Machine Learning

Lecture 01-1: Basics of Probability Theory

Nevin L. Zhang
lzhang@cse.ust.hk

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology

Outline

Random Experiments

- Probability associated with a **random experiment** — a process with uncertain outcomes
- Often kept implicit

Tail Head

Random Experiments

- Probability associated with a **random experiment** — a process with uncertain outcomes
- Often kept implicit

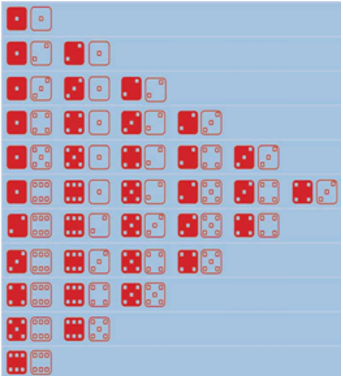
Tail Head

In machine learning, we often assume that data are generated by a hypothetical process (or a model), and task is to determine the structure and parameters of the model from data.

Sample Space

- **Sample space (aka population)** Ω: Set of possible outcomes and a random experiment.
- Example: Rolling two dices.

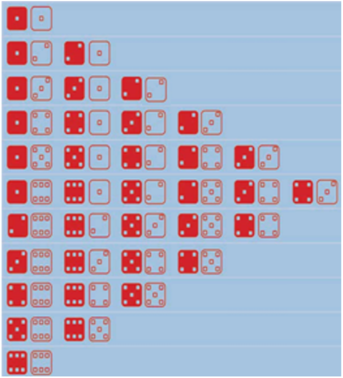
Ω



Sample Space

- **Sample space (aka population)** Ω: Set of possible outcomes and a random experiment.
- Example: Rolling two dices.

Ω

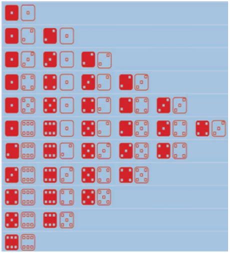


- Elements in a sample space are outcomes.

Events

- **Event**: A subset of the sample space.

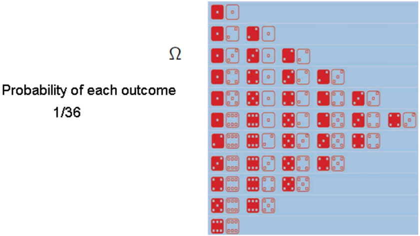
Ω



- Example: The two results add to 4.

Probability Weight Function

- A **probability weight** $P(\omega)$ is assigned to each outcome.



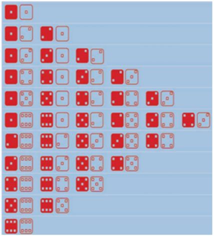
Probability Weight Function

- A **probability weight** $P(\omega)$ is assigned to each outcome.

Ω

Probability of each outcome

1/36



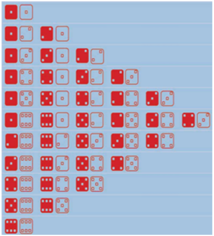
In Machine Learning, we often need to determine the probability weights, or related parameters, from data.

Probability Weight Function

- A **probability weight** $P(\omega)$ is assigned to each outcome.

Ω

Probability of each outcome
1/36



In Machine Learning, we often need to determine the probability weights, or related parameters, from data. This task is called **parameter learning.**

Probability measure

- Probability $P(E)$ of an event E: $P(E) = \sum_{\omega \in E} P(\omega)$

Probability measure

- Probability $P(E)$ of an event E: $P(E) = \sum_{\omega \in E} P(\omega)$
- A **probability measure** is a mapping from the set of **events** to [0, 1]

$$P : 2^{\Omega} \rightarrow [0, 1]$$

Probability measure

- Probability $P(E)$ of an event E: $P(E) = \sum_{\omega \in E} P(\omega)$
- A **probability measure** is a mapping from the set of **events** to [0, 1]

$$P : 2^{\Omega} \rightarrow [0, 1]$$

that satisfies Kolmogorov's axioms:

Probability measure

- Probability $P(E)$ of an event E: $P(E) = \sum_{\omega \in E} P(\omega)$
- A **probability measure** is a mapping from the set of **events** to [0, 1]

$$P : 2^{\Omega} \rightarrow [0, 1]$$

  that satisfies Kolmogorov's axioms:

  1. $P(\Omega) = 1$.

Probability measure

- Probability $P(E)$ of an event E: $P(E) = \sum_{\omega \in E} P(\omega)$
- A **probability measure** is a mapping from the set of **events** to [0, 1]

$$P : 2^{\Omega} \rightarrow [0, 1]$$

that satisfies Kolmogorov's axioms:

1. $P(\Omega) = 1.$
2. $P(A) \geq 0 \; \forall A \subseteq \Omega$

Probability measure

- Probability $P(E)$ of an event E: $P(E) = \sum_{\omega \in E} P(\omega)$

- A **probability measure** is a mapping from the set of **events** to [0, 1]

$$P : 2^{\Omega} \rightarrow [0, 1]$$

that satisfies Kolmogorov's axioms:

1. $P(\Omega) = 1$.
2. $P(A) \geq 0 \ \forall A \subseteq \Omega$
3. **Additivity**: $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$.

Probability measure

- Probability $P(E)$ of an event E: $P(E) = \sum_{\omega \in E} P(\omega)$
- A **probability measure** is a mapping from the set of **events** to [0, 1]

$$P : 2^{\Omega} \rightarrow [0, 1]$$

that satisfies Kolmogorov's axioms:

1. $P(\Omega) = 1$.
2. $P(A) \geq 0 \ \forall A \subseteq \Omega$
3. **Additivity**: $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$.

Probability measure

- Probability $P(E)$ of an event E: $P(E) = \sum_{\omega \in E} P(\omega)$

- A **probability measure** is a mapping from the set of **events** to [0, 1]

$$P : 2^{\Omega} \rightarrow [0, 1]$$

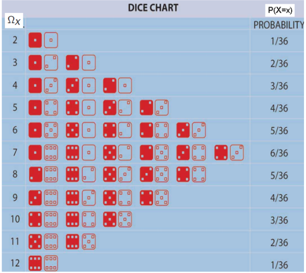
  that satisfies Kolmogorov's axioms:

  1. $P(\Omega) = 1$.
  2. $P(A) \geq 0 \ \forall A \subseteq \Omega$
  3. **Additivity**: $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$.

In a more advanced treatment of Probability Theory, we would start with the concept of probability measure, instead of probability weights.

Random Variables

- A **random variable** is a function over the sample space.
  - Example: X = sum of the two results. $X((2,5)) = 7; X((3,1)) = 4)$



- Why is it random?

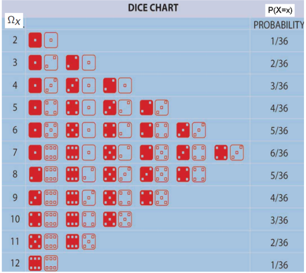
Random Variables

- A **random variable** is a function over the sample space.
  - Example: X = sum of the two results. $X((2,5)) = 7; X((3,1)) = 4)$

- Why is it random? The experiment.

Random Variables

- A **random variable** is a function over the sample space.
  - Example: X = sum of the two results. $X((2, 5)) = 7; X((3, 1)) = 4)$



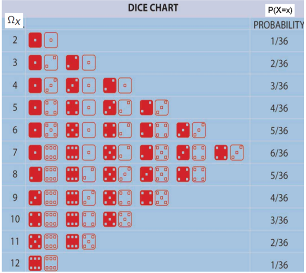
- Why is it random? The experiment.
- **Domain** of a random variable: Set of all its possible values.

$$\Omega_X = \{2, 3, \ldots, 12\}$$

Random Variables and Event

- A random variable X taking a specific value x is an event:

$$\Omega_{X=x} = \{\omega \in \Omega | X(\omega) = x\}$$



- $\Omega_{X=4} = \{(1,3),(2,2,)(3,1)\}$.

Probability Mass Function (Distribution)

- **Probability mass function** $P(X)$: $\Omega_X \rightarrow [0, 1]$

$$P(X = x) = P(\Omega_{X=x})$$

Probability Mass Function (Distribution)

- **Probability mass function** $P(X)$: $\Omega_X \rightarrow [0, 1]$

$$P(X = x) = P(\Omega_{X=x})$$

- $P(X = 4) = P(\{(1, 3), (2, 2,)(3, 1)\}) = \frac{3}{36}$.

Probability Mass Function (Distribution)

- **Probability mass function** $P(X)$: $\Omega_X \rightarrow [0, 1]$

$$P(X = x) = P(\Omega_{X=x})$$

- $P(X = 4) = P(\{(1, 3), (2, 2,)(3, 1)\}) = \frac{3}{36}$.
- If X is continuous, we have a **density function** $p(X)$.

Outline

Frequentist interpretation

- Probabilities are **long term relative frequencies**.
- Example:
  - X is result of coin tossing. $\Omega_X = \{\mathrm{H}, \mathrm{T}\}$

Frequentist interpretation

- Probabilities are **long term relative frequencies**.

- Example:
  - X is result of coin tossing. $\Omega_X = \{\mathrm{H}, \mathrm{T}\}$
  - $P(X{=}\mathrm{H}) = 1/2$ means that

Frequentist interpretation

- Probabilities are **long term relative frequencies**.

- Example:
  - X is result of coin tossing. $\Omega_X = \{\mathrm{H}, \mathrm{T}\}$
  - $P(X{=}\mathrm{H}) = 1/2$ means that
    - *the relative frequency of getting heads* will almost surely approach 1/2 as the number of tosses goes to infinite.

Frequentist interpretation

- Probabilities are **long term relative frequencies**.
- Example:
  - X is result of coin tossing. $\Omega_X = \{\mathrm{H}, \mathrm{T}\}$
  - $P(X{=}\mathrm{H}) = 1/2$ means that
    - *the relative frequency of getting heads* will almost surely approach 1/2 as the number of tosses goes to infinite.
  - Justified by the Law of Large Numbers:

Frequentist interpretation

- Probabilities are **long term relative frequencies**.

- Example:
  - X is result of coin tossing. $\Omega_X = \{\mathrm{H}, \mathrm{T}\}$
  - $P(X{=}\mathrm{H}) = 1/2$ means that
    - *the relative frequency of getting heads* will almost surely approach 1/2 as the number of tosses goes to infinite.
  - Justified by the Law of Large Numbers:
    - X_i: result of the i-th tossing; 1 – H, 0 — T

Frequentist interpretation

- Probabilities are **long term relative frequencies**.

- Example:
  - X is result of coin tossing. $\Omega_X = \{\text{H}, \text{T}\}$
  - $P(X{=}\text{H}) = 1/2$ means that
    - *the relative frequency of getting heads* will almost surely approach 1/2 as the number of tosses goes to infinite.
  - Justified by the Law of Large Numbers:
    - X_i: result of the i-th tossing; 1 – H, 0 — T
    - Law of Large Numbers:

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1}{2} \qquad \text{with probability } 1$$

Frequentist interpretation

- Probabilities are **long term relative frequencies**.
- Example:
  - X is result of coin tossing. $\Omega_X = \{\mathrm{H}, \mathrm{T}\}$
  - $P(X{=}\mathrm{H}) = 1/2$ means that
    - *the relative frequency of getting heads* will almost surely approach 1/2 as the number of tosses goes to infinite.
  - Justified by the Law of Large Numbers:
    - X_i: result of the i-th tossing; 1 – H, 0 — T
    - Law of Large Numbers:

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1}{2} \qquad \text{with probability 1}$$

- The frequentist interpretation is meaningful only when experiment can be repeated under the same condition.

