COMP9418: Advanced Topics in Statistical Machine Learning

Probability Calculus

Instructor: Gustavo Batista

University of New South Wales

Introduction

- In this lecture, we introduce probability calculus
  - Framework for reasoning with uncertain beliefs
  - Each event is assigned a degree of belief which quantifies the belief in that event

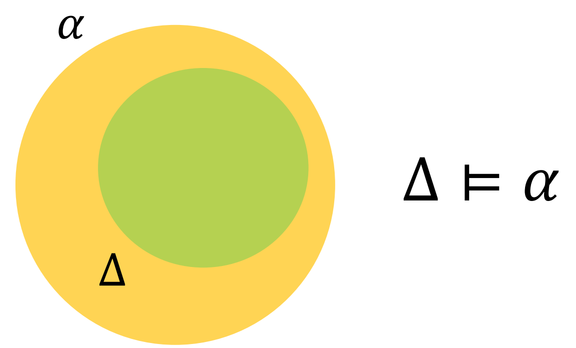
- This topic covered in this lecture will be fundamental to build a framework for reasoning with uncertainty
  - Degree of belief updating
  - The notion of independence

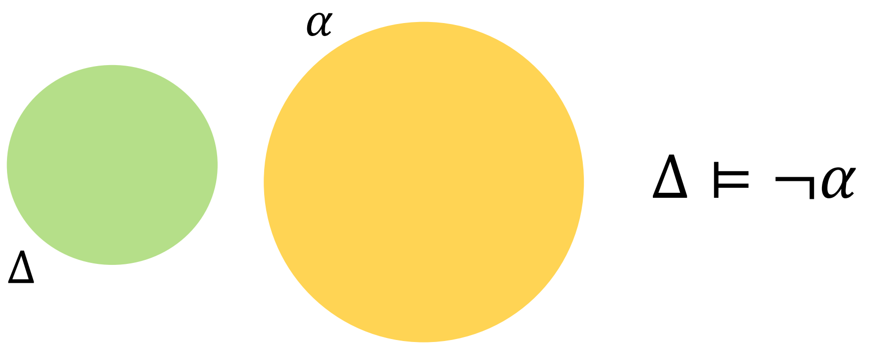
Degrees of Belief

- A propositional knowledge base Δ classifies sentences into three categories
  - Sentences implied by Δ
  - Sentences whose negations are implied by Δ
  - All other sentences
- We can obtain a much finer classification of sentences through a finer classification of worlds
  - Assigning a *degree of belief* or *probability* in $[0,1]$ to each world w ($P(w)$)
  - The belief in, or probability of, a sentence α is defined as

$$P(\alpha) \stackrel{\text{def}}{=} \sum_{w \models \alpha} P(w)$$

  - That is, the sum of probabilities assigned to worlds at which α is true





Degrees of Belief

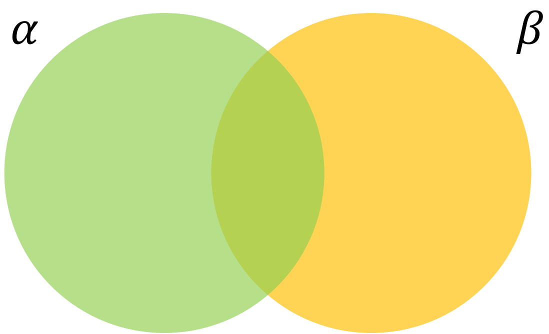
- This table shows a *state of belief* or *joint probability distribution*
  - We require the degrees of belief assigned to all words sum up to 1
  - This is a normalization convention
  - It allows to directly compare the degrees of belief held by different states
- The joint probability distribution is usually too large for direct representation
  - 20 binary variables – 1,048,576 entries
  - 40 binary variables – 1,099,511,627,776 entries
  - We will deal with this issue by using Graphical Models

world	Earthquake	Burglary	Alarm	$P(\cdot)$
w_1	true	true	true	.0190
w_2	true	true	false	.0010
w_3	true	false	true	.0560
w_4	true	false	false	.0240
w_5	false	true	true	.1620
w_6	false	true	false	.0180
w_7	false	false	true	.0072
w_8	false	false	false	.7128

$$P(Earthquake) = P(w_1) + P(w_2) + P(w_3) + P(w_4) = .1$$
$$P(Burglary) = .2$$
$$P(\neg Burglary) = .8$$
$$P(Alarm) = .2442$$

