# Lecture 4: Probability Theory and Probabilistic Modeling

**COMP90049**
**Introduction to Machine Learning**

Semester 1, 2021

Lea Frermann, CIS

# Roadmap

**Last time... Concepts and KNN classification**

- data, features, classes
- K Nearest Neighbors algorithm
- Application to classification

**Today... Probability**

- basics / refresher
- distributions and parameterizations
- why probability in ML?

Estimating confidence in different possible outcomes

# Probability Theory

"The calculus of probability theory provides us with a **formal framework** for considering multiple possible **outcomes** and their **likelihood**. It defines a set of **mutually exclusive** and **exhaustive** possibilities, and associates each of them with a probability — **a number between 0 and 1**, so that the **total probability of all possibilities is 1**. This framework allows us to consider options that are **unlikely, yet not impossible**, without reducing our conclusions to content-free lists of every possibility."

From Probabilistic Graphical Models: Principles and Techniques (2009; Koller and Friedman) `http://pgm.stanford.edu/intro.pdf`

**P**(**A**): **the probability of A** the fraction of times the event A is true in independent trials

$$0 <= P(A) <= 1$$

$$P(True) = 1$$

$$P(False) = 0$$

# (Very) Basics of Probability Theory

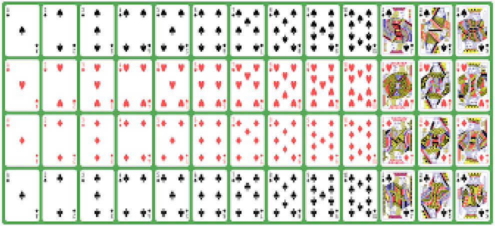
**P(A):** **the probability of A** the fraction of times the event A is true in independent trials

$$0 <= P(A) <= 1 \qquad P(True) = 1$$

$$P(False) = 0$$

**Given a deck of 52 cards**

- 13 ranks (ace, 2-10, jack, queen, king)
- of each of four suits (clubs, spades = black; hearts, diamonds = red)



https://commons.wikimedia.org/wiki/File:English_pattern_playing_cards_deck.svg

# (Very) Basics of Probability Theory

**P(A):** **the probability of A** the fraction of times the event A is true in independent trials

$$0 <= P(A) <= 1 \qquad P(True) = 1 \qquad P(False) = 0$$

**Given a deck of 52 cards**

- 13 ranks (ace, 2-10, jack, queen, king)
- of each of four suits (clubs, spades = black; hearts, diamonds = red)

$$P(\texttt{queen}) = ?\ \frac{1}{13} \qquad P(\texttt{red}) = ?\ \frac{1}{2}$$

$$P(\texttt{heart}) = ?\ \frac{1}{4}$$

# (Very) Basics of Probability Theory

**P(A)**: **the probability of A** — the fraction of times the event A is true in independent trials

$$0 <= P(A) <= 1$$

$$P(True) = 1$$

$$P(False) = 0$$

**Given a deck of 52 cards**

- 13 ranks (ace, 2-10, jack, queen, king)
- of each of four suits (clubs, spades = black; hearts, diamonds = red)

$$P(\texttt{queen}) = \frac{1}{13}$$

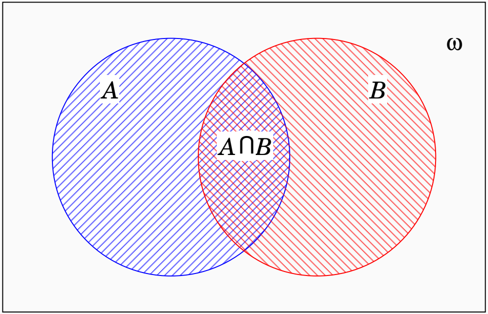
$$P(\texttt{heart}) = \frac{1}{4}$$

$$P(\texttt{red}) = \frac{1}{2}$$

THE UNIVERSITY OF MELBOURNE

# Basics of Probability Theory

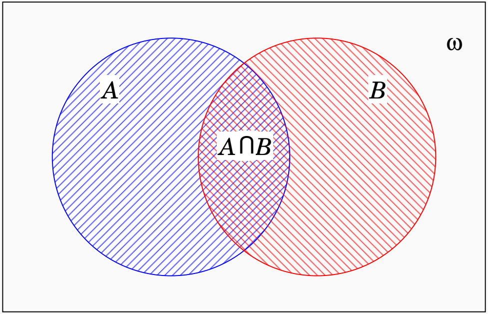
**P(A, B):** **joint probability of A and B** — the probability of both $A$ and $B$ occurring = $P(A \cap B)$