Bayesian interpretation

- Probabilities are **logically consistent degrees of beliefs**.

Bayesian interpretation

- Probabilities are **logically consistent degrees of beliefs**.

- Applicable when experiment not repeatable.

Bayesian interpretation

- Probabilities are **logically consistent degrees of beliefs**.

- Applicable when experiment not repeatable.
- Depends on a person's state of knowledge.

Bayesian interpretation

- Probabilities are **logically consistent degrees of beliefs**.

- Applicable when experiment not repeatable.
- Depends on a person's state of knowledge.
- Example: "probability that Suez canal is longer than the Panama canal".

Bayesian interpretation

- Probabilities are **logically consistent degrees of beliefs**.

- Applicable when experiment not repeatable.
- Depends on a person's state of knowledge.
- Example: "probability that Suez canal is longer than the Panama canal".
  - Doesn't make sense under frequentist interpretation.

Bayesian interpretation

- Probabilities are **logically consistent degrees of beliefs**.

- Applicable when experiment not repeatable.
- Depends on a person's state of knowledge.
- Example: "probability that Suez canal is longer than the Panama canal".
  - Doesn't make sense under frequentist interpretation.
  - Subjectivist: degree of belief based on state of knowledge

Bayesian interpretation

- Probabilities are **logically consistent degrees of beliefs**.
- Applicable when experiment not repeatable.
- Depends on a person's state of knowledge.
- Example: "probability that Suez canal is longer than the Panama canal".
  - Doesn't make sense under frequentist interpretation.
  - Subjectivist: degree of belief based on state of knowledge
    - Primary school student: 0.5
    - Me: 0.8
    - Geographer: 1 or 0
- Arguments such as **Dutch book** are used to explain why one's probability beliefs must satisfy Kolmogorov's axioms.

Interpretations of Probability

- Now both interpretations are accepted. In practice, subjective beliefs and statistical data complement each other.

Interpretations of Probability

- Now both interpretations are accepted. In practice, subjective beliefs and statistical data complement each other.
  - We rely on subjective beliefs **(prior probabilities)** when data are scarce.

Interpretations of Probability

- Now both interpretations are accepted. In practice, subjective beliefs and statistical data complement each other.
  - We rely on subjective beliefs **(prior probabilities)** when data are scarce.
  - As more and more data become available, we rely less and less on subjective beliefs.

Interpretations of Probability

- Now both interpretations are accepted. In practice, subjective beliefs and statistical data complement each other.
  - We rely on subjective beliefs **(prior probabilities)** when data are scarce.
  - As more and more data become available, we rely less and less on subjective beliefs.
  - Often, we also use **prior probabilities** to impose some **bias** on the kind of results we want from a machine learning algorithm.

Interpretations of Probability

- Now both interpretations are accepted. In practice, subjective beliefs and statistical data complement each other.
  - We rely on subjective beliefs **(prior probabilities)** when data are scarce.
  - As more and more data become available, we rely less and less on subjective beliefs.
  - Often, we also use **prior probabilities** to impose some **bias** on the kind of results we want from a machine learning algorithm.

Interpretations of Probability

- Now both interpretations are accepted. In practice, subjective beliefs and statistical data complement each other.
  - We rely on subjective beliefs **(prior probabilities)** when data are scarce.
  - As more and more data become available, we rely less and less on subjective beliefs.
  - Often, we also use **prior probabilities** to impose some **bias** on the kind of results we want from a machine learning algorithm.

- The subjectivist interpretation makes concepts such as conditional independence easy to understand.

Outline

Binomial and Bernoulli Distributions

- Suppose we toss a coin n times. At each time, the probability of getting a head is θ.

Binomial and Bernoulli Distributions

- Suppose we toss a coin n times. At each time, the probability of getting a head is θ.
- Let X be the number of heads. Then X follows the **binomial distribution**, written as $X \sim Bin(n, \theta)$:

$$Bin(X = k|n, \theta) = \begin{cases} \binom{n}{k}\theta^k(1-\theta)^{n-k} & \text{if } 0 \leq k \leq n \\ 0 & \text{if } k < 0 \text{ or } k > n \end{cases}$$

Binomial and Bernoulli Distributions

- Suppose we toss a coin n times. At each time, the probability of getting a head is θ.
- Let X be the number of heads. Then X follows the **binomial distribution**, written as $X \sim Bin(n, \theta)$:

$$Bin(X = k|n, \theta) = \begin{cases} \binom{n}{k}\theta^k(1-\theta)^{n-k} & \text{if } 0 \leq k \leq n \\ 0 & \text{if } k < 0 \text{ or } k > n \end{cases}$$

- If $n = 1$, then X follows the **Bernoulli distribution**, written as $X \sim Ber(\theta)$

$$Ber(X = x|\theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \end{cases}$$

Multinomial Distribution

- Suppose we toss a K-sided die n times. At each time, the probability of getting result j is θ_j. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^\top$.

Multinomial Distribution

- Suppose we toss a K-sided die n times. At each time, the probability of getting result j is θ_j. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^\top$.
- Let $\mathbf{x} = (x_1, ..., x_K)$ be a random vector, where x_j is the number of times side j of the die occurs. Then $\mathbf{x}$ follows the **multinomial distribution**, written as $\mathbf{x} \sim Multi(n, \boldsymbol{\theta})$

$$Multi(\mathbf{x}|n, \boldsymbol{\theta}) = \binom{n}{x_1, \ldots, x_K} \prod_{j=1}^{K} \theta_k^{x_j},$$

where $\binom{n}{x_1, \ldots, x_K} = \frac{n!}{x_1! \ldots x_K!}$ is the multinomial coefficient

Categorical Distribution

- In the previous slide, if $n = 1$, $\mathbf{x} = (x_1, ..., x_K)$ has one component being 1 and the others are 0. In other words, it is a **one-hot** vector.
- In this case, $\mathbf{x}$ follows the **categorical distribution**, written as $\mathbf{x} \sim Cat(\boldsymbol{\theta})$

$$Cat(\mathbf{x}|\boldsymbol{\theta}) = \prod_{j=1}^{K} \theta_j^{\mathbf{1}(x_j=1)},$$

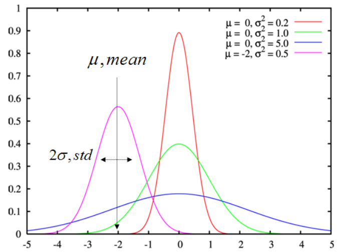
where $\mathbf{1}(x_j = 1)$ is the indicator function, whose value is 1 when $x_j = 1$ and 0 otherwise.

Gaussian (Normal) Distribution

- The most widely used distribution in statistics and machine learning is the Gaussian or normal distribution.
- Its probability density is given by

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Here $\mu = E[X]$ is the mean (and mode), and $\sigma^2 = var[X]$ is the variance



Outline

Joint probability mass function

- **Probability mass function** of a random variable X:

Joint probability mass function

- **Probability mass function** of a random variable X:

$$P(X) : \Omega_X \rightarrow [0, 1]$$

Joint probability mass function

- **Probability mass function** of a random variable X:

$$P(X) : \Omega_X \rightarrow [0, 1]$$

$$P(X = x) = P(\Omega_{X=x}).$$

Joint probability mass function

- **Probability mass function** of a random variable X:

$$P(X) : \Omega_X \rightarrow [0, 1]$$

$$P(X = x) = P(\Omega_{X=x}).$$

- Suppose there are n random variables $X_1, X_2, \ldots, X_n$.

Joint probability mass function

- **Probability mass function** of a random variable X:

$$P(X) : \Omega_X \to [0, 1]$$

$$P(X = x) = P(\Omega_{X=x}).$$

- Suppose there are n random variables X_1, X_2, ..., X_n.
- A **joint probability mass function**, $P(X_1, X_2, \ldots, X_n)$, over those random variables is:

Joint probability mass function

- **Probability mass function** of a random variable X:

$$P(X) : \Omega_X \rightarrow [0,1]$$

$$P(X = x) = P(\Omega_{X=x}).$$

- Suppose there are n random variables X_1, X_2, ..., X_n.
- A **joint probability mass function**, $P(X_1, X_2, \ldots, X_n)$, over those random variables is:
  - a function defined on the Cartesian product of their state spaces:

$$\prod_{i=1}^{n} \Omega_{X_i} \rightarrow [0,1]$$

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(\Omega_{X_1=x_1} \cap \Omega_{X_2=x_2} \cap \ldots \cap \Omega_{X_n=x_n}).$$

Joint probability mass function

- Example:

Joint probability mass function

- Example:
  - Population: Apartments in Hong Kong rental market.

Joint probability mass function

- Example:
  - Population: Apartments in Hong Kong rental market.
  - Random variables: (of a random selected apartment)

Joint probability mass function

- Example:
  - Population: Apartments in Hong Kong rental market.
  - Random variables: (of a random selected apartment)
    - Monthly Rent: {low ($\leq 1k$), medium ((1k, 2k]), upper medium((2k, 4k]), high ($\geq$4k)},

Joint probability mass function

- Example:
  - Population: Apartments in Hong Kong rental market.
  - Random variables: (of a random selected apartment)
    - Monthly Rent: {low ($\leq 1k$), medium ((1k, 2k]), upper medium((2k, 4k]), high (≥4k)},
    - Type: {public, private, others}

Joint probability mass function

- Example:
  - Population: Apartments in Hong Kong rental market.
  - Random variables: (of a random selected apartment)
    - Monthly Rent: {low ($\leq 1k$), medium ((1k, 2k]), upper medium((2k, 4k]), high ($\geq$4k)},
    - Type: {public, private, others}
  - Joint probability distribution $P(\text{Rent}, \text{Type})$:

Joint probability mass function

- Example:
  - Population: Apartments in Hong Kong rental market.
  - Random variables: (of a random selected apartment)
    - Monthly Rent: {low ($\leq 1k$), medium ((1k, 2k]), upper medium((2k, 4k]), high ($\geq$4k)},
    - Type: {public, private, others}
  - Joint probability distribution $P(\text{Rent}, \text{Type})$:

	public	private	others
low	.17	.01	.02
medium	.44	.03	.01
upper medium	.09	.07	.01
high	0	0.14	0.1

Multivariate Gaussian Distributions

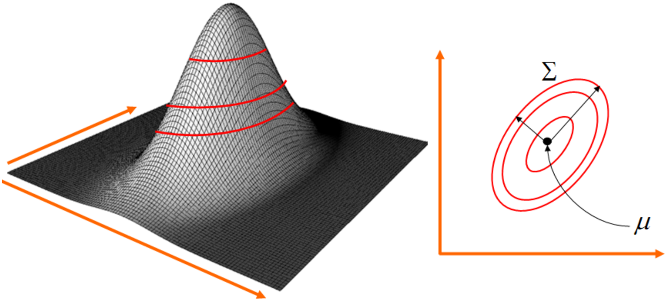
- For continuous variables, the most commonly used joint distribution is the multivariate Gaussian distribution: $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} exp\left[-\frac{(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}{2}\right]$$

  - D: dimensionality.
  - $\mathbf{x}$: vector of D random variables, representing data
  - $\boldsymbol{\mu}$: vector of means
  - Σ: covariance matrix. $|\Sigma|$ denotes the determinant of Σ.

Multivariate Gaussian Distributions

- A 2-D Gaussian distribution.
- μ: center of contours
- Σ: orientation and size of contours



Marginal probability

- What is the probability of a randomly selected apartment being a public one?