Properties of Beliefs

- Some degrees of belief (or simply, beliefs) properties
  - Beliefs are bounded for any sentence α into the [0,1] interval — $0 \leq \mathrm{P}(\alpha) \leq 1$
  - An inconsistent sentence α have belief equal to zero — $\mathrm{P}(\alpha) = 0$ when α is inconsistent
  - A valid sentence α has belief equal to one — $\mathrm{P}(\alpha) = 1$ when α is valid
  - We can compute the belief of a sentence given the belief in its negation — $\mathrm{P}(\alpha) + \mathrm{P}(\neg\alpha) = 1$
  - We can also compute the belief in a disjunction — $\mathrm{P}(\alpha \vee \beta) = \mathrm{P}(\alpha) + P(\beta) - P(\alpha \wedge \beta)$

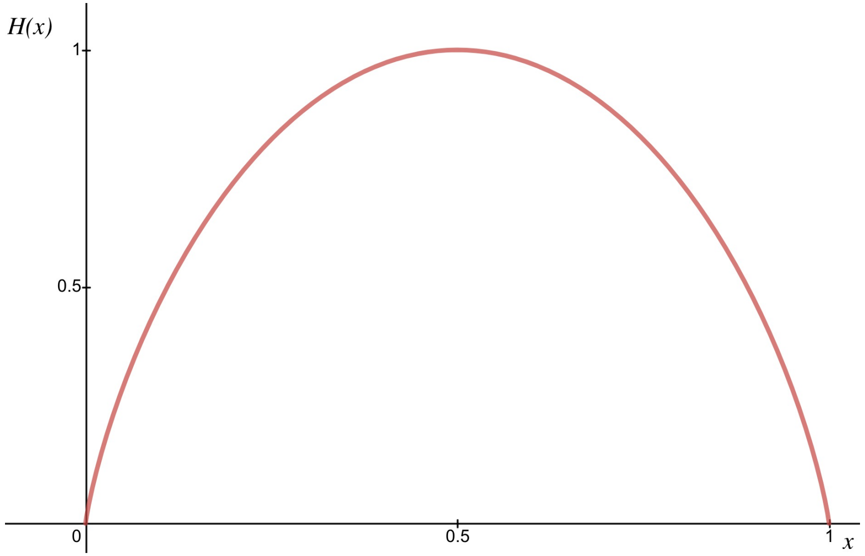


Quantifying Uncertainty

- This tables summarises the beliefs associated with the alarm example
  - The beliefs seem most certain that an Earthquake has occurred
  - We are least certain if an alarm has triggered
- We can quantify uncertainty using *entropy*
  - Where $0 \log 0 = 0$ by convention
- This plot shows the entropy for different probabilities
  - When $p = 0$ or $p = 1$, the entropy is zero, indicating no uncertainty
  - When $p = \frac{1}{2}$, the entropy is at a maximum value

	Earthquake	Burglary	Alarm
true	.1	.2	.2442
false	.9	.8	.7558
$H(\cdot)$	.469	.722	.802

$$H(X) = -\sum_{x} P(X) log_2 P(X)$$



Updating Beliefs

- Suppose now that we know that the alarm has triggered
  - $Alarm = true$
- We need to accommodate the state of belief to this new piece of information we call *evidence*
  - Evidence is represented by an event, say β
  - Our objective is to update the state of belief $P(\cdot)$ into a new state of belief we will denote $P(\cdot|\beta)$

world	Earthquake	Burglary	Alarm	$P(\cdot)$	$P(\cdot\|Alarm)$
w_1	true	true	true	.0190	
w_2	true	true	false	.0010	
w_3	true	false	true	.0560	
w_4	true	false	false	.0240	
w_5	false	true	true	.1620	
w_6	false	true	false	.0180	
w_7	false	false	true	.0072	
w_8	false	false	false	.7128	