$$P(\texttt{ace}, \texttt{heart}) = ?114$$

$$P(\texttt{heart}, \texttt{red}) = ?114$$

| | | |
|---|---|---|
| $\mathbf{P(A, B)}$: | **joint probability of A and B** | the probability of both $A$ and $B$ occurring = $P(A \cap B)$ |

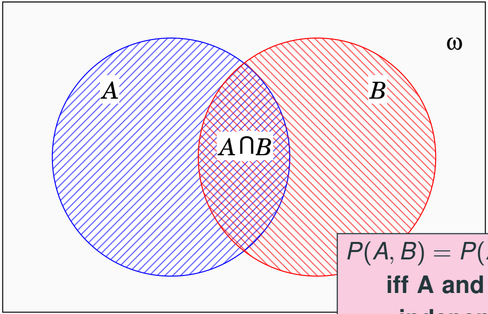



$$P(\mathtt{ace}, \mathtt{heart}) = \frac{1}{52}$$

$$P(\mathtt{heart}, \mathtt{red}) = \frac{1}{4}$$

THE UNIVERSITY OF MELBOURNE

**P(A, B):** **joint probability of A and B** — the probability of both $A$ and $B$ occurring = $P(A \cap B)$



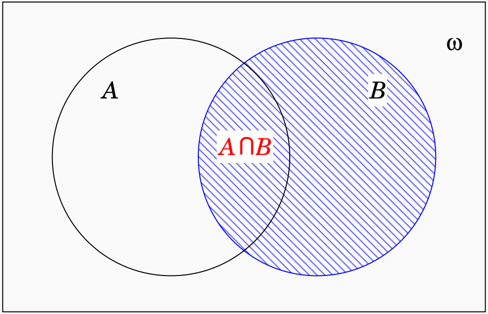

$P(A, B) = P(A) * P(B)$ **iff A and B are independent**

$$P(\texttt{ace}, \texttt{heart}) = \frac{1}{52}$$

$$P(\texttt{heart}, \texttt{red}) = \frac{1}{4}$$

# Conditional Probability

**P(A|B):** **conditional probability** the probability of $A$ given the occurrence of $B = \frac{P(A \cap B)}{P(B)}$

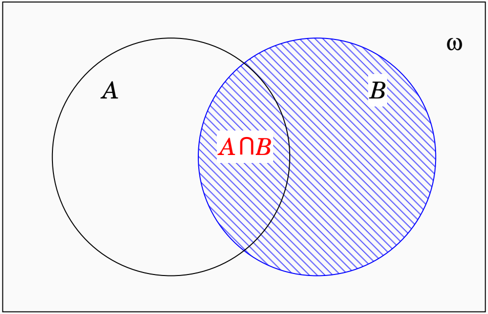



$P(\texttt{ace}|\texttt{heart}) = ?$

$P(\texttt{heart}|\texttt{red}) = ?$

# Conditional Probability

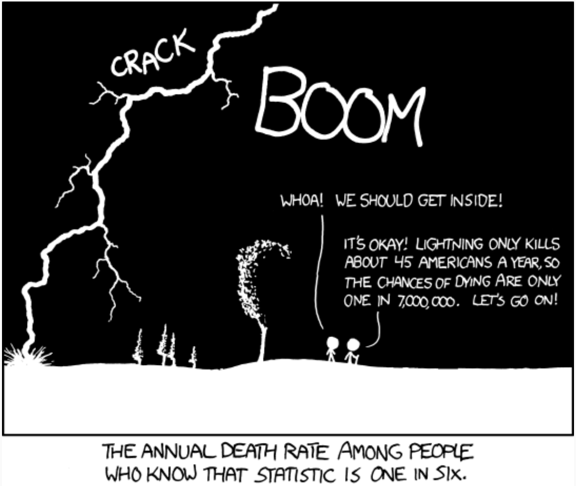
**P(A|B):** **conditional probability** the probability of $A$ given the occurrence of $B = \frac{P(A \cap B)}{P(B)}$




