Warm-up As You Walk In

$g(\vec{x}) = \|\vec{x}\|_2^2$

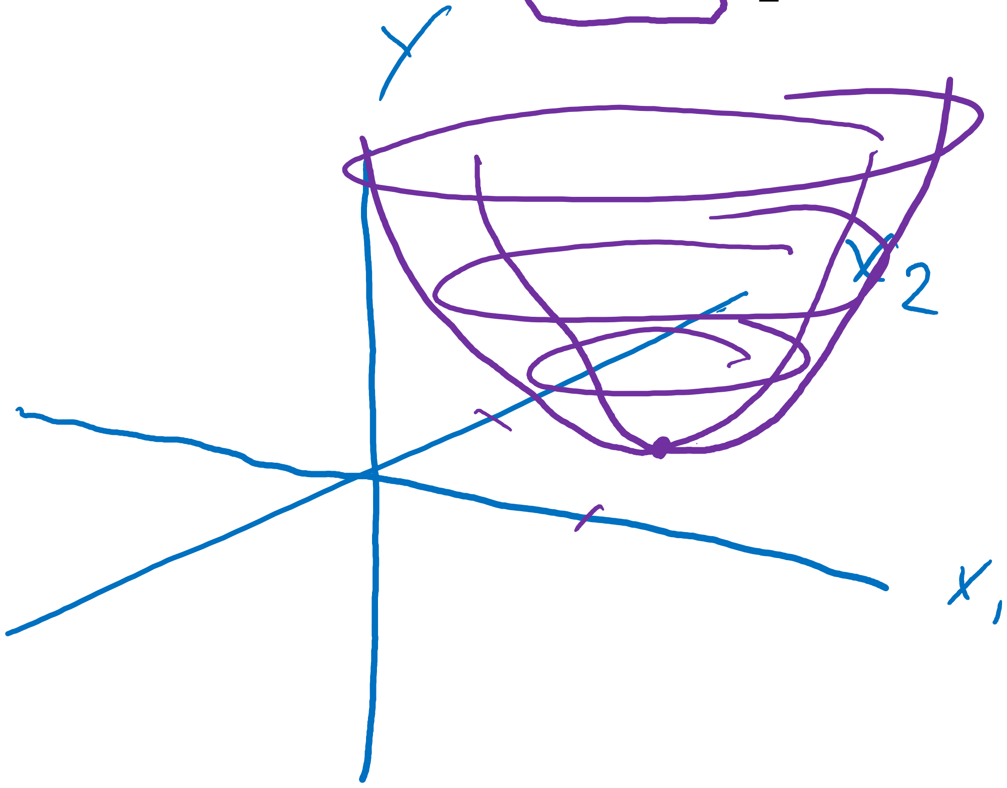
$= x_1^2 + x_2^2$

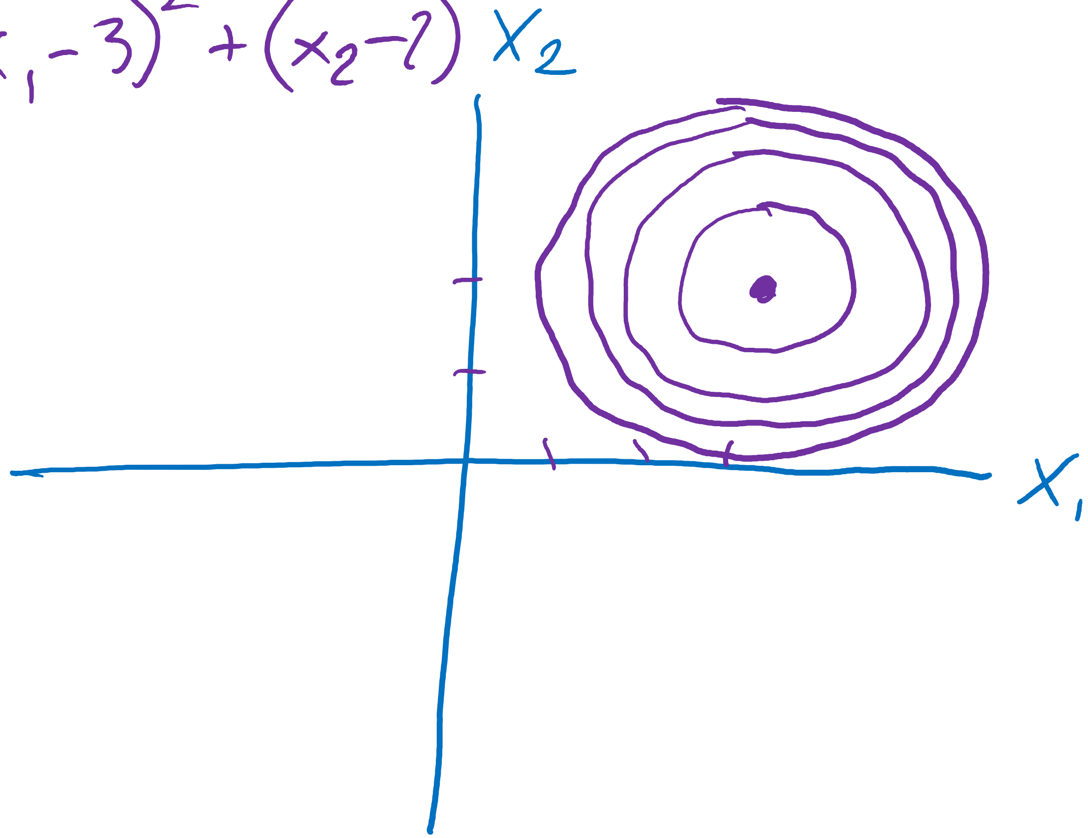
What does this function look like when you plot is as a surface in 3D (y vs x_1, x_2)?

How about as a contour map in 2D (drawn on the x_1, x_2 plane)?

- $y = f(\mathbf{x}) = \left\|\mathbf{x} - \begin{bmatrix} 3 \\ 2 \end{bmatrix}\right\|_2^2, \quad \mathbf{x} \in \mathbb{R}^2$

$(x_1 - 3)^2 + (x_2 - 2)^2$





Announcements

HW1

- Grades can be released after second slip day expires

HW2

- Out this week

Quiz

- Today, last 15 min. of class

Survey

- Thanks for the feedback!