Updating Beliefs

- Given that β is known for sure
  - The new state of belief $P(\cdot\,|\beta)$ should assign a belief 1 to β
  - $P(\beta|\beta) = 1$
- This implies that $P(\neg\beta|\beta) = 0$
  - This implies that every world w that satisfies $\neg\beta$ must be assigned to belief 0

$$P(w|\beta) = 0 \qquad for\ all\ w \vDash \neg\beta$$

world	Earthquake	Burglary	Alarm	$P(\cdot)$	$P(\cdot\,|Alarm)$
w_1	true	true	true	.0190	
w_2	true	true	false	.0010	0
w_3	true	false	true	.0560	
w_4	true	false	false	.0240	0
w_5	false	true	true	.1620	
w_6	false	true	false	.0180	0
w_7	false	false	true	.0072	
w_8	false	false	false	.7128	0

Updating Beliefs

- We know that

$$\sum_{w \models \beta} P(w|\beta) = 1$$

  - But these leaves many options for the $P(w|\beta)$ for w's that satisfies β

- It is reasonable to perturb the beliefs as little as possible, so
  - $P(w|\beta) = 0$ for all w where $P(w) = 0$
  - $\frac{P(w)}{P(w')} = \frac{P(w|\beta)}{P(w'|\beta)}$ for all $w, w' \models \beta, P(w) > 0, P(w') > 0$

- With these three constraints, the only options for the new beliefs is
  - $P(w|\beta) = \frac{P(w)}{P(\beta)}$ for all $w \models \beta$

Updating Beliefs

- The new beliefs are just the normalization of old beliefs
  - The normalization constant is $P(\beta)$

$$P(w|\beta) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } w \models \neg\beta \\ \dfrac{P(w)}{P(\beta)} & \text{if } w \models \beta \end{cases}$$

- The new state of belief is referred as *conditioning* the old state P on evidence β

world	Earthquake	Burglary	Alarm	$P(\cdot)$	$P(\cdot\|Alarm)$
w_1	true	true	true	.0190	.0190/.2442
w_2	true	true	false	.0010	0
w_3	true	false	true	.0560	.0560/.2442
w_4	true	false	false	.0240	0
w_5	false	true	true	.1620	.1620/.2442
w_6	false	true	false	.0180	0
w_7	false	false	true	.0072	.0072/.2442
w_8	false	false	false	.7128	0

Updating Beliefs: Example

- Let us suppose $Alarm = true$ and compute some conditional probabilities
  - $P(Burglary) =$
  - $P(Burglary|Alarm) =$
- Also,
  - $P(Earthquake) =$
  - $P(Earthquake|Alarm) =$

world	Earthquake	Burglary	Alarm	$P(\cdot)$	$P(\cdot\|Alarm)$
w_1	true	true	true	.0190	.0778
w_2	true	true	false	.0010	0
w_3	true	false	true	.0560	.2293
w_4	true	false	false	.0240	0
w_5	false	true	true	.1620	.6634
w_6	false	true	false	.0180	0
w_7	false	false	true	.0072	.0295
w_8	false	false	false	.7128	0

Updating Beliefs: Example

- Let us suppose $Alarm = true$ and compute some conditional probabilities
  - $P(Burglary) = .2$
  - $P(Burglary|Alarm) = .741$
- Also,
  - $P(Earthquake) = .1$
  - $P(Earthquake|Alarm) = .307$

world	Earthquake	Burglary	Alarm	$P(\cdot)$	$P(\cdot\|Alarm)$
w_1	true	true	true	.0190	.0778
w_2	true	true	false	.0010	0
w_3	true	false	true	.0560	.2293
w_4	true	false	false	.0240	0
w_5	false	true	true	.1620	.6634
w_6	false	true	false	.0180	0
w_7	false	false	true	.0072	.0295
w_8	false	false	false	.7128	0