$$P(\texttt{ace}|\texttt{heart}) = \frac{1}{52} / \frac{1}{4} = \frac{1}{13}$$

$$P(\texttt{heart}|\texttt{red}) = \frac{1}{4} / \frac{1}{2} = \frac{1}{2}$$

# What type of probability?




https://imgs.xkcd.com/comics/conditional_risk.png

# Rules of Probability I

- **Independence:** $A$ and $B$ are independent iff $P(A \cap B) = P(A)P(B)$

- **Disjoint events:** The probability of two disjoint events, such that $A \cap B = \emptyset$, is $P(A \text{ or } B) = P(A) + P(B)$

- **Multiplication rule**: $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$

- **Chain rule**:
  $P(A_1 \cap ... \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2 \cap A_1) \; ... \; P(A_n| \cap_{i=1}^{n-1} A_i)$

# Rules of Probability I

- **Independence:** $A$ and $B$ are independent iff $P(A \cap B) = P(A)P(B)$

- **Disjoint events:** The probability of two disjoint events, such that $A \cap B = \emptyset$, is $P(A \text{ or } B) = P(A) +$

e.g., draw an ace or a king: $A$= draw an ace; $B$=draw a king.

- **Multiplication rule**: $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$

- **Chain rule**:
$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2 \cap A_1) \ldots P(A_n| \cap_{i=1}^{n-1} A_i)$

# Rules of Probability I

- **Independence:** $A$ and $B$ are independent iff $P(A \cap B) = P(A)P(B)$

- **Disjoint events:** The probability of two disjoint events, such that $A \cap B = \emptyset$, is $P(A \text{ or } B) = P(A) +$

e.g., draw an ace or a king: $A$= draw an ace; $B$=draw a king.

- **Multiplication rule**: $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$

- **Chain rule**:
$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2 \cap A_1) \ldots P(A_n| \cap_{i=1}^{n-1} A_i)$

**again, we can choose the factorization, e.g., :**

$$P(July, 5^\circ C, sick) = P(July) \times P(5^\circ C|July) \times P(sick|5^\circ C, July)$$

makes sense

$$= P(5^\circ C) \times P(sick|5^\circ C) \times P(July|5^\circ C, sick)$$

???

# Rules of Probability II

**Bayes Rule**

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \qquad \left(\text{cf., } P(A|B) = \frac{P(A \cap B)}{P(B)}\right)$$

**Basic rule of probability**

- Bayes' Rule allows us to compute $P(A|B)$ given knowledge of the 'inverse' probability $P(B|A)$.

**More philosophically,**

- Bayes' Rule allows us to update prior belief with empirical evidence

# Rules of Probability II

**Bayes Rule**

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \qquad \left(\text{cf., } P(A|B) = \frac{P(A \cap B)}{P(B)}\right)$$

**Posterior Probability** $P(A|B)$

- the degree of belief having accounted for $B$.

**Prior Probability** $P(A)$

- the initial degree of belief in $A$.
- the probability of $A$ occurring, given no additional knowledge about $A$

**Likelihood** $P(B|A)$

- the support $B$ provides for $A$

**Normalizing constant ('Evidence')** $P(B) = \sum_A P(B|A)P(A)$

### Bayes Rule

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \qquad \left(\text{cf., } P(A|B) = \frac{P(A \cap B)}{P(B)}\right)$$

### Example

Estimate the probability of a student **being smart** given that (s)he **achieved H1** score, $P(smart|H1)$ from the following information:

$$\begin{aligned} P(Smart) &= 0.3 && \text{prior rate of smart students} \\ P(H1|Smart) &= 0.6 && \text{empirically measured } H1|smart \\ P(H1) &= 0.2 && \text{emprirically measured} \end{aligned}$$

(What if $P(H1) = 0.4$?)