Mathematical Foundations for Machine Learning

Lagrange Multipliers & Probability

Instructor: Pat Virtue

Plan

Last time

End class with minimax game

Today

Constrained optimization

- Formulation
- Solving with Lagrange multipliers

vector norm

Probability

- Vocab
- Properties
- Discrete distributions

Constrained Optimization

Method of Lagrange multipliers

Exercise

Mini-max game

$$L(x, \lambda) = 2x + 9 - \lambda(x^2 - 2)$$

Two teams

Team λ:

- Goes first
- Chooses a value for λ in attempt to *maximize* $L(x, \lambda)$

Team x:

- Goes second
- Chooses a value for x in attempt to *minimize* $L(x, \lambda)$

Constrained Optimization

Notation

$$\min_{\vec{x}} \; f(\vec{x})$$

$$\text{s.t.} \quad g(\vec{x}) = 0$$

$$h(\vec{x}) = c$$

$$\underbrace{h(\vec{x}) - c}_{} = 0$$

Method of Lagrange Multipliers

Goal

scalar

$$\min_{\mathbf{x}} \quad f(\mathbf{x})$$

scalar

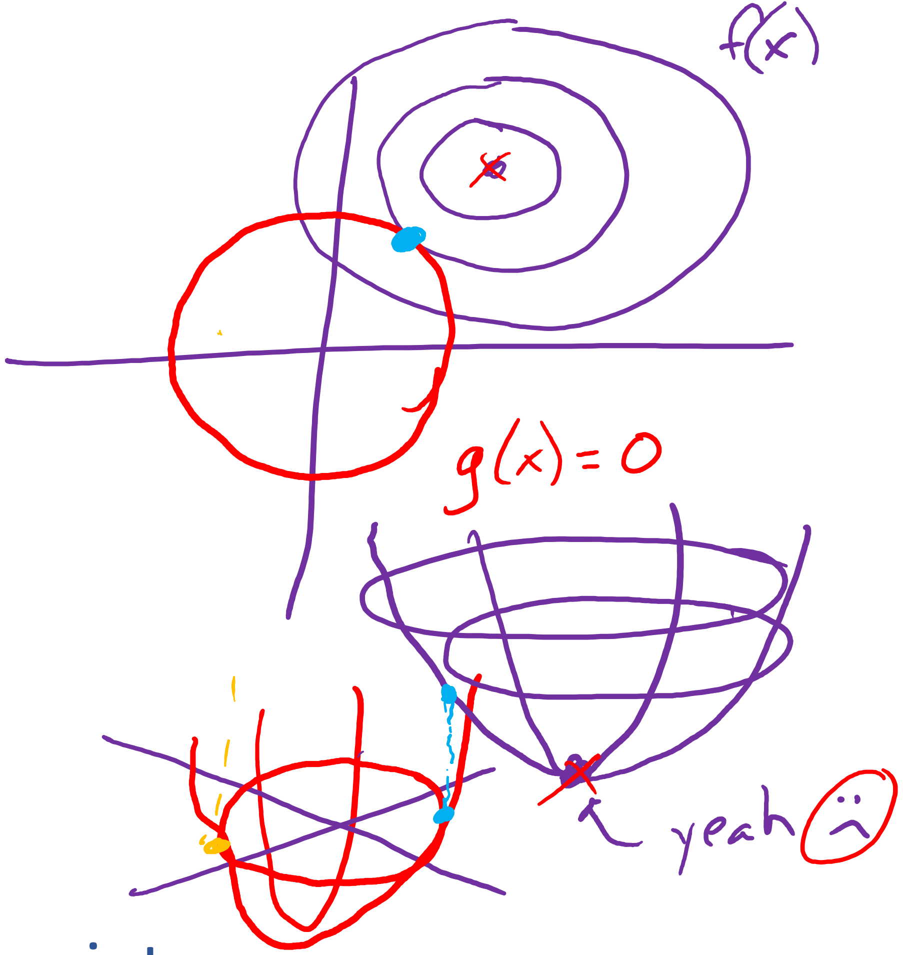
$$\text{s.t.} \quad g(\mathbf{x}) = 0$$

Step 1: Construct Lagrangian

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x})$$

Step 2: Solve

$$\min_{\mathbf{x}} \quad \max_{\lambda} \quad \mathcal{L}(\mathbf{x}, \lambda)$$



Find saddle point:

$$\nabla \mathcal{L}(\mathbf{x}, \lambda) = \mathbf{0}$$

Example

Mini-max game

$$\min_{\mathbf{x}} \quad 2x + 9$$
$$\text{s.t.} \quad x^2 = 2$$

$$\underbrace{x^2 - 2}_{g(x)} = 0$$

$$\mathcal{L}(x, \lambda) = 2x + 9 - \lambda(x^2 - 2)$$

Goal

$$\min_{\mathbf{x}} \quad f(\mathbf{x})$$
$$\text{s.t.} \quad g(\mathbf{x}) = 0$$

Step 1: Construct Lagrangian

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x})$$

Step 2: Solve

$$\min_{\mathbf{x}} \max_{\lambda} \mathcal{L}(\mathbf{x}, \lambda)$$

(Step 2: Find saddle point)

$$\nabla \mathcal{L}(\mathbf{x}, \lambda) = \mathbf{0}$$

Jupyter notebook

Method of Lagrange Multipliers

Lagrangian

$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x})$ $= f(x_1, x_2) - \lambda g(x_1, x_2)$

Find saddle point:

$\nabla \mathcal{L}(\mathbf{x}, \lambda) = \mathbf{0}$

$$\frac{\partial \mathcal{L}}{\partial x_1} = \frac{\partial f}{\partial x_1} - \lambda \frac{\partial g}{\partial x_1} = 0$$

$$\frac{\partial \mathcal{L}}{\partial x_2} = \frac{\partial f}{\partial x_2} - \lambda \frac{\partial g}{\partial x_2} = 0$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = -g(x) = 0$$

$$\left.\begin{aligned} \frac{\partial f}{\partial x_1} &= \lambda \frac{\partial g}{\partial x_1} \\ \frac{\partial f}{\partial x_2} &= \lambda \frac{\partial g}{\partial x_2} \end{aligned}\right] \nabla_x f = \lambda \nabla_x g$$

$$g(x) = 0$$

Method of Lagrange Multipliers

Goal

$$\min_{\mathbf{x}} \quad f(\mathbf{x})$$
$$\text{s.t.} \quad g(\mathbf{x}) = 0$$

Step 1: Construct Lagrangian

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x})$$

Step 2: Solve

$$\min_{\mathbf{x}} \quad \max_{\lambda} \quad \mathcal{L}(\mathbf{x}, \lambda)$$

Find saddle point:

$$\nabla \mathcal{L}(\mathbf{x}, \lambda) = \mathbf{0}$$

Equivalent to solving:

$$\nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x}) \quad \text{and} \quad g(\mathbf{x}) = 0$$

Method of Lagrange Multipliers (Inequality)

Goal

$$\min_{\mathbf{x}} \quad f(\mathbf{x})$$
$$\text{s.t.} \quad g(\mathbf{x}) \leq 0$$

Step 1: Construct Lagrangian

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$$

Step 2: Solve

$$\min_{\mathbf{x}} \quad \max_{\lambda \geq 0} \quad \mathcal{L}(\mathbf{x}, \lambda)$$

Find saddle point:

$$\nabla \mathcal{L}(\mathbf{x}, \lambda) = \mathbf{0} \quad \text{s.t. } \lambda \geq 0$$