Updating Beliefs: Closed Form

- There is a simple closed form to compute the updated belief

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

$$= \sum_{w \models \alpha} P(w|\beta)$$

$$= \sum_{w \models \alpha, w \models \beta} P(w|\beta) + \sum_{w \models \alpha, w \models \neg\beta} P(w|\beta)$$

$$= \sum_{w \models \alpha, w \models \beta} P(w|\beta)$$

$$= \sum_{w \models \alpha \wedge \beta} P(w|\beta)$$

$$= \sum_{w \models \alpha \wedge \beta} P(w)/P(\beta)$$

$$= \frac{1}{P(\beta)} \sum_{w \models \alpha \wedge \beta} P(w)$$

$$= \frac{P(\alpha, \beta)}{P(\beta)}$$

$$P(\alpha|\beta) = \frac{P(\alpha \wedge \beta)}{P(\beta)}$$

This equation is known as *Bayes' conditioning*. It is only defined when $P(\beta) \neq 0$

Bayes Conditioning

- Bayes conditioning follows the following commitments:
  - Worlds that contradict the evidence β will have zero probability
  - Worlds that have zero probability continue to have zero probability
  - Worlds that are consistent with evidence β and have positive probability will maintain their relative beliefs

Earthquake	Burglary	Alarm	$P(\cdot)$
true	true	true	.0190
true	true	false	.0010
true	false	true	.0560
true	false	false	.0240
false	true	true	.1620
false	true	false	.0180
false	false	true	.0072
false	false	false	.7128

	Earthquake	Burglary	Alarm
true	.1	.2	.2442
false	.9	.8	.7558

Bayes Conditioning: Example

- Let us analyse how some beliefs would change given evidence $Earthquake$
  - $P(Burglary) =$
  - $P(Burglary|Earthquake) =$
- Also,
  - $P(Alarm) =$
  - $P(Alarm|Earthquake) =$
- Now, considering evidence $Burglary$
  - $P(Alarm) =$
  - $P(Alarm|Burglary) =$
- Also,
  - $P(Earthquake) =$
  - $P(Earthquake|Burglary) =$

Earthquake	Burglary	Alarm	$P(\cdot)$
true	true	true	.0190
true	true	false	.0010
true	false	true	.0560
true	false	false	.0240
false	true	true	.1620
false	true	false	.0180
false	false	true	.0072
false	false	false	.7128

	Earthquake	Burglary	Alarm
true	.1	.2	.2442
false	.9	.8	.7558

Bayes Conditioning: Example

- Let us analyse how some beliefs would change given evidence $Earthquake$
  - $P(Burglary) = .2$
  - $P(Burglary|Earthquake) = .2$
- Also,
  - $P(Alarm) = .2442$
  - $P(Alarm|Earthquake) \approx .75$
- Now, considering evidence $Burglary$
  - $P(Alarm) = .2442$
  - $P(Alarm|Burglary) \approx .905$
- Also,
  - $P(Earthquake) = .1$
  - $P(Earthquake|Burglary) = .1$

Earthquake	Burglary	Alarm	$P(\cdot)$
true	true	true	.0190
true	true	false	.0010
true	false	true	.0560
true	false	false	.0240
false	true	true	.1620
false	true	false	.0180
false	false	true	.0072
false	false	false	.7128

	Earthquake	Burglary	Alarm
true	.1	.2	.2442
false	.9	.8	.7558

Bayes Conditioning: Example

- These beliefs dynamics are a property of the state of belief in this table
  - It may not hold for other states of belief
  - For instance, we can conceive a state of belief in which evidence $Earthquake$ would change the belief about $Burglary$
- A central question is synthetizing states of beliefs that are *faithful*
  - For instance, those that correspond to the beliefs held by some human expert
  - We will see this is a central issue of modelling a real problem

Earthquake	Burglary	Alarm	$P(\cdot)$
true	true	true	.0190
true	true	false	.0010
true	false	true	.0560
true	false	false	.0240
false	true	true	.1620
false	true	false	.0180
false	false	true	.0072
false	false	false	.7128

	Earthquake	Burglary	Alarm
true	.1	.2	.2442
false	.9	.8	.7558

Bayes Conditioning: Example

- Let us look at one more example as we add more evidence
  - $P(Burglary|Alarm) =$
  - $P(Burglary|Alarm \wedge Earthquake) =$
- Also,
  - $P(Burglary|Alarm \wedge \neg Earthquake) =$