Marginal probability

- What is the probability of a randomly selected apartment being a public one? (Law of total probability)

  P(Type=pulic) = P(Type=public, Rent=low)+

Marginal probability

- What is the probability of a randomly selected apartment being a public one? (Law of total probability)

  P(Type=pulic) = P(Type=public, Rent=low)+P(Type=public, Rent=medium)+

Marginal probability

- What is the probability of a randomly selected apartment being a public one? (Law of total probability)

  P(Type=pulic) = P(Type=public, Rent=low)+P(Type=public, Rent=medium)+ P(Type=public, Rent=upper medium)+

Marginal probability

- What is the probability of a randomly selected apartment being a public one? (Law of total probability)

P(Type=pulic) = P(Type=public, Rent=low)+P(Type=public, Rent=medium)+ P(Type=public, Rent=upper medium)+ P(Type=public, Rent=high)

Marginal probability

- What is the probability of a randomly selected apartment being a public one? (Law of total probability)

P(Type=pulic) = P(Type=public, Rent=low)+P(Type=public, Rent=medium)+ P(Type=public, Rent=upper medium)+ P(Type=public, Rent=high) = .7

Marginal probability

- What is the probability of a randomly selected apartment being a public one? (Law of total probability)

P(Type=pulic) = P(Type=public, Rent=low)+P(Type=public, Rent=medium)+ P(Type=public, Rent=upper medium)+ P(Type=public, Rent=high) = .7

P(Type=private) = P(Type=private, Rent=low)+ P(Type=private, Rent=medium)+ P(Type=private, Rent=upper medium)+ P(Type=private, Rent=high)

Marginal probability

- What is the probability of a randomly selected apartment being a public one? (Law of total probability)

P(Type=pulic) = P(Type=public, Rent=low)+P(Type=public, Rent=medium)+ P(Type=public, Rent=upper medium)+ P(Type=public, Rent=high) = .7

P(Type=private) = P(Type=private, Rent=low)+ P(Type=private, Rent=medium)+ P(Type=private, Rent=upper medium)+ P(Type=private, Rent=high)= .25

Marginal probability

- What is the probability of a randomly selected apartment being a public one? (Law of total probability)

P(Type=pulic) = P(Type=public, Rent=low)+P(Type=public, Rent=medium)+ P(Type=public, Rent=upper medium)+ P(Type=public, Rent=high) = .7

P(Type=private) = P(Type=private, Rent=low)+ P(Type=private, Rent=medium)+ P(Type=private, Rent=upper medium)+ P(Type=private, Rent=high)= .25

	public	private	others	P(Rent)
low	.17	.01	.02	.2
medium	.44	.03	.01	.48
upper medium	.09	.07	.01	.17
high	0	0.14	0.1	.15
P(Type)	.7	.25	.05	

Marginal probability

- What is the probability of a randomly selected apartment being a public one? (Law of total probability)

$P(\text{Type=pulic}) = P(\text{Type=public, Rent=low}) + P(\text{Type=public, Rent=medium}) + P(\text{Type=public, Rent=upper medium}) + P(\text{Type=public, Rent=high}) = .7$

$P(\text{Type=private}) = P(\text{Type=private, Rent=low}) + P(\text{Type=private, Rent=medium}) + P(\text{Type=private, Rent=upper medium}) + P(\text{Type=private, Rent=high}) = .25$

	public	private	others	P(Rent)
low	.17	.01	.02	.2
medium	.44	.03	.01	.48
upper medium	.09	.07	.01	.17
high	0	0.14	0.1	.15
P(Type)	.7	.25	.05	

- Called marginal probability because written on the margins.

Conditional probability

- For events A and B:

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

Conditional probability

- For events A and B:

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

- Meaning:

Conditional probability

- For events A and B:

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

- Meaning:
  - $P(A)$: My probability on A (without any knowledge about B)

Conditional probability

- For events A and B:

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

- Meaning:
  - $P(A)$: My probability on A (without any knowledge about B)
  - $P(A|B)$: My probability on event A assuming that I know event B is true.

Conditional probability

- For events A and B:

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

- Meaning:
  - $P(A)$: My probability on A (without any knowledge about B)
  - $P(A|B)$: My probability on event A assuming that I know event B is true.
- What is the probability of a randomly selected private apartment having "low" rent?

Conditional probability

- For events A and B:

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

- Meaning:
  - $P(A)$: My probability on A (without any knowledge about B)
  - $P(A|B)$: My probability on event A assuming that I know event B is true.
- What is the probability of a randomly selected private apartment having "low" rent?

$$P(Rent{=}low|Type{=}private)$$
$$= \frac{P(Rent{=}Low,\ Type{=}private)}{P(Type{=}private)} = .01/.25 {=} .04$$

Conditional probability

- For events A and B:

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

- Meaning:
  - $P(A)$: My probability on A (without any knowledge about B)
  - $P(A|B)$: My probability on event A assuming that I know event B is true.
- What is the probability of a randomly selected private apartment having "low" rent?

$$P(Rent{=}low|Type{=}private) = \frac{P(Rent{=}Low,\ Type{=}private)}{P(Type{=}private)} = .01/.25 = .04$$

In contrast:

$$P(Rent{=}low) = 0.2.$$

Marginal independence

- Two random variables X and Y are **marginally independent**,

Marginal independence

- Two random variables X and Y are **marginally independent**, written $X \perp Y$, if

Marginal independence

- Two random variables X and Y are **marginally independent**, written $X \perp Y$, if
  - for any state x of X and any state y of Y,

$$P(X{=}x|Y{=}y) = P(X{=}x), \text{ whenever } P(Y = y) \neq 0.$$

- Meaning:

Marginal independence

- Two random variables X and Y are **marginally independent**, written $X \perp Y$, if
  - for any state x of X and any state y of Y,

$$P(X{=}x|Y{=}y) = P(X{=}x), \text{ whenever } P(Y = y) \neq 0.$$

- Meaning: Learning the value of Y does not give me any information about X and vice versa.

Marginal independence

- Two random variables X and Y are **marginally independent**, written $X \perp Y$, if
  - for any state x of X and any state y of Y,

$$P(X{=}x|Y{=}y) = P(X{=}x), \text{ whenever } P(Y = y) \neq 0.$$

- Meaning: Learning the value of Y does not give me any information about X and vice versa. Y contains no information about X and vice versa.

Marginal independence

- Two random variables X and Y are **marginally independent**, written $X \perp Y$, if
  - for any state x of X and any state y of Y,

$$P(X{=}x|Y{=}y) = P(X{=}x), \text{ whenever } P(Y = y) \neq 0.$$

- Meaning: Learning the value of Y does not give me any information about X and vice versa. Y contains no information about X and vice versa.
- Equivalent definition:

$$P(X{=}x, Y{=}y) = P(X{=}x)P(Y{=}y)$$

Marginal independence

- Two random variables X and Y are **marginally independent**, written $X \perp Y$, if
  - for any state x of X and any state y of Y,

$$P(X{=}x|Y{=}y) = P(X{=}x), \text{ whenever } P(Y = y) \neq 0.$$

- Meaning: Learning the value of Y does not give me any information about X and vice versa. Y contains no information about X and vice versa.
- Equivalent definition:

$$P(X{=}x, Y{=}y) = P(X{=}x)P(Y{=}y)$$

- Shorthand for the equations:

Marginal independence

- Two random variables X and Y are **marginally independent**, written $X \perp Y$, if
  - for any state x of X and any state y of Y,

$$P(X{=}x|Y{=}y) = P(X{=}x), \text{ whenever } P(Y = y) \neq 0.$$

- Meaning: Learning the value of Y does not give me any information about X and vice versa. Y contains no information about X and vice versa.
- Equivalent definition:

$$P(X{=}x, Y{=}y) = P(X{=}x)P(Y{=}y)$$

- Shorthand for the equations:

$$P(X|Y) = P(X), P(X,Y) = P(X)P(Y).$$

Marginal independence

- Examples:

Marginal independence

- Examples:
  - X:result of tossing a fair coin for the first time,
    Y: result of second tossing of the same coin.

Marginal independence

- Examples:
  - X:result of tossing a fair coin for the first time,
    Y: result of second tossing of the same coin.
  - X: result of US election, Y: your grades in this course.

Marginal independence

- Examples:
  - X:result of tossing a fair coin for the first time,
    Y: result of second tossing of the same coin.
  - X: result of US election, Y: your grades in this course.
- Counter example:

Marginal independence

- Examples:
  - X:result of tossing a fair coin for the first time,
    Y: result of second tossing of the same coin.
  - X: result of US election, Y: your grades in this course.
- Counter example:X – oral presentation grade , Y – project report grade.

Conditional independence

- Two random variables X and Y are **conditionally independent** given a third variable Z,

Conditional independence

- Two random variables X and Y are **conditionally independent** given a third variable Z,written $X \perp Y|Z$, if

Conditional independence

- Two random variables X and Y are **conditionally independent** given a third variable Z,written $X \perp Y|Z$, if

$$P(X{=}x|Y{=}y, Z{=}z) = P(X{=}x|Z{=}z) \text{ whenever } P(Y{=}y, Z{=}z) \neq 0$$

- Meaning:

Conditional independence

- Two random variables X and Y are **conditionally independent** given a third variable Z,written $X \perp Y|Z$, if

$$P(X{=}x|Y{=}y, Z{=}z) = P(X{=}x|Z{=}z) \text{ whenever } P(Y{=}y, Z{=}z) \neq 0$$

- Meaning:
  - *If I know the state of Z already, then learning the state of Y does not give me additional information about X.*

Conditional independence

- Two random variables X and Y are **conditionally independent** given a third variable Z,written $X \perp Y|Z$, if

$$P(X{=}x|Y{=}y, Z{=}z) = P(X{=}x|Z{=}z) \text{ whenever } P(Y{=}y, Z{=}z) \neq 0$$

- Meaning:
  - *If I know the state of Z already, then learning the state of Y does not give me additional information about X.*
  - Y might contain some information about X.

Conditional independence

- Two random variables X and Y are **conditionally independent** given a third variable Z,written $X \perp Y|Z$, if

$$P(X{=}x|Y{=}y, Z{=}z) = P(X{=}x|Z{=}z) \text{ whenever } P(Y{=}y, Z{=}z) \neq 0$$

- Meaning:
  - *If I know the state of Z already, then learning the state of Y does not give me additional information about X.*
  - Y might contain some information about X.
  - However all the information about X contained in Y are also contained in Z.

Conditional independence

- Two random variables X and Y are **conditionally independent** given a third variable Z,written $X \perp Y|Z$, if

$$P(X{=}x|Y{=}y, Z{=}z) = P(X{=}x|Z{=}z) \text{ whenever } P(Y{=}y, Z{=}z) \neq 0$$

- Meaning:
  - *If I know the state of Z already, then learning the state of Y does not give me additional information about X.*
  - Y might contain some information about X.
  - However all the information about X contained in Y are also contained in Z.

Conditional independence

- Two random variables X and Y are **conditionally independent** given a third variable Z,written $X \perp Y|Z$, if

$$P(X{=}x|Y{=}y, Z{=}z) = P(X{=}x|Z{=}z) \text{ whenever } P(Y{=}y, Z{=}z) \neq 0$$

- Meaning:
  - *If I know the state of Z already, then learning the state of Y does not give me additional information about X.*
  - Y might contain some information about X.
  - However all the information about X contained in Y are also contained in Z.
- Shorthand for the equation:

$$P(X|Y, Z) = P(X|Z)$$

Conditional independence

- Two random variables X and Y are **conditionally independent** given a third variable Z,written $X \perp Y|Z$, if

$$P(X{=}x|Y{=}y, Z{=}z) = P(X{=}x|Z{=}z) \text{ whenever } P(Y{=}y, Z{=}z) \neq 0$$

- Meaning:
  - *If I know the state of Z already, then learning the state of Y does not give me additional information about X.*
  - Y might contain some information about X.
  - However all the information about X contained in Y are also contained in Z.
- Shorthand for the equation:

$$P(X|Y, Z) = P(X|Z)$$

- Equivalent definition:

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

Example of Conditional Independence

- There is a bag of 100 coins.

