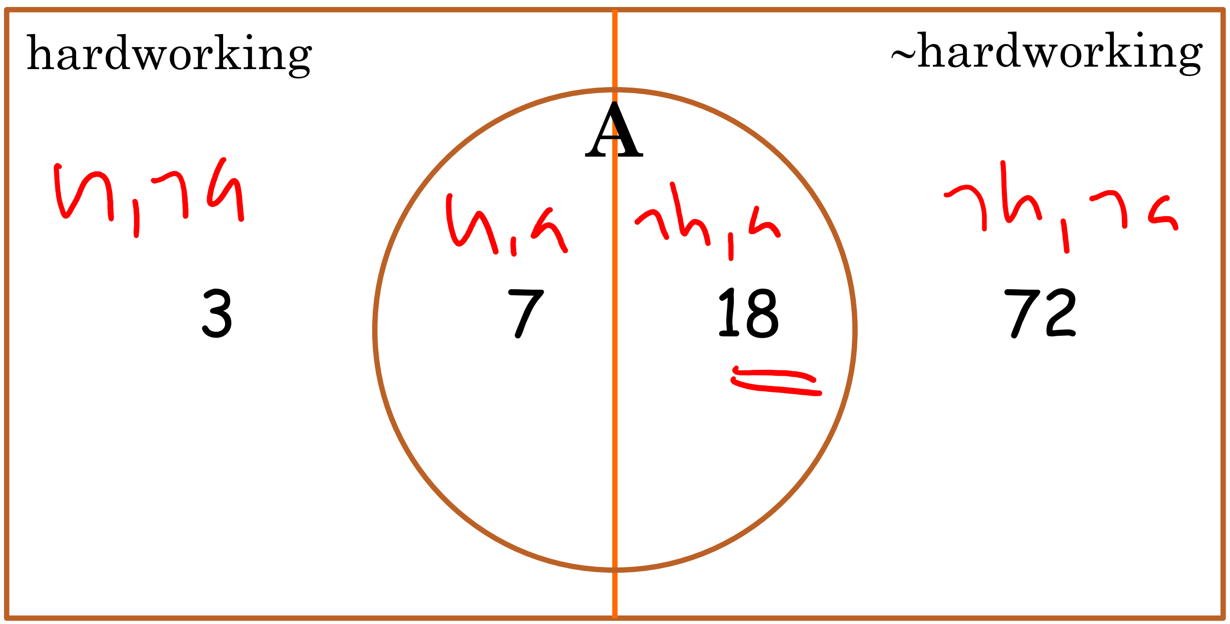
Very Hard Class



P(h | a) = ? $\frac{h, a}{a} = \frac{7}{7} = 1$

Medium Hard Class

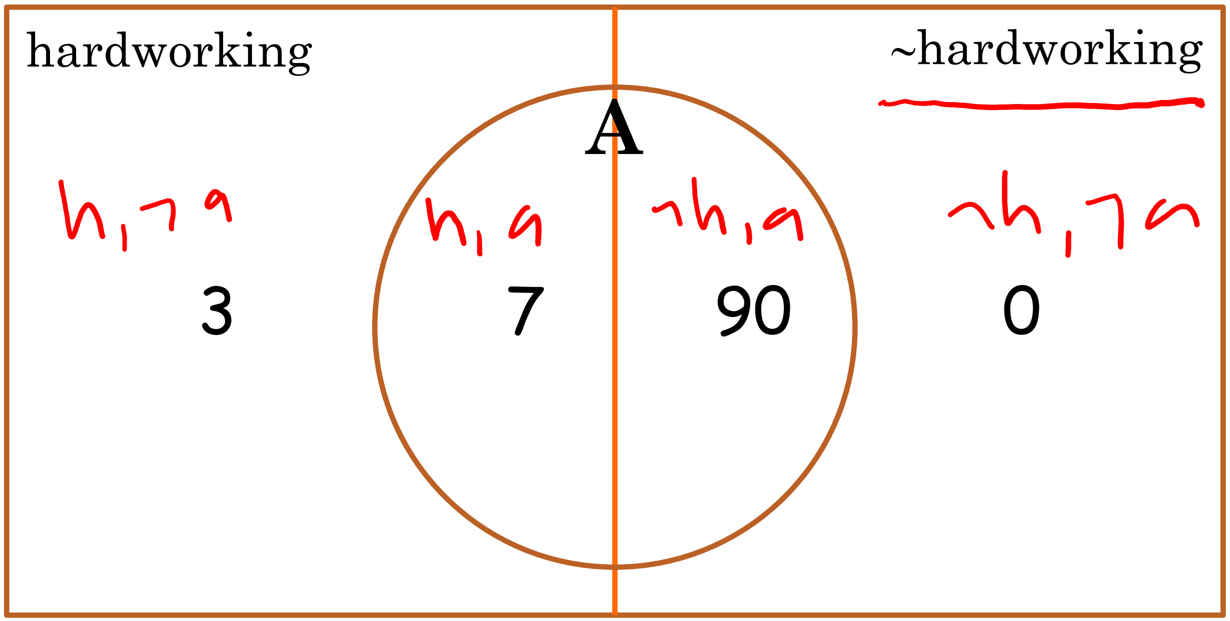
$p(h) = 0.1$



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

$p(h): 0.10$

Weird Class



P(h | a) = ?