Earthquake	Burglary	Alarm	$P(\cdot)$
true	true	true	.0190
true	true	false	.0010
true	false	true	.0560
true	false	false	.0240
false	true	true	.1620
false	true	false	.0180
false	false	true	.0072
false	false	false	.7128

	Earthquake	Burglary	Alarm
true	.1	.2	.2442
false	.9	.8	.7558

Bayes Conditioning: Example

- Let us look at one more example as we add more evidence
  - $P(Burglary|Alarm) \approx .741$
  - $P(Burglary|Alarm \wedge Earthquake) \approx .253$
- Also,
  - $P(Burglary|Alarm \wedge \neg Earthquake) \approx .957$

Earthquake	Burglary	Alarm	$P(\cdot)$
true	true	true	.0190
true	true	false	.0010
true	false	true	.0560
true	false	false	.0240
false	true	true	.1620
false	true	false	.0180
false	false	true	.0072
false	false	false	.7128

	Earthquake	Burglary	Alarm
true	.1	.2	.2442
false	.9	.8	.7558

Conditional Entropy

- The *conditional entropy* of a variable X given another variable Y
  - It quantifies the average uncertainty about X after observing Y

$$H(X|Y) \stackrel{\text{def}}{=} \sum_{y} P(y)H(X|y)$$

$$H(X|y) \stackrel{\text{def}}{=} -\sum_{x} P(x|y)log_2 P(x|y)$$

- We can show that the entropy never increases after conditioning

$$H(X|Y) \leq H(X)$$

  - Although for a specific value y we may have $H(X|y) > H(X)$

	B	$B\|a$	$B\|\neg a$
true	.2	.741	.025
false	.8	.259	.975
$H(\cdot)$	.722	.825	.169

Conditional Entropy

- The *conditional entropy* of a variable X given another variable Y
  - It quantifies the average uncertainty about X after observing Y

$$H(X|Y) \stackrel{\text{def}}{=} \sum_{y} P(y) H(X|y)$$

$$H(X|y) \stackrel{\text{def}}{=} -\sum_{x} P(x|y) log_2 P(x|y)$$

- We can show that the entropy never increases after conditioning

$$H(X|Y) \leq H(X)$$

  - Although for a specific value y we may have $H(X|y) > H(X)$

	B	$B\|a$	$B\|\neg a$
true	.2	.741	.025
false	.8	.259	.975
$H(\cdot)$	.722	.825	.169

  - $H(B|A) = H(B|a)P(a) + H(B|\neg a)P(\neg a)$
    $= .329$

Independence

- We observed that evidence $Burglary$ does not change belief in $Earthquake$
- More generally, we say that event α is independent of event β iff
  - $P(\alpha|\beta) = P(\alpha)$ or
  - $P(\beta) = 0$
- We also found $Burglary$ independent of $Earthquake$
- It is indeed a general property
  - If α is independent of β then β is independent of α
- P finds α and β are independent iff
  - $P(\alpha \wedge \beta) = P(\alpha)P(\beta)$
  - This equation is often taken as the definition of independence

$P(Earthquake) = .1$

$P(Earthquake|Burglary) = .1$

$P(Burglary) = .2$

$P(Burglary|Earthquake) = .2$

Independence

- We observed that evidence $Burglary$ does not change belief in $Earthquake$
- More generally, we say that event α is independent of event β iff
  - $P(\alpha|\beta) = P(\alpha)$ or
  - $P(\beta) = 0$
- We also found $Burglary$ independent of $Earthquake$
- It is indeed a general property
  - If α is independent of β then β is independent of α
- P finds α and β are independent iff
  - $P(\alpha \wedge \beta) = P(\alpha)P(\beta)$
  - This equation is often taken as the definition of independence