Example of Conditional Independence

- There is a bag of 100 coins. 10 coins were made by a malfunctioning machine and are biased toward head.

Example of Conditional Independence

- There is a bag of 100 coins. 10 coins were made by a malfunctioning machine and are biased toward head. Tossing such a coin results in head 80% of the time. The other coins are fair.

Example of Conditional Independence

- There is a bag of 100 coins. 10 coins were made by a malfunctioning machine and are biased toward head. Tossing such a coin results in head 80% of the time. The other coins are fair.
- Randomly draw a coin from the bag and toss it a few time.

Example of Conditional Independence

- There is a bag of 100 coins. 10 coins were made by a malfunctioning machine and are biased toward head. Tossing such a coin results in head 80% of the time. The other coins are fair.
- Randomly draw a coin from the bag and toss it a few time.
- X_i: result of the i-th tossing, Y: whether the coin is produced by the malfunctioning machine.

Example of Conditional Independence

- There is a bag of 100 coins. 10 coins were made by a malfunctioning machine and are biased toward head. Tossing such a coin results in head 80% of the time. The other coins are fair.
- Randomly draw a coin from the bag and toss it a few time.
- X_i: result of the i-th tossing, Y: whether the coin is produced by the malfunctioning machine.
- The X_i's are not marginally independent of each other:

Example of Conditional Independence

- There is a bag of 100 coins. 10 coins were made by a malfunctioning machine and are biased toward head. Tossing such a coin results in head 80% of the time. The other coins are fair.
- Randomly draw a coin from the bag and toss it a few time.
- X_i: result of the i-th tossing, Y: whether the coin is produced by the malfunctioning machine.
- The X_i's are not marginally independent of each other:
    - If I get 9 heads in first 10 tosses, then the coin is probably a biased coin.

Example of Conditional Independence

- There is a bag of 100 coins. 10 coins were made by a malfunctioning machine and are biased toward head. Tossing such a coin results in head 80% of the time. The other coins are fair.
- Randomly draw a coin from the bag and toss it a few time.
- X_i: result of the i-th tossing, Y: whether the coin is produced by the malfunctioning machine.
- The X_i's are not marginally independent of each other:
  - If I get 9 heads in first 10 tosses, then the coin is probably a biased coin. Hence the next tossing will be more likely to result in a head than a tail.
  - Learning the value of X_i gives me some information about

Example of Conditional Independence

- There is a bag of 100 coins. 10 coins were made by a malfunctioning machine and are biased toward head. Tossing such a coin results in head 80% of the time. The other coins are fair.
- Randomly draw a coin from the bag and toss it a few time.
- X_i: result of the i-th tossing, Y: whether the coin is produced by the malfunctioning machine.
- The X_i's are not marginally independent of each other:
    - If I get 9 heads in first 10 tosses, then the coin is probably a biased coin. Hence the next tossing will be more likely to result in a head than a tail.
    - Learning the value of X_i gives me some information about whether the coin is biased,

Example of Conditional Independence

- There is a bag of 100 coins. 10 coins were made by a malfunctioning machine and are biased toward head. Tossing such a coin results in head 80% of the time. The other coins are fair.
- Randomly draw a coin from the bag and toss it a few time.
- X_i: result of the i-th tossing, Y: whether the coin is produced by the malfunctioning machine.
- The X_i's are not marginally independent of each other:
  - If I get 9 heads in first 10 tosses, then the coin is probably a biased coin. Hence the next tossing will be more likely to result in a head than a tail.
  - Learning the value of X_i gives me some information about whether the coin is biased, which in term gives me some information about X_j.

Example of Conditional Independence

- However, they are conditionally independent given Y:

Example of Conditional Independence

- However, they are conditionally independent given Y:
  - If the coin is not biased, the probability of getting a head in one toss is $1/2$ regardless of the results of other tosses.

Example of Conditional Independence

- However, they are conditionally independent given Y:
  - If the coin is not biased, the probability of getting a head in one toss is $1/2$ regardless of the results of other tosses.
  - If the coin is biased, the probability of getting a head in one toss is 80% regardless of the results of other tosses.

Example of Conditional Independence

- However, they are conditionally independent given Y:
  - If the coin is not biased, the probability of getting a head in one toss is $1/2$ regardless of the results of other tosses.
  - If the coin is biased, the probability of getting a head in one toss is 80% regardless of the results of other tosses.
  - If I already knows whether the coin is biased or not,

Example of Conditional Independence

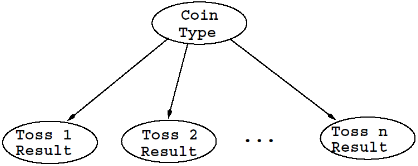
- However, they are conditionally independent given Y:
  - If the coin is not biased, the probability of getting a head in one toss is $1/2$ regardless of the results of other tosses.
  - If the coin is biased, the probability of getting a head in one toss is 80% regardless of the results of other tosses.
  - If I already knows whether the coin is biased or not, learning the value of X_i does not give me additional information about X_j.

Example of Conditional Independence

- However, they are conditionally independent given Y:
  - If the coin is not biased, the probability of getting a head in one toss is $1/2$ regardless of the results of other tosses.
  - If the coin is biased, the probability of getting a head in one toss is 80% regardless of the results of other tosses.
  - If I already knows whether the coin is biased or not, learning the value of X_i does not give me additional information about X_j.
- Here is how the variables are related pictorially. We will return to this picture later.

Example of Conditional Independence

- However, they are conditionally independent given Y:
  - If the coin is not biased, the probability of getting a head in one toss is 1/2 regardless of the results of other tosses.
  - If the coin is biased, the probability of getting a head in one toss is 80% regardless of the results of other tosses.
  - If I already knows whether the coin is biased or not, learning the value of X_i does not give me additional information about X_j.
- Here is how the variables are related pictorially. We will return to this picture later.



Outline

Prior, posterior, and likelihood

- Three important concepts in Bayesian inference.

Prior, posterior, and likelihood

- Three important concepts in Bayesian inference.
- With respect to a piece of evidence: E

Prior, posterior, and likelihood

- Three important concepts in Bayesian inference.
- With respect to a piece of evidence: E
- **Prior probability** $P(H)$:

Prior, posterior, and likelihood

- Three important concepts in Bayesian inference.
- With respect to a piece of evidence: E
- **Prior probability** $P(H)$: belief about a hypothesis before observing evidence.

Prior, posterior, and likelihood

- Three important concepts in Bayesian inference.
- With respect to a piece of evidence: E
- **Prior probability** $P(H)$: belief about a hypothesis before observing evidence.
  - Example: Suppose 10% of people suffer from Hepatitis B. A doctor's prior probability about a new patient suffering from Hepatitis B is 0.1.
- **Posterior probability** $P(H|E)$:

Prior, posterior, and likelihood

- Three important concepts in Bayesian inference.
- With respect to a piece of evidence: E
- **Prior probability** $P(H)$: belief about a hypothesis before observing evidence.
    - Example: Suppose 10% of people suffer from Hepatitis B. A doctor's prior probability about a new patient suffering from Hepatitis B is 0.1.
- **Posterior probability** $P(H|E)$:belief about a hypothesis after obtaining the evidence.

Prior, posterior, and likelihood

- Three important concepts in Bayesian inference.
- With respect to a piece of evidence: E
- **Prior probability** $P(H)$: belief about a hypothesis before observing evidence.
  - Example: Suppose 10% of people suffer from Hepatitis B. A doctor's prior probability about a new patient suffering from Hepatitis B is 0.1.
- **Posterior probability** $P(H|E)$:belief about a hypothesis after obtaining the evidence.
  - If the doctor finds that the eyes of the patient are yellow, his belief about patient suffering from Hepatitis B would be > 0.1.

Prior, posterior, and likelihood

- Suppose a patient is observed to have yellow eyes (E).

Prior, posterior, and likelihood

- Suppose a patient is observed to have yellow eyes (E).
- Consider two possible explanations:

Prior, posterior, and likelihood

- Suppose a patient is observed to have yellow eyes (E).
- Consider two possible explanations:
  1. The patient has Hepatitis B (H_1),

Prior, posterior, and likelihood

- Suppose a patient is observed to have yellow eyes (E).
- Consider two possible explanations:
  1. The patient has Hepatitis B (H_1),
  2. The patient does not have Hepatitis B (H_2)

Prior, posterior, and likelihood

- Suppose a patient is observed to have yellow eyes (E).
- Consider two possible explanations:
  1. The patient has Hepatitis B (H_1),
  2. The patient does not have Hepatitis B (H_2)
- Obviously, H_1 is a better explanation because

Prior, posterior, and likelihood

- Suppose a patient is observed to have yellow eyes (E).
- Consider two possible explanations:
    1. The patient has Hepatitis B (H_1),
    2. The patient does not have Hepatitis B (H_2)
- Obviously, H_1 is a better explanation because $P(E|H_1) > P(E|H_2)$.

Prior, posterior, and likelihood

- Suppose a patient is observed to have yellow eyes (E).
- Consider two possible explanations:
  1. The patient has Hepatitis B (H_1),
  2. The patient does not have Hepatitis B (H_2)
- Obviously, H_1 is a better explanation because $P(E|H_1) > P(E|H_2)$. To state it another way, we say that H_1 is more **likely** than H_2 given E.

Prior, posterior, and likelihood

- Suppose a patient is observed to have yellow eyes (E).
- Consider two possible explanations:
  1. The patient has Hepatitis B (H_1),
  2. The patient does not have Hepatitis B (H_2)
- Obviously, H_1 is a better explanation because $P(E|H_1) > P(E|H_2)$. To state it another way, we say that H_1 is more **likely** than H_2 given E.
- In general, the **likelihood** of a hypothesis H given evidence E is a measure of how well H explains E.

Prior, posterior, and likelihood

- Suppose a patient is observed to have yellow eyes (E).
- Consider two possible explanations:
  1. The patient has Hepatitis B (H_1),
  2. The patient does not have Hepatitis B (H_2)
- Obviously, H_1 is a better explanation because $P(E|H_1) > P(E|H_2)$. To state it another way, we say that H_1 is more **likely** than H_2 given E.
- In general, the **likelihood** of a hypothesis H given evidence E is a measure of how well H explains E. Mathematically, it is

$$L(H|E) = P(E|H)$$

Prior, posterior, and likelihood

- Suppose a patient is observed to have yellow eyes (E).
- Consider two possible explanations:
  1. The patient has Hepatitis B (H_1),
  2. The patient does not have Hepatitis B (H_2)
- Obviously, H_1 is a better explanation because $P(E|H_1) > P(E|H_2)$. To state it another way, we say that H_1 is more **likely** than H_2 given E.
- In general, the **likelihood** of a hypothesis H given evidence E is a measure of how well H explains E. Mathematically, it is

$$L(H|E) = P(E|H)$$

- In Machine Learning, we often talk about the likelihood of a model M given data D.

Prior, posterior, and likelihood

- Suppose a patient is observed to have yellow eyes (E).
- Consider two possible explanations:
  1. The patient has Hepatitis B (H_1),
  2. The patient does not have Hepatitis B (H_2)
- Obviously, H_1 is a better explanation because $P(E|H_1) > P(E|H_2)$. To state it another way, we say that H_1 is more **likely** than H_2 given E.
- In general, the **likelihood** of a hypothesis H given evidence E is a measure of how well H explains E. Mathematically, it is

$$L(H|E) = P(E|H)$$

- In Machine Learning, we often talk about the likelihood of a model M given data D. It is a measure of how well the model M explains the data D.