# Binomial Distributions

- A **binomial distribution** results from a
  series of independent trials with only two outcomes (aka **Bernoulli trials**)

  *e.g. multiple coin tosses* ($\langle H, T, H, H, ..., T \rangle$)

# Binomial Distributions

- A **binomial distribution** results from a series of independent trials with only two outcomes (aka **Bernoulli trials**)

  *e.g. multiple coin tosses* ($\langle H, T, H, H, ..., T \rangle$)

- The probability $P$ of an event with probability $p$ occurring exactly $m$ out of $n$ times is given by

$$P(m, n, p) = \binom{n}{m} p^m (1-p)^{n-m}$$

$$P(m, n, p) = \underbrace{\frac{n!}{m!(n-m)!}}_{\substack{\text{possible distributions} \\ \text{of } m \text{ successes} \\ \text{over } n \text{ trials}}} \underbrace{p^m}_{m \text{ successes}} \underbrace{(1-p)^{n-m}}_{n-m \text{ failures}}$$

# Binomial Distributions

- A **binomial distribution** results from a series of independent trials with only two outcomes (aka **Bernoulli trials**)

  *e.g. multiple coin tosses* $(\langle H, T, H, H, ..., T \rangle)$

- The probability $P$ of an event with probability $p$ occurring exactly $m$ out of $n$ times is given by

$$P(m, n, p) \quad = \quad \binom{n}{m} p^m (1-p)^{n-m}$$

$$P(m, n, p) \quad = \quad \underbrace{\frac{n!}{m!(n-m)!}}_{\substack{\text{possible distributions} \\ \text{of } m \text{ successes} \\ \text{over } n \text{ trials}}} \quad \underbrace{p^m}_{m \text{ successes}} \quad \underbrace{(1-p)^{n-m}}_{n-m \text{ failures}}$$

What is the probability of getting times 2 heads out of 3 tosses of a fair coin?

# Binomial Example: Coin Toss

What is the probability of getting times 2 heads out of 3 tosses of a fair coin?

# Binomial Example: Coin Toss

What is the probability of getting times 2 heads out of 3 tosses of a fair coin?

1. $m = 2$ successes (heads) when flipping coin $n = 3$ times; $P(X = 2)$

# Binomial Example: Coin Toss

What is the probability of getting times 2 heads out of 3 tosses of a fair coin?

1. $m = 2$ successes (heads) when flipping coin $n = 3$ times; $P(X = 2)$

2. number of possible outcomes $e$ from 3 coin flips:

$$2 * 2 * 2 = 2^3 = 8 \qquad \text{each with } P(e) = \frac{1}{8}$$

THE UNIVERSITY OF MELBOURNE

# Binomial Example: Coin Toss

What is the probability of getting times 2 heads out of 3 tosses of a fair coin?

1. $m = 2$ successes (heads) when flipping coin $n = 3$ times; $P(X = 2)$
2. number of possible outcomes $e$ from 3 coin flips:

$$2 * 2 * 2 = 2^3 = 8 \qquad \text{each with } P(e) = \frac{1}{8}$$

3. Choose 2 out of 3: $C(3, 2) = \frac{3!}{2!1!} = 3$

# Binomial Example: Coin Toss

What is the probability of getting times 2 heads out of 3 tosses of a fair coin?

1. $m = 2$ successes (heads) when flipping coin $n = 3$ times; $P(X = 2)$

2. number of possible outcomes $e$ from 3 coin flips:

$$2 * 2 * 2 = 2^3 = 8 \qquad \text{each with } P(e) = \frac{1}{8}$$

3. Choose 2 out of 3: $C(3, 2) = \frac{3!}{2!1!} = 3$

4. 3 possible outcomes, $\frac{1}{8}$ for each: $P(X = 2) = \frac{3}{8}$

# Binomial Example: Coin Toss

What is the probability of getting times 2 heads out of 3 tosses of a fair coin?