$P(Earthquake) = .1$

$P(Earthquake|Burglary) = .1$

$P(Burglary) = .2$

$P(Burglary|Earthquake) = .2$

Independence and Mutual Exclusiveness

- Independence and mutual exclusiveness are *not* the same notion
- Two events α and β are mutually exclusive (logically disjoint) iff they do not share any models
  - $Mods(\alpha) \cap Mods(\beta) = \emptyset$
  - They cannot hold together at the same world
- Events α and β are independent iff $P(\alpha \wedge \beta) = P(\alpha)P(\beta)$

Conditional Independence

- Independence is a dynamic notion
  - Two independent events may become dependent after some evidence
  - For example, we saw that Burglary was independent of Earthquake. However, these events are dependent after accepting evidence Alarm
  - This is expected since $Earthquake$ and $Burglary$ are competing explanations to $Alarm$
- Consider this table for another example
  - We have two sensors that can detect the current state of temperature
  - The sensors are noisy and have different reliabilities

$P(Burglary|Alarm) \approx .741$
$P(Burglary|Alarm \wedge Earthquake) \approx .253$

world	Temp	Sensor1	Sensor2	$P(\cdot)$
w_1	normal	normal	normal	.576
w_2	normal	normal	extreme	.144
w_3	normal	extreme	normal	.064
w_4	normal	extreme	extreme	.016
w_5	extreme	normal	normal	.008
w_6	extreme	normal	extreme	.032
w_7	extreme	extreme	normal	.032
w_8	extreme	extreme	extreme	.128

Conditional Independence

- We have the following initial beliefs
  - $P(Temp = normal) = .80$
  - $P(Sensor1 = normal) = .76$
  - $P(Sensor2 = normal) = .68$
- If the first sensor reads normal
  - Our belief that the second sensor also reads normal increases
  - $P(Sensor2 = normal | Sensor1 = normal) \approx .768$
  - Therefore the sensors readings are dependent
- However, they will become independent if we observe the temperature is normal
  - $P(Sensor2 = normal | Temp = normal) = .80$
  - $P(Sensor2 = normal | Temp = normal, Sensor1 = normal) = .80$

world	Temp	Sensor1	Sensor2	$P(\cdot)$
w_1	normal	normal	normal	.576
w_2	normal	normal	extreme	.144
w_3	normal	extreme	normal	.064
w_4	normal	extreme	extreme	.016
w_5	extreme	normal	normal	.008
w_6	extreme	normal	extreme	.032
w_7	extreme	extreme	normal	.032
w_8	extreme	extreme	extreme	.128

Conditional Independence

- In general,
  - Independent event may become dependent given new evidence
  - Dependent events may become independent given new evidence

- Event α is conditionally independent of event β given event γ iff

  $P(\alpha|\beta \wedge \gamma) = P(\alpha|\gamma)$ or

  $P(\gamma) = 0$
- Conditional independence is also symmetric
  - α is conditionally independent of β given γ iff β is conditionally independent of α given γ
- This equation is often taken as definition of conditional independence

  $P(\alpha \wedge \beta|\gamma) = P(\alpha|\gamma)P(\beta|\gamma)$ or

  $P(\gamma) = 0$

Variable Independence

- It is useful to talk about independence of sets of variables
  - Let $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ be three disjoint sets of variables
- $\boldsymbol{X}$ is independent of $\boldsymbol{Y}$ given $\boldsymbol{Z}$ is denoted by
  - It means that $\boldsymbol{x}$ is independent of $\boldsymbol{y}$ given $\boldsymbol{z}$ for all instantiations of $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$
- For example, suppose that $\boldsymbol{X} = \{A, B\}$, $\boldsymbol{Y} = \{C\}$ and $\boldsymbol{Z} = \{D, E\}$
  - Where A, B, C, D and E are propositional variables
  - The statement $\boldsymbol{X} \perp \boldsymbol{Y}|\boldsymbol{Z}$ is a compact notation for several statements about independence

$\boldsymbol{X} \perp \boldsymbol{Y}|\boldsymbol{Z}$

$I_P(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$

$A \wedge$ B is independent of C given $D \wedge$ E
$A \wedge \neg$B is independent of C given $D \wedge$ E
$\neg A \wedge \neg$B is independent of C given $D \wedge$ E
...
$\neg A \wedge \neg$B is independent of $\neg C$ given $\neg D \wedge \neg$E