Equivalent to solving:

$$\nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x}) \quad \text{s.t. } \lambda \geq 0 \text{ and } g(\mathbf{x}) = 0$$

Method of Lagrange Multipliers (multiple constraints)

Goal

$$\min_{\mathbf{x}} \quad f(\mathbf{x})$$

$$\text{s.t.} \quad g_1(\mathbf{x}) = 0$$

$$g_2(\mathbf{x}) = 0$$

Step 1: Construct Lagrangian

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda_1 g_1(\mathbf{x}) - \lambda_2 g_2(\mathbf{x})$$

Step 2: Solve

$$\min_{\mathbf{x}} \quad \max_{\lambda_1, \lambda_2} \quad \mathcal{L}(\mathbf{x}, \lambda_1, \lambda_2)$$

Find saddle point:

$$\nabla \mathcal{L}(\mathbf{x}, \lambda_1, \lambda_2) = \mathbf{0}$$

Vector Norms

$\mathbf{u} \in \mathbb{R}^M$

p-norm (general)

$$\|\mathbf{u}\|_p = \left(\sum_i^M |u_i^p|\right)^{1/p}$$

L2 norm (Euclidean norm)

$$\|\mathbf{u}\|_2 = \left(\sum_i^M u_i^2\right)^{1/2} = \left(\boldsymbol{u}^T \boldsymbol{u}\right)^{1/2}$$

L1 norm

$$\|\mathbf{u}\|_1 = \sum_i^M |u_i|$$

sum of abs entries

L0 "norm" (not really a norm)

$$\|\mathbf{u}\|_0 = |\{ u_i \mid u_i \neq 0 \}|$$

Number of non-zero entries

Probability

Probability Vocab

Sample Space Ω

Outcomes $o_1 \quad o_2 \quad o_3 \quad o_4$

Events $E = \{o_1, o_3\}$

Probability $P(E)$

Random variable $X: \Omega \to \mathbb{R}$

Discrete random variable $\to$

Continuous random variable

Probability mass function

Probability density function

$X(HH) \to 2$

$X(HT) \to 1$

Probability Vocab

Outcomes

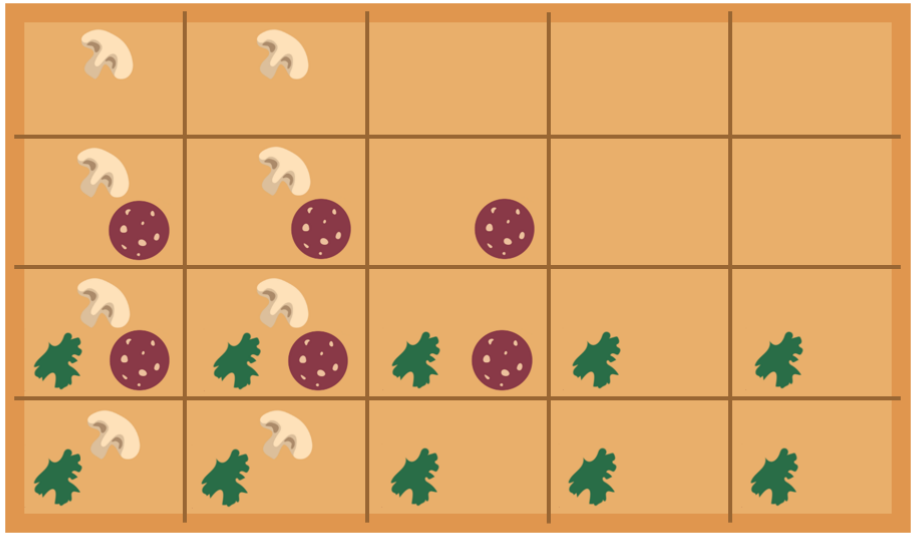
Events

Probability

Random variable

Discrete random variable

Probability mass function



Probability Toolbox

- Algebra
- Three axioms of probability
- Theorem of total probability
- Definition of conditional probability
- Product rule
- Bayes' theorem
- Chain rule
- Independence
- Conditional independence

Omega Pizzeria

Three questions: What is the probability of getting a slice with:

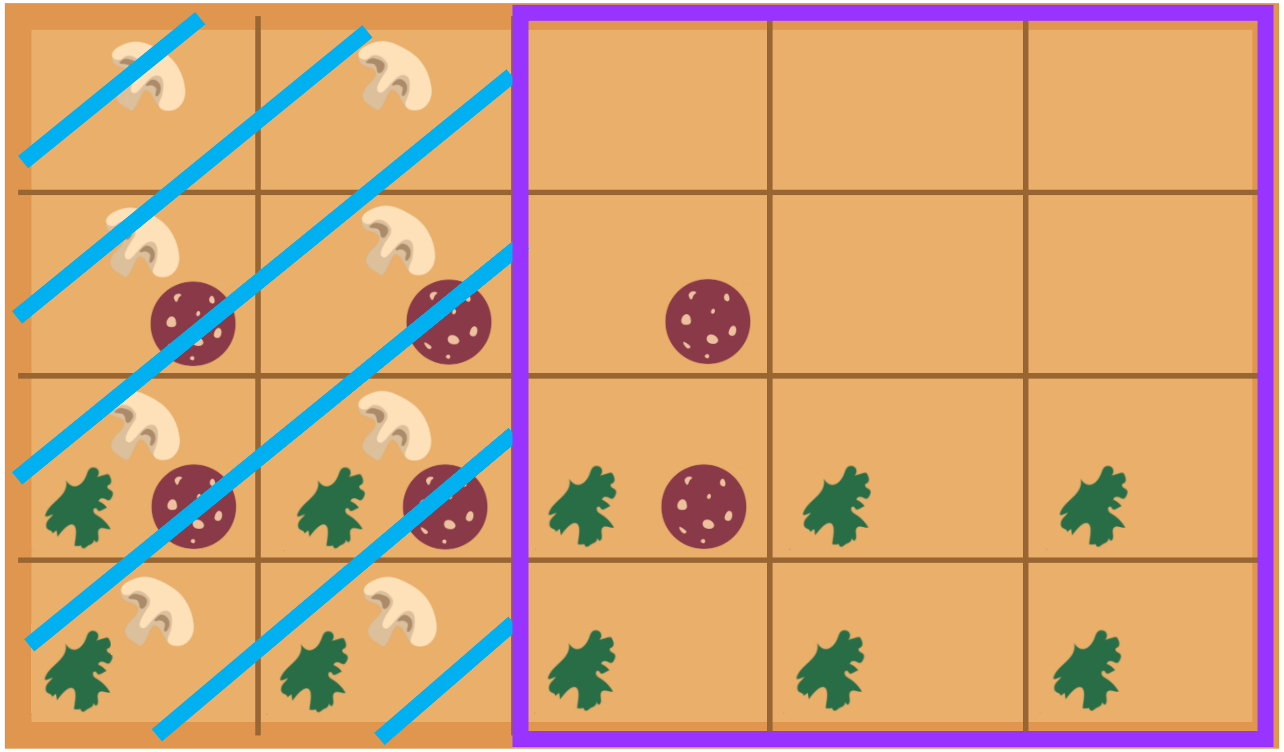
1) No mushrooms

2) Spinach and no mushrooms

3) Spinach, when asking for slice with no mushrooms

New information (condition)