Prior, posterior, and likelihood

- Suppose a patient is observed to have yellow eyes (E).
- Consider two possible explanations:
  1. The patient has Hepatitis B (H_1),
  2. The patient does not have Hepatitis B (H_2)
- Obviously, H_1 is a better explanation because $P(E|H_1) > P(E|H_2)$. To state it another way, we say that H_1 is more **likely** than H_2 given E.
- In general, the **likelihood** of a hypothesis H given evidence E is a measure of how well H explains E. Mathematically, it is

$$L(H|E) = P(E|H)$$

- In Machine Learning, we often talk about the likelihood of a model M given data D. It is a measure of how well the model M explains the data D. Mathematically, it is

$$L(M|D) = P(D|M)$$

Prior, posterior, and likelihood

- Suppose a patient is observed to have yellow eyes (E).
- Consider two possible explanations:
  1. The patient has Hepatitis B (H_1),
  2. The patient does not have Hepatitis B (H_2)
- Obviously, H_1 is a better explanation because $P(E|H_1) > P(E|H_2)$. To state it another way, we say that H_1 is more **likely** than H_2 given E.
- In general, the **likelihood** of a hypothesis H given evidence E is a measure of how well H explains E. Mathematically, it is

$$L(H|E) = P(E|H)$$

- In Machine Learning, we often talk about the likelihood of a model M given data D. It is a measure of how well the model M explains the data D. Mathematically, it is

$$L(M|D) = P(D|M)$$

Bayes' Theorem/Bayes Rule

- **Bayes' Theorem**:

Bayes' Theorem/Bayes Rule

- **Bayes' Theorem**: relates prior probability, likelihood, and posterior probability:

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)}$$

Bayes' Theorem/Bayes Rule

- **Bayes' Theorem**: relates prior probability, likelihood, and posterior probability:

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)} \propto P(H)L(H|E)$$

Bayes' Theorem/Bayes Rule

- **Bayes' Theorem**: relates prior probability, likelihood, and posterior probability:
$$P(H|E) = \frac{P(H)P(E|H)}{P(E)} \propto P(H)L(H|E)$$
where $P(E)$ is normalization constant to ensure $\sum_{h \in \Omega_H} P(H=h|E) = 1$.

Bayes' Theorem/Bayes Rule

- **Bayes' Theorem**: relates prior probability, likelihood, and posterior probability:

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)} \propto P(H)L(H|E)$$

where $P(E)$ is normalization constant to ensure $\sum_{h \in \Omega_H} P(H = h|E) = 1$.

Bayes' Theorem/Bayes Rule

- **Bayes' Theorem**: relates prior probability, likelihood, and posterior probability:

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)} \propto P(H)L(H|E)$$

where $P(E)$ is normalization constant to ensure $\sum_{h \in \Omega_H} P(H = h|E) = 1$.

That is: posterior $\propto$ prior $\times$ likelihood

Outline

A Simple Problem

- Let X be the result of tossing a thumbtack and $\Omega_X = \{H, T\}$.

A Simple Problem

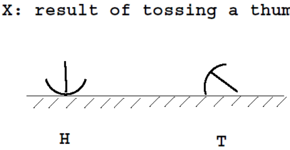
- Let X be the result of tossing a thumbtack and $\Omega_X = \{H, T\}$.
- Data instances:
  $D_1 = H, D_2 = T, D_3 = H, \ldots, D_m = H$

A Simple Problem

- Let X be the result of tossing a thumbtack and $\Omega_X = \{H, T\}$.
- Data instances:
  $D_1 = H,\ D_2 = T,\ D_3 = H,\ \ldots,\ D_m = H$
- Data set: $\mathcal{D} = \{D_1, D_2, D_3, \ldots, D_m\}$

A Simple Problem

- Let X be the result of tossing a thumbtack and $\Omega_X = \{H, T\}$.
- Data instances:
  $D_1 = H,\ D_2 = T,\ D_3 = H,\ \ldots,\ D_m = H$
- Data set: $\mathcal{D} = \{D_1, D_2, D_3, \ldots, D_m\}$
- Task: To estimate parameter $\theta = P(X{=}H)$.



Likelihood

- Data: $\mathcal{D} = \{H, T, H, T, T, H, T\}$

Likelihood

- Data: $\mathcal{D} = \{H, T, H, T, T, H, T\}$
- As possible values of θ, which of the following is the most likely? Why?
  - $\theta = 0$
  - $\theta = 0.01$
  - $\theta = 0.5$
- $\theta = 0$ contradicts data because $P(\mathcal{D}|\theta = 0) = 0$.

Likelihood

- Data: $\mathcal{D} = \{H, T, H, T, T, H, T\}$
- As possible values of θ, which of the following is the most likely? Why?
    - $\theta = 0$
    - $\theta = 0.01$
    - $\theta = 0.5$
- $\theta = 0$ contradicts data because $P(\mathcal{D}|\theta = 0) = 0$.It cannot explain the data at all.

Likelihood

- Data: $\mathcal{D} = \{H, T, H, T, T, H, T\}$
- As possible values of θ, which of the following is the most likely? Why?
  - $\theta = 0$
  - $\theta = 0.01$
  - $\theta = 0.5$
- $\theta = 0$ contradicts data because $P(\mathcal{D}|\theta = 0) = 0$.It cannot explain the data at all.
- $\theta = 0.01$ almost contradicts with the data. It does not explain the data well.

  However, it is more consistent with the data than $\theta = 0$ because $P(\mathcal{D}|\theta = 0.01) > P(\mathcal{D}|\theta = 0)$.

Likelihood

- Data: $\mathcal{D} = \{H, T, H, T, T, H, T\}$
- As possible values of θ, which of the following is the most likely? Why?
    - $\theta = 0$
    - $\theta = 0.01$
    - $\theta = 0.5$
- $\theta = 0$ contradicts data because $P(\mathcal{D}|\theta = 0) = 0$.It cannot explain the data at all.
- $\theta = 0.01$ almost contradicts with the data. It does not explain the data well.

  However, it is more consistent with the data than $\theta = 0$ because $P(\mathcal{D}|\theta = 0.01) > P(\mathcal{D}|\theta = 0)$.
- So $\theta = 0.5$ is more consistent with the data than $\theta = 0.01$ because

Likelihood

- Data: $\mathcal{D} = \{H, T, H, T, T, H, T\}$
- As possible values of θ, which of the following is the most likely? Why?
  - $\theta = 0$
  - $\theta = 0.01$
  - $\theta = 0.5$
- $\theta = 0$ contradicts data because $P(\mathcal{D}|\theta = 0) = 0$.It cannot explain the data at all.
- $\theta = 0.01$ almost contradicts with the data. It does not explain the data well.

  However, it is more consistent with the data than $\theta = 0$ because $P(\mathcal{D}|\theta = 0.01) > P(\mathcal{D}|\theta = 0)$.
- So $\theta = 0.5$ is more consistent with the data than $\theta = 0.01$ because $P(\mathcal{D}|\theta = 0.5) > P(\mathcal{D}|\theta = 0.01)$

  It explains the data the best, and is hence the most likely.

Maximum Likelihood Estimation

- In general, the larger $P(\mathcal{D}|\theta)$ is,

Maximum Likelihood Estimation

- In general, the larger $P(\mathcal{D}|\theta)$ is, the more likely the value θ is.

Maximum Likelihood Estimation

- In general, the larger $P(\mathcal{D}|\theta)$ is, the more likely the value θ is.
- Likelihood of parameter θ given data set:

$$L(\theta|\mathcal{D}) = P(\mathcal{D}|\theta)$$

Maximum Likelihood Estimation

- In general, the larger $P(\mathcal{D}|\theta)$ is, the more likely the value θ is.
- Likelihood of parameter θ given data set:

$$L(\theta|\mathcal{D}) = P(\mathcal{D}|\theta)$$

- The **maximum likelihood estimation (MLE)** θ^* is

Maximum Likelihood Estimation

- In general, the larger $P(\mathcal{D}|\theta)$ is, the more likely the value θ is.
- Likelihood of parameter θ given data set:

$$L(\theta|\mathcal{D}) = P(\mathcal{D}|\theta)$$

- The **maximum likelihood estimation (MLE)** θ^* is

$$L(\theta^*|\mathcal{D}) = \arg\max_{\theta} L(\theta|\mathcal{D}).$$

Maximum Likelihood Estimation

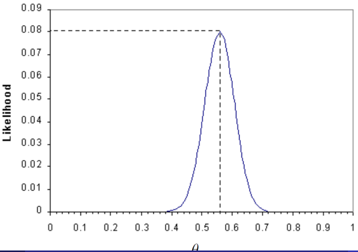
- In general, the larger $P(\mathcal{D}|\theta)$ is, the more likely the value θ is.
- Likelihood of parameter θ given data set:

$$L(\theta|\mathcal{D}) = P(\mathcal{D}|\theta)$$

- The **maximum likelihood estimation (MLE)** θ^* is

$$L(\theta^*|\mathcal{D}) = \arg\max_{\theta} L(\theta|\mathcal{D}).$$

MLE best explains data or best fits data.



i.i.d and Likelihood

- Assume the data instances $D_1, \ldots, D_m$ are independent given θ:

$$P(D_1, \ldots, D_m|\theta) = \prod_{i=1}^{m} P(D_i|\theta)$$

i.i.d and Likelihood

- Assume the data instances D_1, ..., D_m are independent given θ:

$$P(D_1, \ldots, D_m|\theta) = \prod_{i=1}^{m} P(D_i|\theta)$$

- Assume the data instances are identically distributed:

$$P(D_i = H) = \theta, P(D_i = T) = 1-\theta \qquad \text{for all i}$$

i.i.d and Likelihood

- Assume the data instances $D_1, \ldots, D_m$ are independent given θ:

$$P(D_1, \ldots, D_m|\theta) = \prod_{i=1}^{m} P(D_i|\theta)$$

- Assume the data instances are identically distributed:

$$P(D_i = H) = \theta, P(D_i = T) = 1-\theta \qquad \text{for all i}$$

(Note: i.i.d means independent and identically distributed)

i.i.d and Likelihood

- Assume the data instances $D_1, \ldots, D_m$ are independent given θ:

$$P(D_1, \ldots, D_m|\theta) = \prod_{i=1}^{m} P(D_i|\theta)$$

- Assume the data instances are identically distributed:

$$P(D_i = H) = \theta, P(D_i = T) = 1-\theta \qquad \text{for all i}$$

(Note: i.i.d means independent and identically distributed)

- Then

$$L(\theta|\mathcal{D}) = P(\mathcal{D}|\theta)$$

i.i.d and Likelihood

- Assume the data instances $D_1, \ldots, D_m$ are independent given θ:

$$P(D_1, \ldots, D_m|\theta) = \prod_{i=1}^{m} P(D_i|\theta)$$

- Assume the data instances are identically distributed:

$$P(D_i = H) = \theta, P(D_i = T) = 1-\theta \qquad \text{for all i}$$

(Note: i.i.d means independent and identically distributed)

- Then

$$L(\theta|\mathcal{D}) \quad = \quad P(\mathcal{D}|\theta) = P(D_1, \ldots, D_m|\theta)$$

i.i.d and Likelihood

- Assume the data instances $D_1, \ldots, D_m$ are independent given θ:

$$P(D_1, \ldots, D_m|\theta) = \prod_{i=1}^{m} P(D_i|\theta)$$

- Assume the data instances are identically distributed:

$$P(D_i = H) = \theta, P(D_i = T) = 1-\theta \qquad \text{for all i}$$

(Note: i.i.d means independent and identically distributed)