1. $m = 2$ successes (heads) when flipping coin $n = 3$ times; $P(X = 2)$

2. number of possible outcomes $e$ from 3 coin flips:

$$2 * 2 * 2 = 2^3 = 8 \qquad \text{each with } P(e) = \frac{1}{8}$$

3. Choose 2 o

$$P(m, n, p) = \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m}$$

4. 3 possible outcomes, $\frac{1}{8}$ for each: $P(X = 2) = \frac{3}{8}$

$$P\left(2, 3, \frac{1}{2}\right) = \frac{3!}{2!(3-2)!} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{3-2} = 3 \left(\frac{1}{4}\right) \left(\frac{1}{2}\right)$$

# Multinomial Distributions

- A **multinomial distribution** models the probability of **counts** of different events from a series of independent trials with **more than two possible outcomes**, e.g.,

  - a fair 6-sided dice is rolled 5 times
  - what is the probability of observing exactly 3 ’ones’ and 2 ’fives’?
  - what is the probability of observing 5 ’threes’?

# Multinomial Distributions

- A **multinomial distribution** models the probability of **counts** of different events from a series of independent trials with **more than two possible outcomes**, e.g.,

  - a fair 6-sided dice is rolled 5 times
  - what is the probability of observing exactly 3 'ones' and 2 'fives'?
  - what is the probability of observing 5 'threes'?

- The probability of events $X_1, X_2, ..., X_n$ with probabilities $\mathbf{p} = p_1, p_2, ..., p_n$ occurring exactly $x_1, x_2, ..., x_n$ times, respectively, is given by

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n; \mathbf{p}) = \frac{(\sum_i x_i)!}{x_1!...x_n!} p_1^{x_1} \times p_2^{x_2} \times \cdots \times p_n^{x_n}$$
$$= \frac{(\sum_i x_i)!}{x_1!...x_n!} \prod_i p_i^{x_i}$$

# Categorical Distributions

- The **categorical distribution** models the probability of **events** resulting from a single trial with **more than two possible outcomes**, e.g.,

  - we roll a fair-sided dice once
  - what is the probability of observing a 'five'?

- The probability of events $X_1, X_2, ..., X_n$ with probabilities $\mathbf{p} = p_1, p_2, ..., p_n$ occurring exactly $x_1, x_2, ..., x_n$ times, respectively, is given by

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n; \mathbf{p}) &= p_1^{x_1} \times p_2^{x_2} \times \cdots \times p_n^{x_n} \\ &= \prod_i p_i^{x_i} \end{aligned}$$

# Marginalization

**Intuition**

We want to know the probability of an event $A$ *irrespective of* the outcome of another event $B$. We can obtain it, by summing over all possible outcomes $\mathcal{B}$ of B.

- Take an event $B$. The set of *all* possible *individual* outcomes of $B$, $\mathcal{B}$ is the **partition** of the outcome space
- E.g., $\mathcal{B} =$\{head, tail\} for a coin flip; $\mathcal{B} =$\{king, heart, diamond, spades\} for card suits
- We can **marginalize** over the set of outcomes of $B$ as follows

$$P(A) = \sum_{b \in \mathcal{B}} P(A, B = b)$$

or equivalently (remember the multiplication rule?)

$$P(A) = \sum_{b \in \mathcal{B}} P(A|B = b)P(B = b)$$

and even for conditional probabilities

$$P(A|C) = \sum_{b \in \mathcal{B}} P(A|C, B = b)P(B = b|C)$$

# Marginalization

**Example**

We want to know the probability of success of movies of a specific genre ($\mathcal{A} = \{comedy, thriller, romance\}$). But we only have data on movie success probabilities in a specific market, namely ($\mathcal{B} = \{EU, NA, AUS\}$).