Adjust sample space

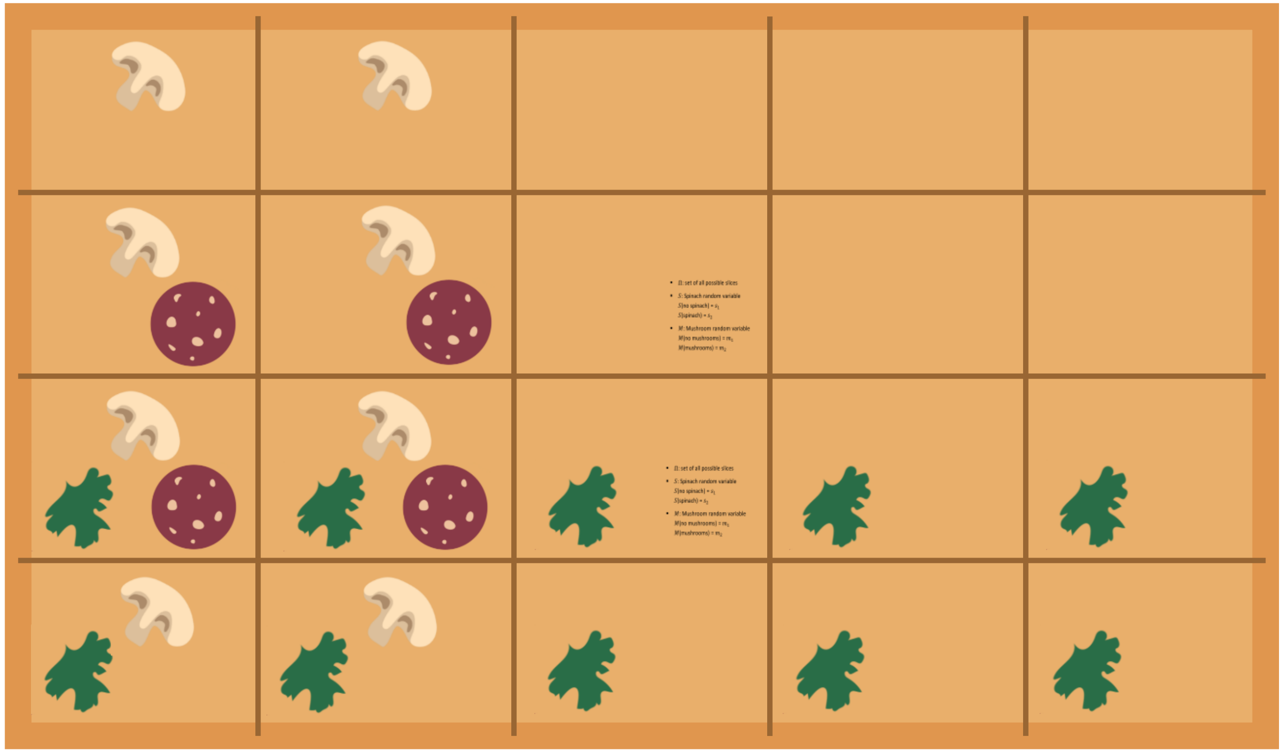


Icons: CC, https://openclipart.org/detail/296791/pizza-slice

Omega Pizzeria

Formalize this a bit

- Ω: set of all possible slices
- S: Spinach random variable
  S(no spinach) = s_1
  S(spinach) = s_2
- M: Mushroom random variable
  M(no mushrooms) = m_1
  M(mushrooms) = m_2



Icons: CC, https://openclipart.org/detail/296791/pizza-slice

Omega Pizzeria

Formalize this a bit

- Ω: whole pizza
- S: Spinach random variable

  S(no spinach) = s_1

  S(spinach) = s_2
- M: Mushroom random variable

  M(no mushrooms) = m_1

  M(mushrooms) = m_2

1) No mushrooms

   $P(M = m_1)$
2) Spinach and no mushrooms

   $P(S = s_2, M = m_1)$
3) Spinach, when asking for slice with no mushrooms

   $P(S = s_2 \mid M = m_1)$

Vocab alert!

Probability Vocab

Marginal

Joint

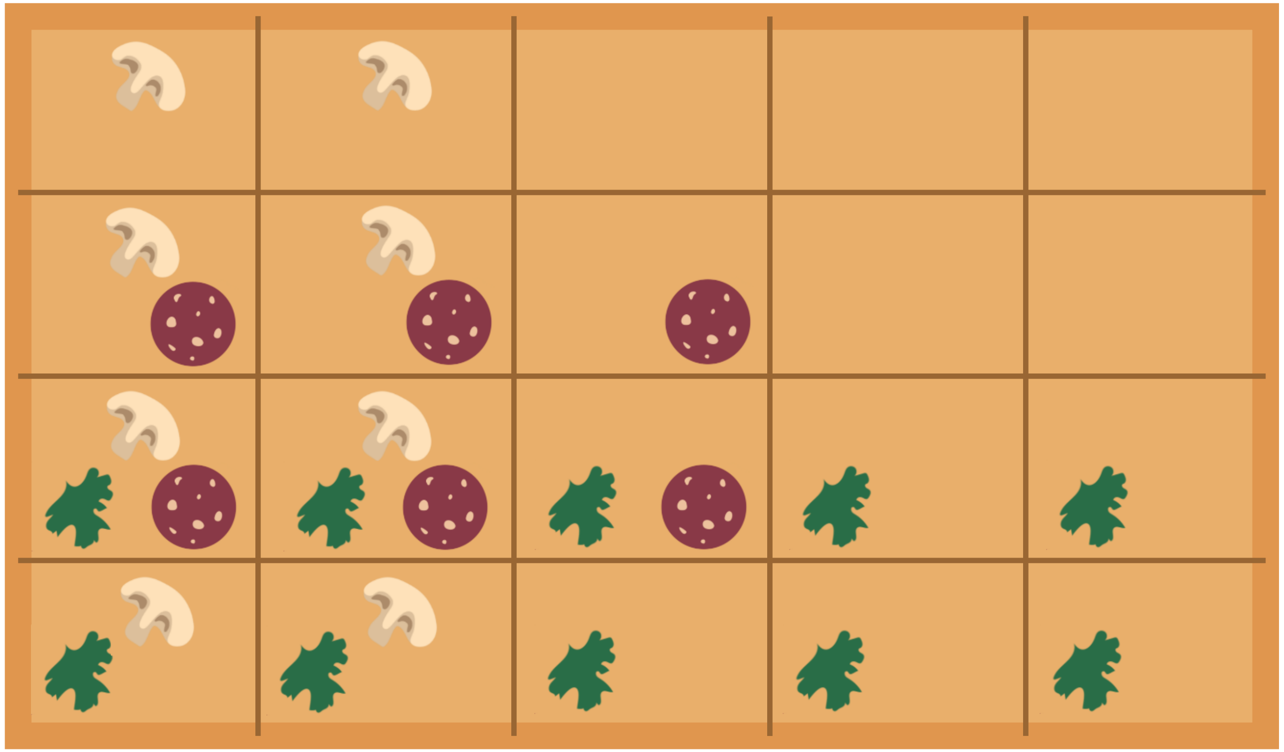
Conditional

Omega Pizzeria

More questions: What is the probability of getting a slice with:

1) No mushrooms
2) Spinach and no mushrooms
3) Spinach, when asking for slice with no mushrooms

- Mushrooms
- Spinach
- No spinach
- No spinach and mushrooms
- No spinach when asking for no mushrooms
- No spinach when asking for mushrooms
- Spinach when asking for mushrooms
- No mushrooms and no spinach



Icons: CC, https://openclipart.org/detail/296791/pizza-slice

Omega Pizzeria

You can fill out all of these probability mass functions

$p_M(m)$

m_1	12/20
m_2	

$p_S(s)$

s_1	
s_2	

$p_{M,S}(m, s)$

m_1	s_1	
m_1	s_2	6/20
m_2	s_1	
m_2	s_2	

$p_{M|S}(m \mid s_1)$

m_1	
m_2	

$p_{M|S}(m|s_2)$

m_1	
m_2	

$p_{S|M}(s \mid m_1)$

s_1	
s_2	6/12

$p_{S|M}(s|m_2)$

s_1	
s_2	

Definition of Conditional Probability

Definition:

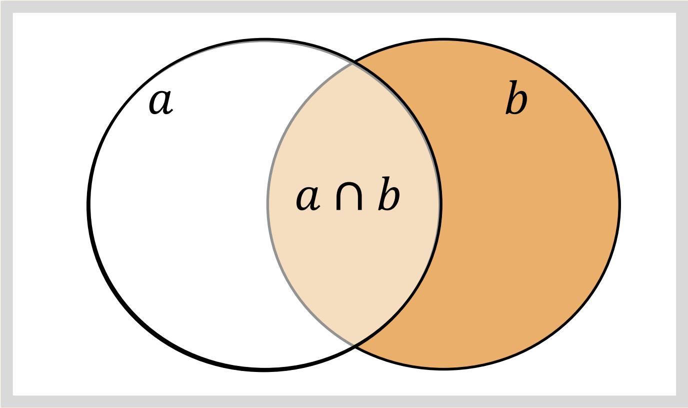
If $P(b) > 0$, then the **conditional probability** *of a given b is:*

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

Counting: proportions

$$P(a) = \frac{Count(a)}{Count(\Omega)}$$

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



Omega Pizzeria

Apply definition of conditional probability

- No mushrooms

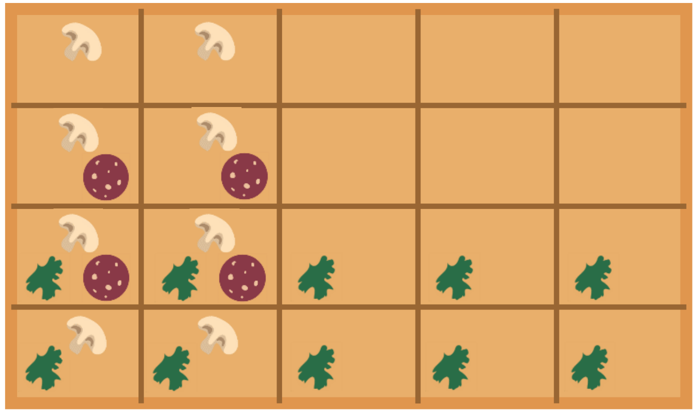
$$p(m_1) = \frac{12}{20}$$

- Spinach and no mushrooms

$$p(s_2, m_1) = \frac{6}{20}$$

- Spinach, when asking for slice with no mushrooms

$$p(s_2 | m_1) = \frac{6}{12}$$

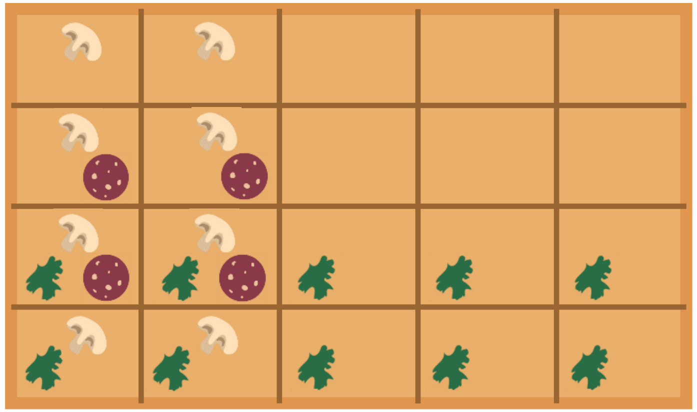


Conditional Probability:

$$p(a|b) = \frac{p(a,b)}{p(b)}$$

Omega Pizzeria

Apply definition of conditional probability



- No mushrooms

$$\mathrm{p}(\mathrm{m}_1) = \frac{12}{20}$$

- Spinach and no mushrooms

$$\mathrm{p}(\mathrm{s}_2, \mathrm{m}_1) = \frac{6}{20}$$

Conditional Probability: $p(a|b) = \dfrac{p(a,b)}{p(b)}$

- Spinach, when asking for slice with no mushrooms

$$\mathrm{p}(\mathrm{s}_2|\mathrm{m}_1) = \frac{6}{12}$$

$$p(s_2|m_1) = \frac{p(s_2, s_1)}{p(s_1)} = \frac{\frac{6}{20}}{\frac{12}{20}} = \frac{6}{12}$$

Definition of Conditional Probability

Definition:

If $P(B) > 0$, *then the* **conditional probability** *of* A *given* B *is:*

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

http://achievementgen.com/360/

Normalization Trick

P(X | Y=0) ?

$$P(X,Y)$$

X	Y	P
1	1	0.2
1	0	0.3
0	1	0.4
0	0	0.1

SELECT the joint probabilities matching the evidence →

NORMALIZE the selection (make it sum to one) →

To Normalize

(Dictionary) To bring or restore to a normal condition

All entries sum to ONE

Procedure:

- Step 1: Compute Z = sum over all entries
- Step 2: Divide every entry by Z

Example 1

W	$p_{W,T}$ (w,1)
sun	0.2
rain	0.3

Normalize

Z = 0.5

W	p(w)
sun	0.4
rain	0.6

- Example 2

T	W	Count
hot	sun	20
hot	rain	5
cold	sun	10
cold	rain	15

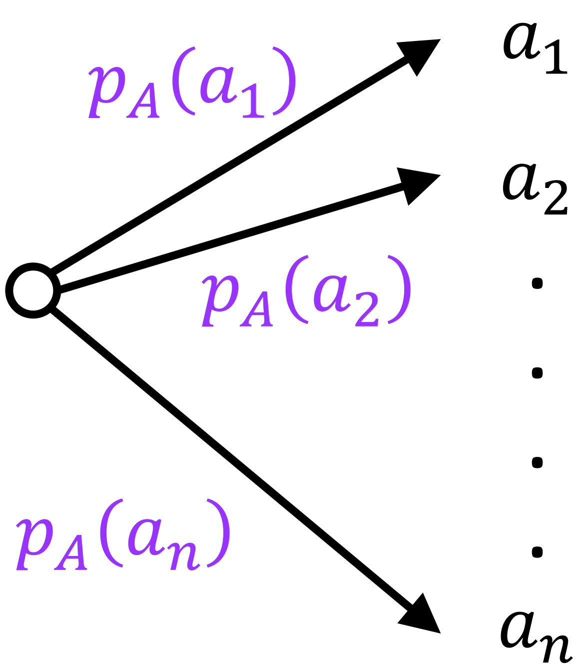
Normalize

Z = 50

T	W	P(t,w)
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

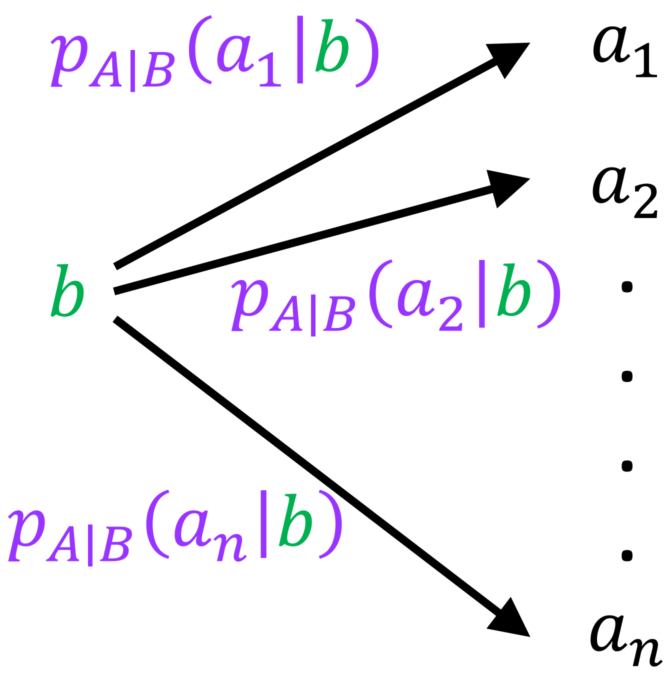
Sum over all values of a discrete random variable

For all possible discrete real values of a random variable A: $a_1, a_2, \ldots a_n$
$\sum_{i=1}^{n} p_A(a_i) = 1.$



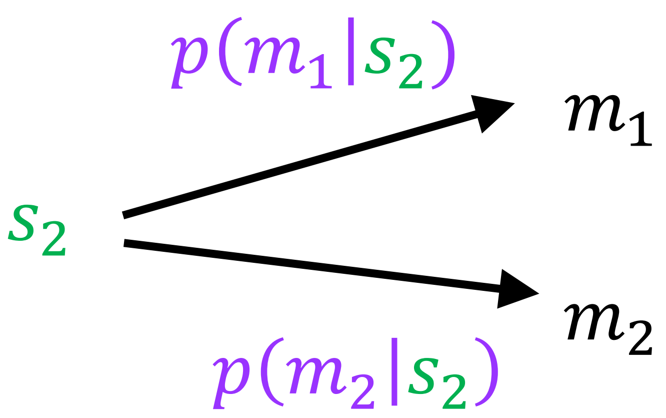
Partition given Event, Still Sums to One

For a given value of random variable $\mathrm{B} = b$ and all possible discrete real values of a random variable A: $a_1, a_2, \ldots a_n$, $\sum_{i=1}^{n} p_{A|B}(a_i \mid b) = 1$:

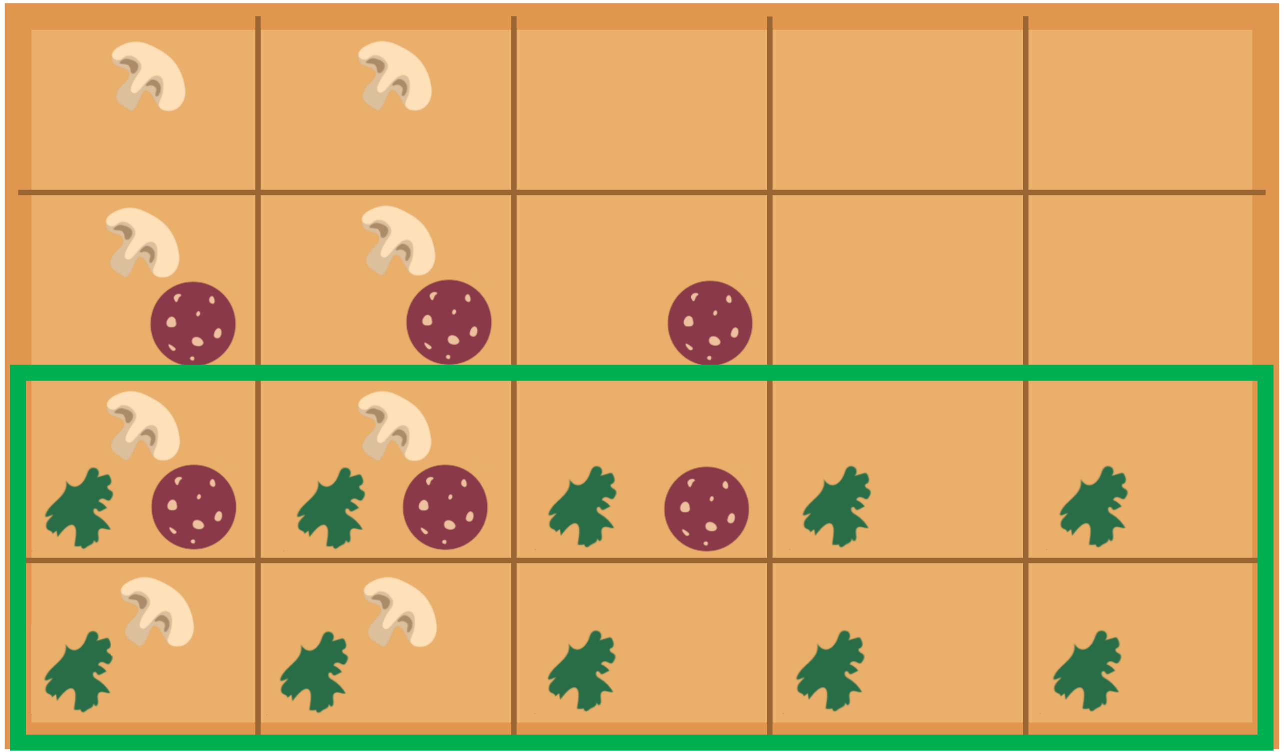


Partition given Event, Still Sums to One

For a given value of random variable $\mathrm{B} = b$ and all possible discrete real values of a random variable A: $a_1, a_2, \ldots a_n$, $\sum_{i=1}^{n} p_{A|B}(a_i \mid b) = 1$:



$$p(m_1|s_2) + p(m_2|s_2)$$



Icons: CC, https://openclipart.org/detail/296791/pizza-slice

Product Rule and Bayes’ Theorem

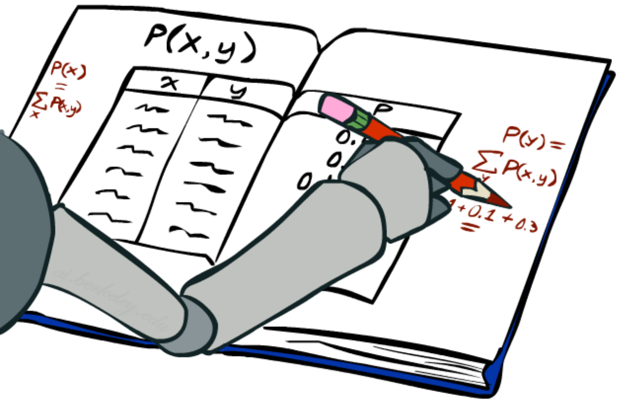
Reformulations of definition of conditional probability

Product rule:

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

Bayes’ theorem:

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

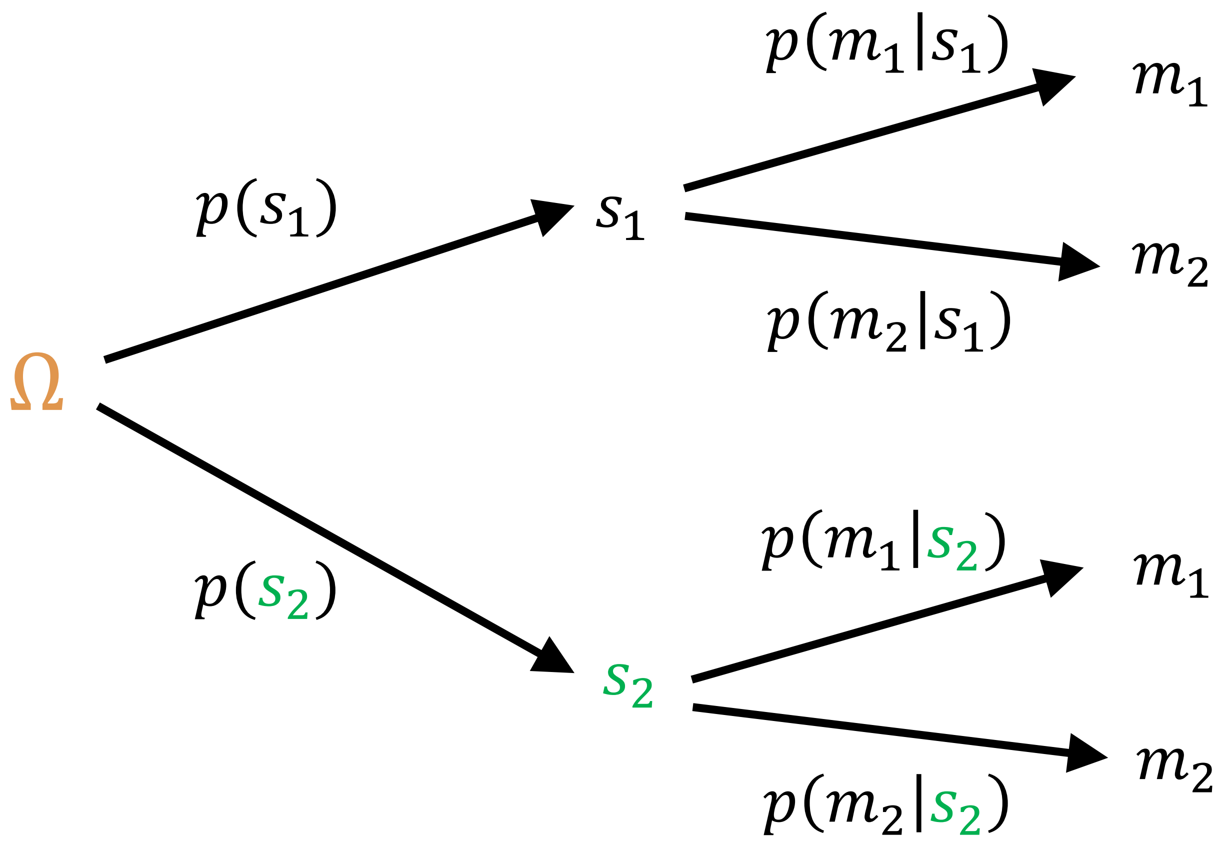


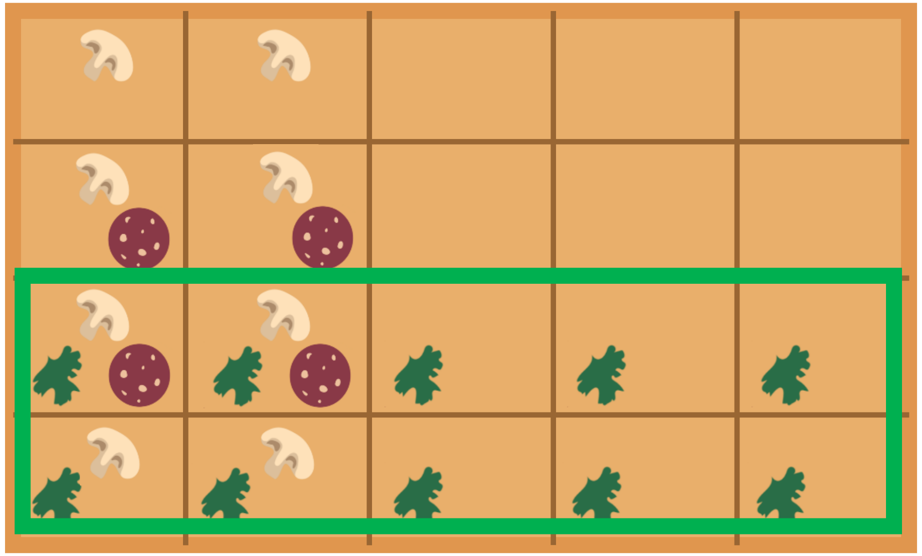
http://achievementgen.com/360/

Product Rule: Tree

Product rule:

$p(a, b) = p(a|b)p(b)$





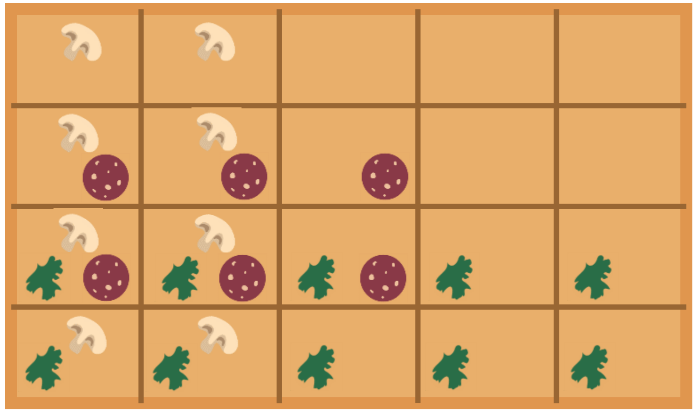
$$p(s_2)\, p(m_1|s_2) = \frac{1}{2} \cdot \frac{6}{10}$$

$$p(m_1, s_2) = \frac{6}{20}$$

Exercise: Product Rule: Tree

Demonstrate, using trees, that product rule works both ways:

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



Icons: CC, https://openclipart.org/detail/296791/pizza-slice

Bayes' Theorem

Bayes' theorem:

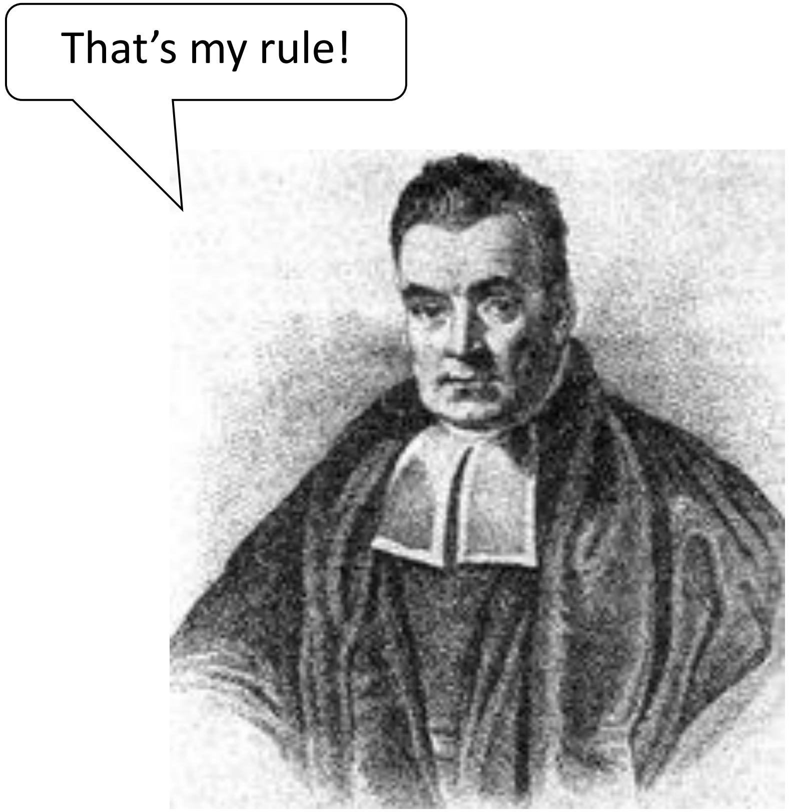
$$P(a_1|b) = \frac{P(b|a_1)P(a_1)}{P(b)}$$

Also:

$$P(a_1|b) = \frac{P(b|a_1)P(a_1)}{\sum_{i=1}^{n} P(b|a_i)P(a_i)}$$

Why is this at all helpful?

- Lets us build one conditional from its reverse
- Often one conditional is tricky but the other one is simple
- Describes an "update" step from prior $P(a)$ to posterior $P(a \mid b)$
- Foundation of many probabilistic systems



Inference with Bayes' Theorem

Example: Diagnostic probability from *causal probability:*

$$P(cause \mid effect) = \frac{P(effect \mid cause)\, P(cause)}{P(effect)}$$

Example:

- Your friend has a stiff neck ($+s$)
- Knowledge:

$$P(+s) = 0.01$$
$$P(+m) = 0.0001$$
$$P(+s \mid +m) = 0.8$$

$$P(+m \mid +s) = \frac{P(+s|+m)\, P(+m)}{P(+s)}$$

$$= \frac{0.8 \times 0.0001}{0.01} = 0.008$$

- What are the chances your friend has meningitis ($+m$)?

Tools Summary

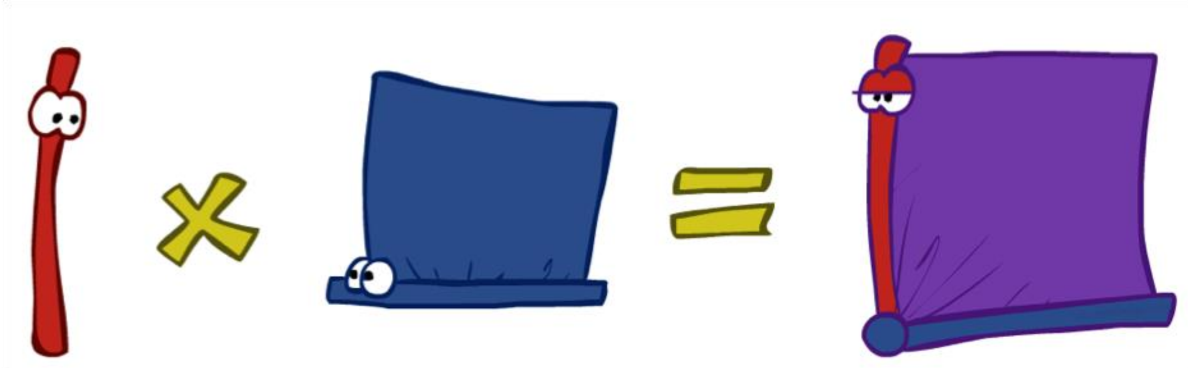
Adding to our toolbox

1. Definition of conditional probability $$P(A|B) = \frac{P(A,B)}{P(B)}$$
2. Product Rule $$P(A,B) = P(A|B)P(B)$$
3. Bayes' theorem $$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$
4. Chain Rule…

The Product Rule

Sometimes have conditional distributions but want the joint

$$P(y)P(x|y) = P(x, y) \qquad \Longleftrightarrow \qquad P(x|y) = \frac{P(x,y)}{P(y)}$$



The Product Rule

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

Example:

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

$P(D,W)$

D	W	P
wet	sun	
dry	sun	
wet	rain	
dry	rain	

The Chain Rule

More generally, can always write any joint distribution as an incremental product of conditional distributions

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

$$P(x_1, x_2, \ldots x_n) = \prod_i P(x_i|x_1 \ldots x_{i-1})$$