- Then

$$\begin{aligned} L(\theta|\mathcal{D}) &= P(\mathcal{D}|\theta) = P(D_1, \ldots, D_m|\theta) \\ &= \prod_{i=1}^{m} P(D_i|\theta) \end{aligned}$$

i.i.d and Likelihood

- Assume the data instances $D_1, \ldots, D_m$ are independent given θ:

$$P(D_1, \ldots, D_m|\theta) = \prod_{i=1}^{m} P(D_i|\theta)$$

- Assume the data instances are identically distributed:

$$P(D_i = H) = \theta, P(D_i = T) = 1-\theta \qquad \text{for all i}$$

(Note: i.i.d means independent and identically distributed)

- Then

$$\begin{aligned} L(\theta|\mathcal{D}) &= P(\mathcal{D}|\theta) = P(D_1, \ldots, D_m|\theta) \\ &= \prod_{i=1}^{m} P(D_i|\theta) = \theta^{m_h}(1-\theta)^{m_t} \end{aligned} \tag{1}$$

where m_h is the number of heads and m_t is the number of tail.

i.i.d and Likelihood

- Assume the data instances $D_1, \ldots, D_m$ are independent given θ:

$$P(D_1, \ldots, D_m|\theta) = \prod_{i=1}^{m} P(D_i|\theta)$$

- Assume the data instances are identically distributed:

$$P(D_i = H) = \theta, P(D_i = T) = 1-\theta \qquad \text{for all i}$$

(Note: i.i.d means independent and identically distributed)

- Then

$$\begin{aligned} L(\theta|\mathcal{D}) &= P(\mathcal{D}|\theta) = P(D_1, \ldots, D_m|\theta) \\ &= \prod_{i=1}^{m} P(D_i|\theta) = \theta^{m_h}(1-\theta)^{m_t} \end{aligned} \qquad (1)$$

where m_h is the number of heads and m_t is the number of tail. **Binomial likelihood**.

Example of Likelihood Function

- Example: $\mathcal{D} = \{D_1 = H, D_2 T, D_3 = H, D_4 = H, D_5 = T\}$

Example of Likelihood Function

- Example: $\mathcal{D} = \{D_1 = H, D_2 T, D_3 = H, D_4 = H, D_5 = T\}$

$$L(\theta|\mathcal{D}) \quad = \quad P(\mathcal{D}|\theta)$$

Example of Likelihood Function

- Example: $\mathcal{D} = \{D_1 = H, D_2 T, D_3 = H, D_4 = H, D_5 = T\}$

$$
\begin{aligned}
L(\theta|\mathcal{D}) &= P(\mathcal{D}|\theta) \\
&= P(D_1 = H|\theta)P(D_2 = T|\theta)P(D_3 = H|\theta)P(D_4 = H|\theta)P(D_5 = T|\theta)
\end{aligned}
$$

Example of Likelihood Function

- Example: $\mathcal{D} = \{D_1 = H, D_2 T, D_3 = H, D_4 = H, D_5 = T\}$

$$
\begin{aligned}
L(\theta|\mathcal{D}) &= P(\mathcal{D}|\theta) \\
&= P(D_1 = H|\theta)P(D_2 = T|\theta)P(D_3 = H|\theta)P(D_4 = H|\theta)P(D_5 = T|\theta) \\
&= \theta(1-\theta)\theta\theta(1-\theta)
\end{aligned}
$$

Example of Likelihood Function

- Example: $\mathcal{D} = \{D_1 = H, D_2 T, D_3 = H, D_4 = H, D_5 = T\}$

$$
\begin{aligned}
L(\theta|\mathcal{D}) &= P(\mathcal{D}|\theta) \\
&= P(D_1 = H|\theta)P(D_2 = T|\theta)P(D_3 = H|\theta)P(D_4 = H|\theta)P(D_5 = T|\theta) \\
&= \theta(1-\theta)\theta\theta(1-\theta) \\
&= \theta^3(1-\theta)^2.
\end{aligned}
$$

Sufficient Statistic

- A **sufficient statistic** is a function $s(\mathcal{D})$ of data that summarizing the relevant information for computing the likelihood.

Sufficient Statistic

- A **sufficient statistic** is a function $s(\mathcal{D})$ of data that summarizing the relevant information for computing the likelihood. That is

$$s(\mathcal{D}) = s(\mathcal{D}') \Rightarrow L(\theta|\mathcal{D}) = L(\theta|\mathcal{D}')$$

Sufficient Statistic

- A **sufficient statistic** is a function $s(\mathcal{D})$ of data that summarizing the relevant information for computing the likelihood. That is

$$s(\mathcal{D}) = s(\mathcal{D}') \Rightarrow L(\theta|\mathcal{D}) = L(\theta|\mathcal{D}')$$

- Sufficient statistics tell us all there is to know about data.

Sufficient Statistic

- A **sufficient statistic** is a function $s(\mathcal{D})$ of data that summarizing the relevant information for computing the likelihood. That is

$$s(\mathcal{D}) = s(\mathcal{D}') \Rightarrow L(\theta|\mathcal{D}) = L(\theta|\mathcal{D}')$$

- Sufficient statistics tell us all there is to know about data.
- Since $L(\theta|\mathcal{D}) = \theta^{m_h}(1-\theta)^{m_t}$,

Sufficient Statistic

- A **sufficient statistic** is a function $s(\mathcal{D})$ of data that summarizing the relevant information for computing the likelihood. That is

$$s(\mathcal{D}) = s(\mathcal{D}') \Rightarrow L(\theta|\mathcal{D}) = L(\theta|\mathcal{D}')$$

- Sufficient statistics tell us all there is to know about data.
- Since $L(\theta|\mathcal{D}) = \theta^{m_h}(1-\theta)^{m_t}$,
  the pair (m_h, m_t) is a **sufficient statistic**.

Loglikelihood

- **Loglikelihood**:

$$l(\theta|\mathcal{D}) = logL(\theta|\mathcal{D})$$

Loglikelihood

- **Loglikelihood**:

$$l(\theta|\mathcal{D}) = log L(\theta|\mathcal{D}) = log \theta^{m_h}(1-\theta)^{m_t}$$

Loglikelihood

- **Loglikelihood**:

$$l(\theta|\mathcal{D}) = logL(\theta|\mathcal{D}) = log\theta^{m_h}(1-\theta)^{m_t} = m_h log\theta + m_t log(1-\theta)$$

Loglikelihood

- **Loglikelihood**:

$$l(\theta|\mathcal{D}) = logL(\theta|\mathcal{D}) = log\theta^{m_h}(1-\theta)^{m_t} = m_h log\theta + m_t log(1-\theta)$$

Maximizing likelihood is the same as maximizing loglikelihood. The latter is easier.

Loglikelihood

- **Loglikelihood**:

$$l(\theta|\mathcal{D}) = logL(\theta|\mathcal{D}) = log\theta^{m_h}(1-\theta)^{m_t} = m_h log\theta + m_t log(1-\theta)$$

  Maximizing likelihood is the same as maximizing loglikelihood. The latter is easier.

- Taking the derivative of $\frac{dl(\theta|\mathcal{D})}{d\theta}$ and setting it to zero, we get

Loglikelihood

- **Loglikelihood**:

$$l(\theta|\mathcal{D}) = log L(\theta|\mathcal{D}) = log\theta^{m_h}(1-\theta)^{m_t} = m_h log\theta + m_t log(1-\theta)$$

  Maximizing likelihood is the same as maximizing loglikelihood. The latter is easier.

- Taking the derivative of $\frac{dl(\theta|\mathcal{D})}{d\theta}$ and setting it to zero, we get

$$\theta^* = \frac{m_h}{m_h + m_t} = \frac{m_h}{m}$$

Loglikelihood

- **Loglikelihood**:

$$l(\theta|\mathcal{D}) = logL(\theta|\mathcal{D}) = log\theta^{m_h}(1-\theta)^{m_t} = m_h log\theta + m_t log(1-\theta)$$

  Maximizing likelihood is the same as maximizing loglikelihood. The latter is easier.

- Taking the derivative of $\frac{dl(\theta|\mathcal{D})}{d\theta}$ and setting it to zero, we get

$$\theta^* = \frac{m_h}{m_h + m_t} = \frac{m_h}{m}$$

- MLE is intuitive.

Loglikelihood

- **Loglikelihood**:

$$l(\theta|\mathcal{D}) = logL(\theta|\mathcal{D}) = log\theta^{m_h}(1-\theta)^{m_t} = m_h log\theta + m_t log(1-\theta)$$

  Maximizing likelihood is the same as maximizing loglikelihood. The latter is easier.

- Taking the derivative of $\frac{dl(\theta|\mathcal{D})}{d\theta}$ and setting it to zero, we get

$$\theta^* = \frac{m_h}{m_h + m_t} = \frac{m_h}{m}$$

- MLE is intuitive.
- It also has nice properties:

Loglikelihood

- **Loglikelihood**:

$$l(\theta|\mathcal{D}) = logL(\theta|\mathcal{D}) = log\theta^{m_h}(1-\theta)^{m_t} = m_h log\theta + m_t log(1-\theta)$$

  Maximizing likelihood is the same as maximizing loglikelihood. The latter is easier.

- Taking the derivative of $\frac{dl(\theta|\mathcal{D})}{d\theta}$ and setting it to zero, we get

$$\theta^* = \frac{m_h}{m_h + m_t} = \frac{m_h}{m}$$

- MLE is intuitive.
- It also has nice properties:
  - E.g. **Consistence**:

Loglikelihood

- **Loglikelihood**:

$$l(\theta|\mathcal{D}) = logL(\theta|\mathcal{D}) = log\theta^{m_h}(1-\theta)^{m_t} = m_h log\theta + m_t log(1-\theta)$$

  Maximizing likelihood is the same as maximizing loglikelihood. The latter is easier.

- Taking the derivative of $\frac{dl(\theta|\mathcal{D})}{d\theta}$ and setting it to zero, we get

$$\theta^* = \frac{m_h}{m_h + m_t} = \frac{m_h}{m}$$

- MLE is intuitive.
- It also has nice properties:
  - E.g. **Consistence**: θ^* approaches the true value of θ with probability 1 as m goes to infinity.

Drawback of MLE

Drawback of MLE

- Thumbtack tossing:
  - $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.

Drawback of MLE

- Thumbtack tossing:
  - $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
  - Reasonable. Data suggest that the thumbtack is biased toward tail.

Drawback of MLE

- Thumbtack tossing:
  - $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
  - Reasonable. Data suggest that the thumbtack is biased toward tail.
- Coin tossing:
  - Case 1: $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.

Drawback of MLE

- Thumbtack tossing:
  - $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
  - Reasonable. Data suggest that the thumbtack is biased toward tail.
- Coin tossing:
  - Case 1: $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
    - Not reasonable.

Drawback of MLE

- Thumbtack tossing:
  - $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
  - Reasonable. Data suggest that the thumbtack is biased toward tail.
- Coin tossing:
  - Case 1: $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
    - Not reasonable.
    - Our experience (prior) suggests strongly that coins are fair, hence $\theta = 1/2$.

Drawback of MLE

- Thumbtack tossing:
  - $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
  - Reasonable. Data suggest that the thumbtack is biased toward tail.
- Coin tossing:
  - Case 1: $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
    - Not reasonable.
    - Our experience (prior) suggests strongly that coins are fair, hence θ=1/2.
    - The size of the data set is too small to convince us this particular coin is biased.