$$P(A) = \sum_{b \in \mathcal{B}} P(A, B = b)$$

| $A$ | $B$ | $P(A, B)$ |
|---|---|---|
| romance | EU | 0.05 |
| romance | NA | 0.1 |
| romance | AUS | 0.3 |
| thriller | EU | 0.1 |
| thriller | NA | 0.2 |
| thriller | AUS | 0.1 |
| comedy | EU | 0.1 |
| comedy | NA | 0.025 |
| comedy | AUS | 0.025 |
| | | 1.0 |

**Example**

We want to know the probability of success of movies of a specific genre ($\mathcal{A} = \{comedy, thriller, romance\}$). But we only have data on movie success probabilities in a specific market, namely ($\mathcal{B} = \{EU, NA, AUS\}$).

$$P(A) = \sum_{b \in \mathcal{B}} P(A, B = b)$$

| $A$ | $B$ | $P(A, B)$ |
|---|---|---|
| romance | EU | 0.05 |
| romance | NA | 0.1 |
| romance | AUS | 0.3 |
| thriller | EU | 0.1 |
| thriller | NA | 0.2 |
| thriller | AUS | 0.1 |
| comedy | EU | 0.1 |
| comedy | NA | 0.025 |
| comedy | AUS | 0.025 |
| | | 1.0 |

$\sum$

| $A$ | $P(A)$ |
|---|---|
| romance | |
| thriller | |
| comedy | |

# Marginalization

**Example**

We want to know the probability of success of movies of a specific genre ($\mathcal{A} = \{comedy, thriller, romance\}$). But we only have data on movie success probabilities in a specific market, namely ($\mathcal{B} = \{EU, NA, AUS\}$).

$$P(A) = \sum_{b \in \mathcal{B}} P(A, B = b)$$

| $A$ | $B$ | $P(A, B)$ |
|---|---|---|
| romance | EU | 0.05 |
| romance | NA | 0.1 |
| romance | AUS | 0.3 |
| thriller | EU | 0.1 |
| thriller | NA | 0.2 |
| thriller | AUS | 0.1 |
| comedy | EU | 0.1 |
| comedy | NA | 0.025 |
| comedy | AUS | 0.025 |
| | | 1.0 |

$\sum$

| $A$ | $P(A)$ |
|---|---|
| romance | 0.45 |
| thriller | |
| comedy | |

# Marginalization

**Example**

We want to know the probability of success of movies of a specific genre ($\mathcal{A} = \{comedy, thriller, romance\}$). But we only have data on movie success probabilities in a specific market, namely ($\mathcal{B} = \{EU, NA, AUS\}$).

$$P(A) = \sum_{b \in \mathcal{B}} P(A, B = b)$$

| $A$ | $B$ | $P(A, B)$ |
|---|---|---|
| romance | EU | 0.05 |
| romance | NA | 0.1 |
| romance | AUS | 0.3 |
| thriller | EU | 0.1 |
| thriller | NA | 0.2 |
| thriller | AUS | 0.1 |
| comedy | EU | 0.1 |
| comedy | NA | 0.025 |
| comedy | AUS | 0.025 |
| | | 1.0 |

$\sum$

$\sum$

| $A$ | $P(A)$ |
|---|---|
| romance | 0.45 |
| thriller | |
| comedy | |

**Example**

We want to know the probability of success of movies of a specific genre ($\mathcal{A} = \{comedy, thriller, romance\}$). But we only have data on movie success probabilities in a specific market, namely ($\mathcal{B} = \{EU, NA, AUS\}$).

$$P(A) = \sum_{b \in \mathcal{B}} P(A, B = b)$$

| $A$ | $B$ | $P(A, B)$ |
|---|---|---|
| romance | EU | 0.05 |
| romance | NA | 0.1 |
| romance | AUS | 0.3 |
| thriller | EU | 0.1 |
| thriller | NA | 0.2 |
| thriller | AUS | 0.1 |
| comedy | EU | 0.1 |
| comedy | NA | 0.025 |
| comedy | AUS | 0.025 |
| | | 1.0 |

