2.3. Probabilities

Lecture based on **https://github.com/gwthomas/math4ml** (Garrett Thomas, 2018)

Prof. Dr. Christoph Lippert

Digital Health & Machine Learning

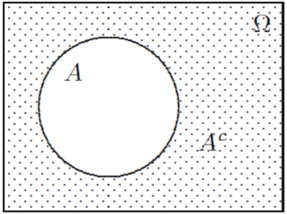
Why probabilities?

- Inferences from data are intrinsically uncertain.
- Probability theory: model uncertainty instead of ignoring it!
- Applications in Statistics, Machine Learning, Data Mining, Pattern Recognition, etc.
- Goal of this part of the course
    - Basic probability theory
    - estimation
    - probabilistic modeling

Sample space

Suppose we have some sort of randomized experiment (e.g. a coin toss, die roll) that has a fixed set of possible **outcomes** ω.
This set is called the **sample space** and denoted Ω.

- We define probabilities for some **events**, which are subsets of Ω.
- The set of events is denoted $\mathcal{F}$.
- The **complement** of the event A is another event, $A^{\mathsf{c}} = \Omega \setminus A$.

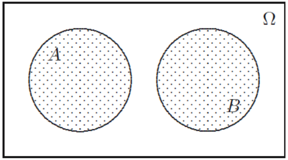


Then we can define a **probability measure** $P : \mathcal{F} \to [0, 1]$ which must satisfy

ⓘ $P(\Omega) = 1$

ⓘⓘ **Countable additivity**: for any countable collection of **disjoint** sets $\{A_i\} \subseteq \mathcal{F}$,

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$$



The triple $(\Omega, \mathcal{F}, P)$ is called a **probability space**.

Proposition

Let A be an event.
Then

(i) $P(A^c) = 1 - P(A)$.

(ii) *If A is an event and $A \subseteq B$, then $P(A) \leq P(B)$.*

(iii) $0 = P(\varnothing) \leq P(A) \leq P(\Omega) = 1$

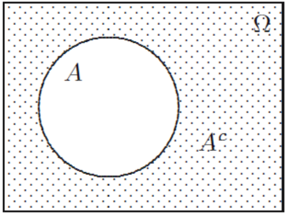
Proof.

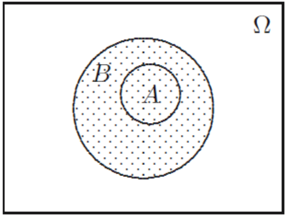
(i) Using the countable additivity of P, we have
$P(A) + P(A^c) = P(A \dot{\cup} A^c) = P(\Omega) = 1$

(ii) suppose $A \in \mathcal{F}$ and $A \subseteq B$.
Then $P(B) = P(A \dot{\cup} (B \setminus A)) = P(A) + P(B \setminus A) \geq P(A)$.

(iii) the middle inequality follows from (ii) since $\varnothing \subseteq A \subseteq \Omega$.
We also have $P(\varnothing) = P(\varnothing \dot{\cup} \varnothing) = P(\varnothing) + P(\varnothing)$ by countable additivity, which shows $P(\varnothing) = 0$.

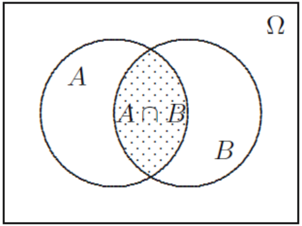
$\square$





Proposition

If A and B are events, then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.



Proof.

The key is to break the events up into their various overlapping and non-overlapping parts.

$$\begin{aligned} P(A \cup B) &= P((A \cap B) \mathbin{\dot\cup} (A \setminus B) \mathbin{\dot\cup} (B \setminus A)) \\ &= P(A \cap B) + P(A \setminus B) + P(B \setminus A) \\ &= P(A \cap B) + P(A) - P(A \cap B) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

$\square$

Conditional Probability and Chain Rule

- The **conditional probability** of event A given that event B has occurred is written $P(A|B)$ and defined as

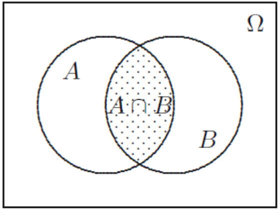
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

assuming $P(B) > 0$.

- Another very useful tool, the **chain rule**, follows from this definition:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

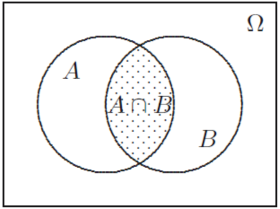
.



Bayes' rule

Taking the equality from above one step further, we arrive at the simple but crucial **Bayes' rule**:

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



It is sometimes beneficial to omit the normalizing constant and write

$$P(A|B) \propto P(A)P(B|A)$$

Under this formulation, $P(A)$ is often referred to as the **prior**, $P(A|B)$ as the **posterior**, and $P(B|A)$ as the **likelihood**.
In the context of machine learning, we can use Bayes' rule to update our "beliefs" (e.g. values of our model parameters) given some data that we've observed.