Mutual Information

- Independence is a special case of a more general notion known as *mutual information*
  - Mutual information quantifies the impact of observing one variable on the uncertainty in another

$$MI(X;Y) \stackrel{\text{def}}{=} \sum_{x,y} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- Mutual information is
  - Non-negative
  - Equal to zero only if X and Y are independent
  - It measures the extent to which observing one variable will reduce the uncertainty in another

$$MI(X;Y) = H(X) - H(X|Y)$$

$$MI(X;Y) = H(Y) - H(Y|X)$$

Conditional Mutual Information

- Conditional mutual information can be defined as

$$MI(X;Y|Z) \overset{\text{def}}{=} \sum_{x,y,z} P(x,y,z) \log_2 \frac{P(x,y|z)}{P(x|z)P(y|z)}$$

- It has the following properties

$$MI(X;Y|Z) = H(X|Z) - H(X|Y,Z)$$

$$MI(X;Y|Z) = H(Y|Z) - H(Y|X,Z)$$

- Entropy and mutual information can be extended to sets of variables
  - For instance, entropy can be generalized to a set of variables $\boldsymbol{X}$ as follows

$$H(\boldsymbol{X}) = -\sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log_2 P(\boldsymbol{x})$$

Conditional Probability for Multiple Variables

- We can extend the definition for conditional probabilities for multiple variables

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

Bayes' conditioning

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

Product rule

- For three variables A, B and C, we have

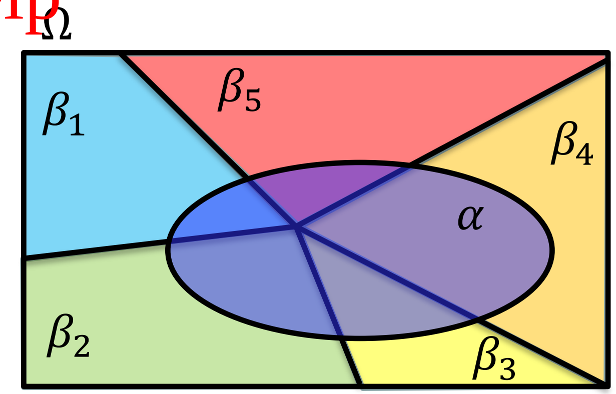
$$P(A,B|C) = \frac{P(A,B,C)}{P(C)}$$

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

Case Analysis

- Also known as *law of total probability*

$$P(\alpha) = \sum_{i=1}^{n} P\,(\alpha \wedge \beta_i)$$



The events $\beta_1, \dots, \beta_n$ are mutually exclusive and exhaustive

Chain Rule

- It is the repeated application of Bayes conditioning

$$P(\alpha_1 \wedge \alpha_2 \wedge \cdots \wedge \alpha_n) = P(\alpha_1 | \alpha_2 \wedge \cdots \wedge \alpha_n) P(\alpha_2 | \alpha_3 \wedge \cdots \wedge \alpha_n) \ldots P(\alpha_n)$$

Bayes Rule

- *Bayes rule* or *Bayes theorem*
  - The classical usage of this rule is when event α is perceived to be a cause of event β
  - For example, α is a disease and β is a symptom
  - Our goal is to assess our belief in the cause given the effect

$$P(\alpha|\beta) = \frac{P(\beta|\alpha)P(\alpha)}{P(\beta)}$$

- Example
  - Suppose that we have a patient who was just tested for a particular disease and the test came out positive. We know that one in every thousand people has this disease. We also know that the test is not reliable: it has a false positive rate of 2% and a false negative rate of 5%. Our goal is then to assess our belief in the patient having the disease given that the test came out positive.

Conclusion

- In this lecture, we discussed some fundamental aspects of probabilistic calculus
  - Belief update
  - Independence and Conditional Independence
  - Bayes conditioning, case analysis,
  - Chain rule, Bayes rule

- Tasks
  - Read Chapter 3, but Sections 3.6 and 3.7 from the textbook (Darwiche)