$\sum$

$\sum$

| $A$ | $P(A)$ |
|---|---|
| romance | 0.45 |
| thriller | 0.4 |
| comedy | |

# Marginalization

**Example**

We want to know the probability of success of movies of a specific genre ($\mathcal{A} = \{comedy, thriller, romance\}$). But we only have data on movie success probabilities in a specific market, namely ($\mathcal{B} = \{EU, NA, AUS\}$).

$$P(A) = \sum_{b \in \mathcal{B}} P(A, B = b)$$

| $A$ | $B$ | $P(A, B)$ |
|---|---|---|
| romance | EU | 0.05 |
| romance | NA | 0.1 |
| romance | AUS | 0.3 |
| thriller | EU | 0.1 |
| thriller | NA | 0.2 |
| thriller | AUS | 0.1 |
| comedy | EU | 0.1 |
| comedy | NA | 0.025 |
| comedy | AUS | 0.025 |
| | | 1.0 |

| $A$ | $P(A)$ |
|---|---|
| romance | 0.45 |
| thriller | 0.4 |
| comedy | |

**Example**

We want to know the probability of success of movies of a specific genre ($\mathcal{A} = \{comedy, thriller, romance\}$). But we only have data on movie success probabilities in a specific market, namely ($\mathcal{B} = \{EU, NA, AUS\}$).

$$P(A) = \sum_{b \in \mathcal{B}} P(A, B = b)$$

| $A$ | $B$ | $P(A,B)$ |
|---|---|---|
| romance | EU | 0.05 |
| romance | NA | 0.1 |
| romance | AUS | 0.3 |
| thriller | EU | 0.1 |
| thriller | NA | 0.2 |
| thriller | AUS | 0.1 |
| comedy | EU | 0.1 |
| comedy | NA | 0.025 |
| comedy | AUS | 0.025 |
| | | 1.0 |

$\sum$ $\sum$ $\sum$

| $A$ | $P(A)$ |
|---|---|
| romance | 0.45 |
| thriller | 0.4 |
| comedy | 0.15 |
| | 1.0 |

# Quiz!

Please go to

`https://pollev.com/iml2021`

for a quick quiz on probabilities!

# Probability and Machine Learning

We probably all agree that probabilities are useful for thinking about card games or coin flips

... but why should we care in machine learning?

Consider typical classification problems

- document $\rightarrow$ $\{$spam, no spam$\}$
- hand-written digit $\rightarrow \{0,1,2,3,4,5,6,7,8,9\}$
- purchase history $\rightarrow$ recommend $\{$book a, book b, book c, ...$\}$

# Probability and Machine Learning

We probably all agree that probabilities are useful for thinking about card games or coin flips

... but why should we care in machine learning?

Consider typical classification problems

- document $\rightarrow$ {spam, no spam}
- hand-written digit $\rightarrow$ {0,1,2,3,4,5,6,7,8,9}
- purchase history $\rightarrow$ recommend {book a, book b, book c, ...}

- **uncertainty**, due to few observations, noisy data, ...
- model features as following certain **probability distributions**
- **soft predictions** ("we are 60% confident that Bob will like *Harry Potter* given his purchase history")
- ...

> "All models are wrong, but some are useful."
> (George Box, Statistician)

## Probabilistic Models

- allow to reason about random events in a **principled** way.
- allow to formalize hypotheses as different types of probability distributions, and use the laws of probability to derive predictions