$\frac{h,a}{a} : \frac{7}{97} \approx 0.07\ldots$

Chain Rule

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

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

C, A, D, B

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

B, D, C, A

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

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

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

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

Height & Knowledge

$H \perp K$? No

$H \perp K \mid A$ Yes

Number of Parameters

- Assuming everything is binary
- $P(X_1)$ requires
  - 1 independent parameter
- $P(X_1, X_2, \ldots, X_n)$ requires
  - 2^n-1 independent parameters
- $P(X_1 | X_2)$ requires
  - 2 independent parameters
- $P(X_1, X_2, \ldots, X_n \mid X_{n+1}, X_{n+2}, \ldots, X_{n+m})$ requires
  - $2^m \times (2^n-1)$ independent parameters

Number of Parameters

- Assuming everything is binary
- $P(X_1)$ requires
  - 1 independent parameter
- $P(X_1, X_2, \ldots, X_n)$ requires
  - 2^n-1 independent parameters
- $P(X_1 \mid X_2)$ requires
  - 2 independent parameters
- $P(X_1,X_2, \ldots, X_n \mid X_{n+1}, X_{n+2}, \ldots, X_{n+m})$ requires
  - $2^m \times (2^n\text{-}1)$ independent parameters

X_1	$P(X_1)$
T	–
F	–

$2^n - 1$

X_1	$X_1 \mid X_2 = T$
T	–
F	–

X_1	$X_1 \mid X_2 = F$
T	–
F	–

# of tables

# of entries

$(2^n - 1) \times 2^m$

X : T, F
y : R, G, B

x	y	
T	R	P_1
T	B	P_2
T	G	P_3
F	R	P_4
F	B	P_5
F	G	P_6

$\sum P_i = 1$

ind params
$2 \times 3 - 1$

$P(y|x)$

y	$P(y \mid x=T)$
R	—
G	—
B	—

$\overline{+} \quad 1$

y	$P(y \mid x=F)$
R	—
G	—
B	—

$\overline{+} \quad 1$

2+2 ind

x	$P(x \mid y=R)$
T	—
F	—

$\overline{+} \quad 1$

x	$P(x : y=G)$
T	—
F	—

$\overline{+} \quad 1$

x	$P(x \mid y=B)$
T	—
F	—

$\overline{+} \quad 1$

1+1+1 = 3 ind

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) = ?
- Answer: P(X=0.5) = 0.

Probability Density Function

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

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

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

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?

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]?

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]?

Some Distributions

BINOMIAL DISTRIBUTION

- Two parameters
  - n – number of independent experiments each measuring a binary outcome (e.g., Yes/No, Heads/Tails, Positive/Negative, …)
  - p – “success” probability for each individual experiment (e.g., Yes, Heads, Positive, …)
- Probability of exactly k successes
  - $P(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$, where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- Expected value: np
- Variance: $np(1-p)$
- An important use for binomial distribution: estimate a binary measure using a sample
  - Given k success in n experiments, estimate p and a confidence interval for p

Uniform Distribution

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

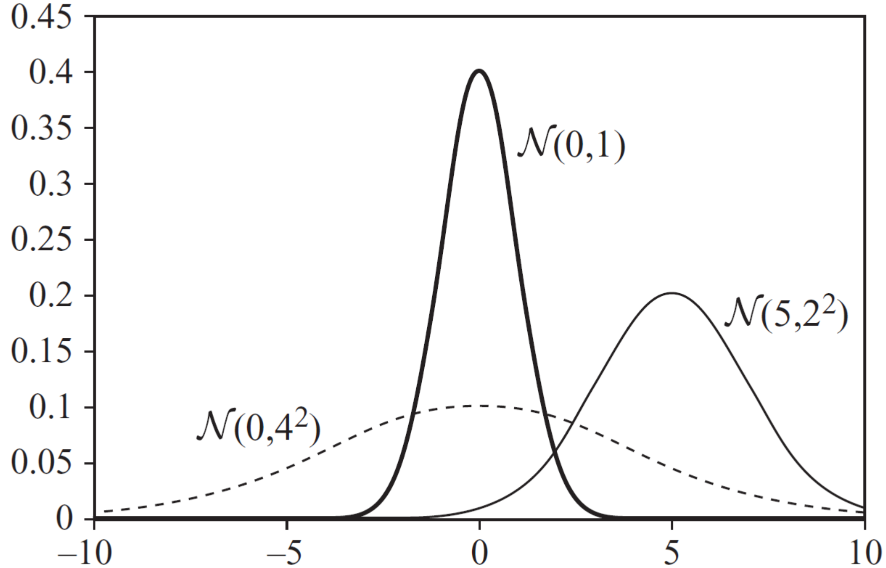
$$p(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & otherwise \end{cases}$$

Check and make sure that p(x) integrates to 1.
What are the mean and variances of this distribution?

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?

OTHER TOPICS

- Information theory
- Parameter estimation
- Decision-making under uncertainty