Drawback of MLE

- Thumbtack tossing:
  - $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
  - Reasonable. Data suggest that the thumbtack is biased toward tail.
- Coin tossing:
  - Case 1: $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
    - Not reasonable.
    - Our experience (prior) suggests strongly that coins are fair, hence θ=1/2.
    - The size of the data set is too small to convince us this particular coin is biased.
    - The fact that we get (3, 7) instead of (5, 5) is probably due to randomness.
  - Case 2: $(m_h, m_t) = (30,000, 70,000)$. MLE: $\theta = 0.3$.

Drawback of MLE

- Thumbtack tossing:
  - $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
  - Reasonable. Data suggest that the thumbtack is biased toward tail.
- Coin tossing:
  - Case 1: $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
    - Not reasonable.
    - Our experience (prior) suggests strongly that coins are fair, hence θ=1/2.
    - The size of the data set is too small to convince us this particular coin is biased.
    - The fact that we get (3, 7) instead of (5, 5) is probably due to randomness.
  - Case 2: $(m_h, m_t) = (30,000, 70,000)$. MLE: $\theta = 0.3$.
    - Reasonable.

Drawback of MLE

- Thumbtack tossing:
  - $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
  - Reasonable. Data suggest that the thumbtack is biased toward tail.
- Coin tossing:
  - Case 1: $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
    - Not reasonable.
    - Our experience (prior) suggests strongly that coins are fair, hence $\theta{=}1/2$.
    - The size of the data set is too small to convince us this particular coin is biased.
    - The fact that we get (3, 7) instead of (5, 5) is probably due to randomness.
  - Case 2: $(m_h, m_t) = (30, 000, 70, 000)$. MLE: $\theta = 0.3$.
    - Reasonable.
    - Data suggest that the coin is after all biased, overshadowing our prior.

Drawback of MLE

- Thumbtack tossing:
  - $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
  - Reasonable. Data suggest that the thumbtack is biased toward tail.
- Coin tossing:
  - Case 1: $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
    - Not reasonable.
    - Our experience (prior) suggests strongly that coins are fair, hence θ=1/2.
    - The size of the data set is too small to convince us this particular coin is biased.
    - The fact that we get (3, 7) instead of (5, 5) is probably due to randomness.
  - Case 2: $(m_h, m_t) = (30,000, 70,000)$. MLE: $\theta = 0.3$.
    - Reasonable.
    - Data suggest that the coin is after all biased, overshadowing our prior.
  - MLE does not differentiate between those two instances.

Drawback of MLE

- Thumbtack tossing:
  - $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
  - Reasonable. Data suggest that the thumbtack is biased toward tail.
- Coin tossing:
  - Case 1: $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
    - Not reasonable.
    - Our experience (prior) suggests strongly that coins are fair, hence θ=1/2.
    - The size of the data set is too small to convince us this particular coin is biased.
    - The fact that we get (3, 7) instead of (5, 5) is probably due to randomness.
  - Case 2: $(m_h, m_t) = (30,000, 70,000)$. MLE: $\theta = 0.3$.
    - Reasonable.
    - Data suggest that the coin is after all biased, overshadowing our prior.
  - MLE does not differentiate between those two instances. It doe not take prior information into account.

Two Views on Parameter Estimation

MLE:

Two Views on Parameter Estimation

MLE:

- Assumes that θ is unknown but fixed parameter.

Two Views on Parameter Estimation

MLE:

- Assumes that θ is unknown but fixed parameter.
- Estimates it using θ^*, the value that maximizes the likelihood function

Two Views on Parameter Estimation

MLE:

- Assumes that θ is unknown but fixed parameter.
- Estimates it using θ^*, the value that maximizes the likelihood function
- Makes prediction based on the estimation: $P(D_{m+1} = H|\mathcal{D}) = \theta^*$

Two Views on Parameter Estimation

MLE:

- Assumes that θ is unknown but fixed parameter.
- Estimates it using θ^*, the value that maximizes the likelihood function
- Makes prediction based on the estimation: $P(D_{m+1} = H|\mathcal{D}) = \theta^*$

Bayesian Estimation:

Two Views on Parameter Estimation

MLE:

- Assumes that θ is unknown but fixed parameter.
- Estimates it using θ^*, the value that maximizes the likelihood function
- Makes prediction based on the estimation: $P(D_{m+1} = H|\mathcal{D}) = \theta^*$

Bayesian Estimation:

- Treats θ as a random variable.

Two Views on Parameter Estimation

MLE:

- Assumes that θ is unknown but fixed parameter.
- Estimates it using θ^*, the value that maximizes the likelihood function
- Makes prediction based on the estimation: $P(D_{m+1} = H|\mathcal{D}) = \theta^*$

Bayesian Estimation:

- Treats θ as a random variable.
- Assumes a prior probability of θ: $p(\theta)$

Two Views on Parameter Estimation

MLE:

- Assumes that θ is unknown but fixed parameter.
- Estimates it using θ^*, the value that maximizes the likelihood function
- Makes prediction based on the estimation: $P(D_{m+1} = H|\mathcal{D}) = \theta^*$

Bayesian Estimation:

- Treats θ as a random variable.
- Assumes a prior probability of θ: $p(\theta)$
- Uses data to get posterior probability of θ: $p(\theta|\mathcal{D})$

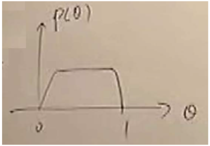
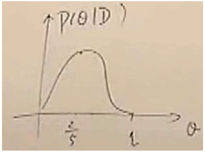
Two Views on Parameter Estimation

MLE:

- Assumes that θ is unknown but fixed parameter.
- Estimates it using θ^*, the value that maximizes the likelihood function
- Makes prediction based on the estimation: $P(D_{m+1} = H|\mathcal{D}) = \theta^*$

Bayesian Estimation:

- Treats θ as a random variable.
- Assumes a prior probability of θ: $p(\theta)$
- Uses data to get posterior probability of θ: $p(\theta|\mathcal{D})$

	Before Seeing Data	After Seeing Data: {2H, 3T}
MLE	?	2/5
Bayesian Estimation		

Two Views on Parameter Estimation

Bayesian Estimation:

- Predicting D_{m+1}

$$P(D_{m+1} = H|\mathcal{D}) = \int P(D_{m+1} = H, \theta|\mathcal{D})d\theta$$

Two Views on Parameter Estimation

Bayesian Estimation:

- Predicting D_{m+1}

$$
\begin{aligned}
P(D_{m+1} = H|\mathcal{D}) &= \int P(D_{m+1} = H, \theta|\mathcal{D}) d\theta \\
&= \int P(D_{m+1} = H|\theta, \mathcal{D}) p(\theta|\mathcal{D}) d\theta
\end{aligned}
$$

Two Views on Parameter Estimation

Bayesian Estimation:

- Predicting D_{m+1}

$$
\begin{aligned}
P(D_{m+1} = H|\mathcal{D}) &= \int P(D_{m+1} = H, \theta|\mathcal{D})d\theta \\
&= \int P(D_{m+1} = H|\theta, \mathcal{D})p(\theta|\mathcal{D})d\theta \\
&= \int P(D_{m+1} = H|\theta)p(\theta|\mathcal{D})d\theta
\end{aligned}
$$

Two Views on Parameter Estimation

Bayesian Estimation:

- Predicting D_{m+1}

$$
\begin{aligned}
P(D_{m+1} = H|\mathcal{D}) &= \int P(D_{m+1} = H, \theta|\mathcal{D}) d\theta \\
&= \int P(D_{m+1} = H|\theta, \mathcal{D}) p(\theta|\mathcal{D}) d\theta \\
&= \int P(D_{m+1} = H|\theta) p(\theta|\mathcal{D}) d\theta \\
&= \int \theta p(\theta|\mathcal{D}) d\theta.
\end{aligned}
$$

Full Bayesian: Take expectation over θ.

Two Views on Parameter Estimation

Bayesian Estimation:

- Predicting D_{m+1}

$$\begin{aligned} P(D_{m+1} = H|\mathcal{D}) &= \int P(D_{m+1} = H, \theta|\mathcal{D}) d\theta \\ &= \int P(D_{m+1} = H|\theta, \mathcal{D}) p(\theta|\mathcal{D}) d\theta \\ &= \int P(D_{m+1} = H|\theta) p(\theta|\mathcal{D}) d\theta \\ &= \int \theta p(\theta|\mathcal{D}) d\theta. \end{aligned}$$

Full Bayesian: Take expectation over θ.

Two Views on Parameter Estimation

Bayesian Estimation:

- Predicting D_{m+1}

$$
\begin{aligned}
P(D_{m+1} = H|\mathcal{D}) &= \int P(D_{m+1} = H, \theta|\mathcal{D})d\theta \\
&= \int P(D_{m+1} = H|\theta, \mathcal{D})p(\theta|\mathcal{D})d\theta \\
&= \int P(D_{m+1} = H|\theta)p(\theta|\mathcal{D})d\theta \\
&= \int \theta p(\theta|\mathcal{D})d\theta.
\end{aligned}
$$

  Full Bayesian: Take expectation over θ.

- **Bayesian MAP**:

Two Views on Parameter Estimation

Bayesian Estimation:

- Predicting D_{m+1}

$$\begin{aligned} P(D_{m+1}=H|\mathcal{D}) &= \int P(D_{m+1}=H,\theta|\mathcal{D})d\theta \\ &= \int P(D_{m+1}=H|\theta,\mathcal{D})p(\theta|\mathcal{D})d\theta \\ &= \int P(D_{m+1}=H|\theta)p(\theta|\mathcal{D})d\theta \\ &= \int \theta p(\theta|\mathcal{D})d\theta. \end{aligned}$$

  Full Bayesian: Take expectation over θ.

- **Bayesian MAP**:

$$P(D_{m+1}=H|\mathcal{D}) = \theta^* = \arg\max p(\theta|\mathcal{D})$$

Calculating Bayesian Estimation

- Posterior distribution:

$$p(\theta|\mathcal{D}) \quad \propto \quad p(\theta)L(\theta|\mathcal{D})$$

Calculating Bayesian Estimation

- Posterior distribution:

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto p(\theta)L(\theta|\mathcal{D}) \\ &= \theta^{m_h}(1-\theta)^{m_t}p(\theta) \end{aligned}$$

where the equation follows from (1)

Calculating Bayesian Estimation

- Posterior distribution:

$$
\begin{aligned}
p(\theta|\mathcal{D}) &\propto p(\theta)L(\theta|\mathcal{D}) \\
&= \theta^{m_h}(1-\theta)^{m_t}p(\theta)
\end{aligned}
$$

  where the equation follows from (1)

- To facilitate analysis, assume prior has **Beta distribution** $B(\alpha_h, \alpha_t)$

$$p(\theta) \propto \theta^{\alpha_h-1}(1-\theta)^{\alpha_t-1}$$

Calculating Bayesian Estimation

- Posterior distribution:

$$
\begin{aligned}
p(\theta|\mathcal{D}) &\propto p(\theta)L(\theta|\mathcal{D}) \\
&= \theta^{m_h}(1-\theta)^{m_t}p(\theta)
\end{aligned}
$$

where the equation follows from (1)

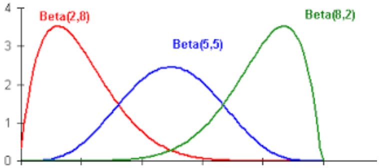
- To facilitate analysis, assume prior has **Beta distribution** $B(\alpha_h, \alpha_t)$

$$
p(\theta) \propto \theta^{\alpha_h - 1}(1-\theta)^{\alpha_t - 1}
$$

- Then

$$
p(\theta|\mathcal{D}) \propto \theta^{m_h+\alpha_h-1}(1-\theta)^{m_t+\alpha_t-1} \tag{2}
$$

Beta Distribution

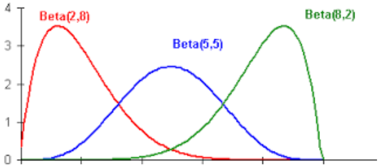


- The normalization constant for the Beta distribution $B(\alpha_h, \alpha_t)$

$$\frac{\Gamma(\alpha_t + \alpha_h)}{\Gamma(\alpha_t)\Gamma(\alpha_h)}$$

where $\Gamma(.)$ is the **Gamma function**. For any integer α, $\Gamma(\alpha) = (\alpha - 1)!$. It is also defined for non-integers.