## Example: Spam classification

- An email is a random event with two possible outcomes: `spam`, `not spam`
- The probability of observing a spam email $P(spam) = \theta$, and trivially $P(not\ spam) = 1 - \theta$.
- We might care about a random variable $X$ as the number of spam emails in an inbox of 100 emails. $X$ is distributed according to the **binomial distribution**, and depends on the **parameters** $\theta$ and $N = 100$

$$X \sim Binomial(\theta, N = 100)$$

## Learning Probabilistic Models I

$X$ is distributed according to the **binomial distribution**, and depends on the **parameters** $\theta$ and $N = 100$

$$X \sim Binomial(\theta, N = 100)$$

- In order to make predictions of $X$ we need to know the parameters $\theta$. **How do we learn them?**
- Typically, $\theta$ is unknown, but if we have **data** available we can **estimate** $\theta$
- One common choice is to pick $\theta$ that maximizes the probability of the observed data

  $$\hat{\theta} = \underset{\theta}{\text{argmax}}\, P(X; \theta, N)$$

  That is the **maximum likelihood estimate (MLE)** of $\theta$.
- Once we have estimated $\theta$ we can use it to **predict** values for unseen data

## Learning Probabilistic Models II

The **maximum likelihood principle**

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(X; \theta, N) \tag{1}$$

- Consider a **data set** consisting of 100 emails, 20 of which are spam.
- Following from the binomial distribution

$$\mathcal{L}(\theta) = P(X; \theta, N) = \binom{n}{m} \theta^{x} (1-\theta)^{N-x}$$

the **likelihood of the data**[1] is $\propto \; \theta^{20}(1-\theta)^{100-20}$

- What do you think would be a good value for $\theta = p(spam = 1)$? Why?

- Next lecture, we will see how to derive this value in a principled way

[1] $\propto$ means 'proportional to'. $\binom{n}{m}$ can be ignored because it is independent of $\theta$.

## Maximum likelihood is only one choice of estimator among many

- Consider a data set of one inbox with no spam email. MLE: $\theta = 1$, and hence $P(\textit{not spam}) = \theta = 1$ and $P(\textit{spam}) = 1 - \theta = 0$.
  $\rightarrow$ "spam emails don't exist"
- We could modify this estimate with our **prior belief**. E.g., we might believe that about 80 of 100 emails are not spam. We 'nudge' $\theta$ from $\theta = 1$ towards $\theta = 0.80$
- We can combine our prior belief with the estimate from the data to arrive at a **posterior probability distribution** over $\theta$: $P(\theta)$.

$$P(\theta|x) = \frac{P(\theta)P(x|\theta)}{P(x)} \propto P(\theta)P(x|\theta) \qquad \text{(looks familiar)?}$$

- The **maximum a posteriori estimate** is then

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(\theta)P(x|\theta)$$

# Summary

**Probability underlies many modern knowledge technologies**

- estimate the (conditional, joint) probability of observations
- Bayes rule
- Expectations and marginalization
- Probabilistic models
- Maximum likelihood estimation (taster)
- Maximum aposteriori estimation (taster)

**Next Lecture(s):**

- Optimization
- Naive Bayes Classification

# References

Chris Bishop. Pattern Recognition and Machine Learning. Chapters: 1.2 (intro), 1.2.3, 2 (intro), 2.1 (up to 2.1.1), 2.2 (up to 2.2.1)

THE UNIVERSITY OF MELBOURNE

# Optional / If time permits: Expectations

The **expectation** of a function (like a probability distribution) is the **weighted average** of all possible outcomes, weighted by their respective probability.

- For functions with discrete outputs

$$E[f(x)] = \sum_{x \in \mathcal{X}} f(x)P(x)$$

- For functions with continuous outputs

$$E[f(x)] = \int_{\mathcal{X}} f(x)P(x)dx$$

# Optional / If time permits: Expectations

The **expectation** of a function (like a probability distribution) is the **weighted average** of all possible outcomes, weighted by their respective probability.

- On sunny days Bob watches 1 movie
- On rainy days Bob watches 3 movies
- Bob lives in Melbourne, it rains on 70% of all days
- What is the expected number of movies Bob watches per day?

# Optional / If time permits: Expectations

The **expectation** of a function (like a probability distribution) is the **weighted average** of all possible outcomes, weighted by their respective probability.

- On sunny days Bob watches 1 movie
- On rainy days Bob watches 3 movies
- Bob lives in Melbourne, it rains on 70% of all days
- What is the expected number of movies Bob watches per day?

$$1 * 0.3 + 3 * 0.7 = 2.4$$