Beta Distribution



- Density function of prior Beta distribution $B(\alpha_h, \alpha_t)$,

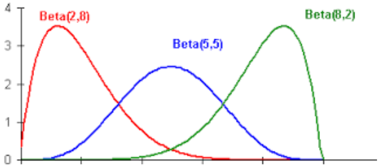
$$p(\theta) = \frac{\Gamma(\alpha_t + \alpha_h)}{\Gamma(\alpha_t)\Gamma(\alpha_h)} \theta^{\alpha_h - 1}(1-\theta)^{\alpha_t}$$

- The normalization constant for the Beta distribution $B(\alpha_h, \alpha_t)$

$$\frac{\Gamma(\alpha_t + \alpha_h)}{\Gamma(\alpha_t)\Gamma(\alpha_h)}$$

where $\Gamma(.)$ is the **Gamma function**. For any integer α, $\Gamma(\alpha) = (\alpha - 1)!$. It is also defined for non-integers.

Beta Distribution



- The normalization constant for the Beta distribution $B(\alpha_h, \alpha_t)$

$$\frac{\Gamma(\alpha_t + \alpha_h)}{\Gamma(\alpha_t)\Gamma(\alpha_h)}$$

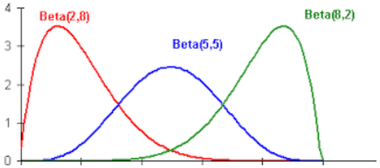
where $\Gamma(.)$ is the **Gamma function**. For any integer α, $\Gamma(\alpha) = (\alpha - 1)!$. It is also defined for non-integers.

- Density function of prior Beta distribution $B(\alpha_h, \alpha_t)$,

$$p(\theta) = \frac{\Gamma(\alpha_t + \alpha_h)}{\Gamma(\alpha_t)\Gamma(\alpha_h)}\theta^{\alpha_h - 1}(1-\theta)^{\alpha_t}$$

- The **hyperparameters** α_h and α_t can be thought of as "imaginary" counts from our prior experiences.

Beta Distribution



- The normalization constant for the Beta distribution $B(\alpha_h, \alpha_t)$

$$\frac{\Gamma(\alpha_t + \alpha_h)}{\Gamma(\alpha_t)\Gamma(\alpha_h)}$$

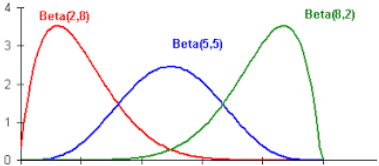
where $\Gamma(.)$ is the **Gamma function**. For any integer α, $\Gamma(\alpha) = (\alpha - 1)!$. It is also defined for non-integers.

- Density function of prior Beta distribution $B(\alpha_h, \alpha_t)$,

$$p(\theta) = \frac{\Gamma(\alpha_t + \alpha_h)}{\Gamma(\alpha_t)\Gamma(\alpha_h)} \theta^{\alpha_h - 1}(1-\theta)^{\alpha_t}$$

- The **hyperparameters** α_h and α_t can be thought of as "imaginary" counts from our prior experiences.
- Their sum $\alpha = \alpha_h + \alpha_t$ is called **equivalent sample size**.

Beta Distribution



- The normalization constant for the Beta distribution $B(\alpha_h, \alpha_t)$

$$\frac{\Gamma(\alpha_t + \alpha_h)}{\Gamma(\alpha_t)\Gamma(\alpha_h)}$$

where $\Gamma(.)$ is the **Gamma function**. For any integer α, $\Gamma(\alpha) = (\alpha - 1)!$. It is also defined for non-integers.

- Density function of prior Beta distribution $B(\alpha_h, \alpha_t)$,

$$p(\theta) = \frac{\Gamma(\alpha_t + \alpha_h)}{\Gamma(\alpha_t)\Gamma(\alpha_h)} \theta^{\alpha_h - 1}(1 - \theta)^{\alpha_t}$$

- The **hyperparameters** α_h and α_t can be thought of as "imaginary" counts from our prior experiences.
- Their sum $\alpha = \alpha_h + \alpha_t$ is called **equivalent sample size**.
- The larger the equivalent sample size, the more confident we are in our prior.

Conjugate Families

- Binomial Likelihood: $\theta^{m_h}(1-\theta)^{m_t}$

Conjugate Families

- Binomial Likelihood: $\theta^{m_h}(1-\theta)^{m_t}$
- Beta Prior: $\theta^{\alpha_h-1}(1-\theta)^{\alpha_t-1}$

Conjugate Families

- Binomial Likelihood: $\theta^{m_h}(1-\theta)^{m_t}$
- Beta Prior: $\theta^{\alpha_h-1}(1-\theta)^{\alpha_t-1}$

Conjugate Families

- Binomial Likelihood: $\theta^{m_h}(1-\theta)^{m_t}$
- Beta Prior: $\theta^{\alpha_h-1}(1-\theta)^{\alpha_t-1}$
- Beta Posterior: $\theta^{m_h+\alpha_h-1}(1-\theta)^{m_t+\alpha_t-1}$.

Conjugate Families

- Binomial Likelihood: $\theta^{m_h}(1-\theta)^{m_t}$
- Beta Prior: $\theta^{\alpha_h-1}(1-\theta)^{\alpha_t-1}$
- Beta Posterior: $\theta^{m_h+\alpha_h-1}(1-\theta)^{m_t+\alpha_t-1}$.
- Beta distributions are hence called a **conjugate family** for Binomial likelihood.

Conjugate Families

- Binomial Likelihood: $\theta^{m_h}(1-\theta)^{m_t}$
- Beta Prior: $\theta^{\alpha_h-1}(1-\theta)^{\alpha_t-1}$
- Beta Posterior: $\theta^{m_h+\alpha_h-1}(1-\theta)^{m_t+\alpha_t-1}$.
- Beta distributions are hence called a **conjugate family** for Binomial likelihood.
- Conjugate families allow closed-form for posterior distribution of parameters and closed-form solution for prediction.

Calculating Prediction

- We have

Calculating Prediction

- We have

$$P(D_{m+1} = H|\mathcal{D}) = \int \theta p(\theta|\mathcal{D}) d\theta$$

Calculating Prediction

- We have

$$\begin{aligned} P(D_{m+1} = H|\mathcal{D}) &= \int \theta p(\theta|\mathcal{D}) d\theta \\ &= c \int \theta \theta^{m_h+\alpha_h-1}(1-\theta)^{m_t+\alpha_t-1} d\theta \end{aligned}$$

Calculating Prediction

- We have

$$
\begin{aligned}
P(D_{m+1} = H|\mathcal{D}) &= \int \theta p(\theta|\mathcal{D}) d\theta \\
&= c \int \theta \theta^{m_h+\alpha_h-1}(1-\theta)^{m_t+\alpha_t-1} d\theta \\
&= \frac{m_h+\alpha_h}{m+\alpha}
\end{aligned}
$$

where c is the normalization constant, $m = m_h + m_t$, $\alpha = \alpha_h + \alpha_t$.

Calculating Prediction

- We have

$$
\begin{aligned}
P(D_{m+1} = H|\mathcal{D}) &= \int \theta p(\theta|\mathcal{D}) d\theta \\
&= c \int \theta \theta^{m_h+\alpha_h-1}(1-\theta)^{m_t+\alpha_t-1} d\theta \\
&= \frac{m_h+\alpha_h}{m+\alpha}
\end{aligned}
$$

where c is the normalization constant, $m{=}m_h{+}m_t$, $\alpha = \alpha_h{+}\alpha_t$.

- Consequently,

$$
P(D_{m+1} = T|\mathcal{D}) = \frac{m_t+\alpha_t}{m+\alpha}
$$

Calculating Prediction

- We have

$$\begin{aligned} P(D_{m+1} = H|\mathcal{D}) &= \int \theta p(\theta|\mathcal{D}) d\theta \\ &= c \int \theta \theta^{m_h+\alpha_h-1}(1-\theta)^{m_t+\alpha_t-1} d\theta \\ &= \frac{m_h+\alpha_h}{m+\alpha} \end{aligned}$$

  where c is the normalization constant, $m = m_h + m_t$, $\alpha = \alpha_h + \alpha_t$.

- Consequently,

$$P(D_{m+1} = T|\mathcal{D}) = \frac{m_t+\alpha_t}{m+\alpha}$$

- After taking data $\mathcal{D}$ into consideration, now our **updated belief** on $X = T$ is $\frac{m_t+\alpha_t}{m+\alpha}$.

MLE and Bayesian estimation

- As m goes to infinity, $P(D_{m+1} = H|\mathcal{D})$ approaches the MLE $\frac{m_h}{m_h+m_t}$,

MLE and Bayesian estimation

- As m goes to infinity, $P(D_{m+1} = H|\mathcal{D})$ approaches the MLE $\frac{m_h}{m_h+m_t}$,which approaches the true value of θ with probability 1.

MLE and Bayesian estimation

- As m goes to infinity, $P(D_{m+1} = H|\mathcal{D})$ approaches the MLE $\frac{m_h}{m_h+m_t}$,which approaches the true value of θ with probability 1.
- Coin tossing example revisited:

MLE and Bayesian estimation

- As m goes to infinity, $P(D_{m+1} = H|\mathcal{D})$ approaches the MLE $\frac{m_h}{m_h+m_t}$,which approaches the true value of θ with probability 1.
- Coin tossing example revisited:
  - Suppose $\alpha_h = \alpha_t = 100$. Equivalent sample size: 200

MLE and Bayesian estimation

- As m goes to infinity, $P(D_{m+1} = H|\mathcal{D})$ approaches the MLE $\frac{m_h}{m_h+m_t}$,which approaches the true value of θ with probability 1.
- Coin tossing example revisited:
  - Suppose $\alpha_h = \alpha_t = 100$. Equivalent sample size: 200
  - In case 1,

$$P(D_{m+1} = H|\mathcal{D}) = \frac{3+100}{10+100+100} \approx 0.5$$

  Our prior prevails.

MLE and Bayesian estimation

- As m goes to infinity, $P(D_{m+1} = H|\mathcal{D})$ approaches the MLE $\frac{m_h}{m_h+m_t}$,which approaches the true value of θ with probability 1.
- Coin tossing example revisited:
  - Suppose $\alpha_h = \alpha_t = 100$. Equivalent sample size: 200
  - In case 1,

    $$P(D_{m+1} = H|\mathcal{D}) = \frac{3+100}{10+100+100} \approx 0.5$$

    Our prior prevails.
  - In case 2,

    $$P(D_{m+1} = H|\mathcal{D}) = \frac{30,000+100}{100,0000+100+100} \approx 0.3$$

    Data prevail.

MLE vs Bayesian Estimation

- Much of Machine Learning is about parameter estimation.

MLE vs Bayesian Estimation

- Much of Machine Learning is about parameter estimation.
- In all case, both MLE and Bayesian estimations can used, although the latter is harder mathematically.

MLE vs Bayesian Estimation

- Much of Machine Learning is about parameter estimation.
- In all case, both MLE and Bayesian estimations can used, although the latter is harder mathematically.
- In this course, we will focus on MLE.