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

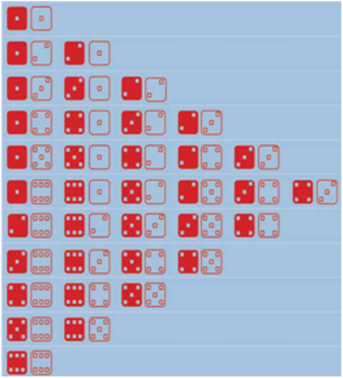
Tail Head

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 dice.

Ω

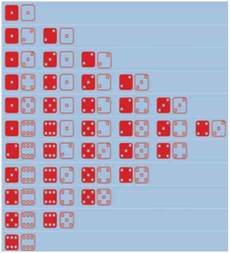


- Elements in a sample space are outcomes.

Events

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

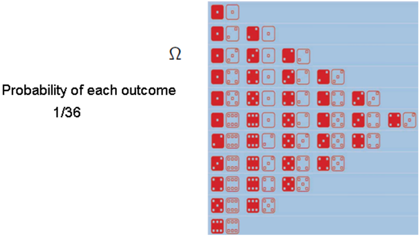
Ω



- Example: The two results add to 4.

Probability Weight Function

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



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: $P(E) = \sum_{\omega \in E} P(\omega)$

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

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

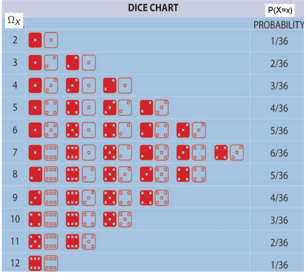
that satisfies Kolmogorov's axioms:

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. $X((2,5)) = 7; X((3,1)) = 4)$



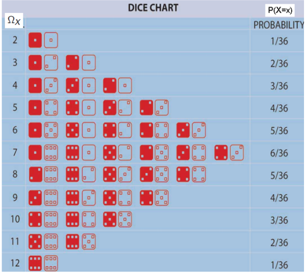
- Why is it random? The experiment.
- **Domain** of a random variable: Set of all its possible values.

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

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

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

- $P(X = 4) = P(\{(1, 3), (2, 2,)(3, 1)\}) = \frac{3}{36}$.
- 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. $\Omega_X = \{\text{H}, \text{T}\}$
  - $P(X{=}\text{H}) = 1/2$ 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:

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

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

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.
  - 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

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).
- 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**: relates prior probability, likelihood, and posterior probability:

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

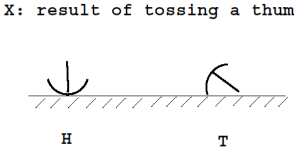
where $P(E)$ is normalization constant to ensure $\sum_{h \in \Omega_H} P(H = h|E) = 1$.

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

Outline

A Simple Problem

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

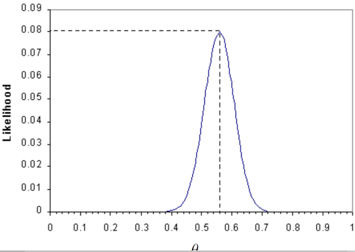
- In general, the larger $P(\mathcal{D}|\theta)$ is, the more likely the value θ is.
- Likelihood of parameter θ given data set:

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

- 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:

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

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

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

Sufficient Statistic

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

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

  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

- 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.

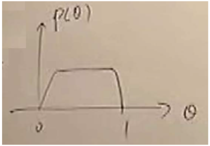
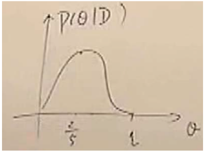
Two Views on Parameter Estimation

MLE:

- 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}

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

  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:

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

  where the equation follows from (1)

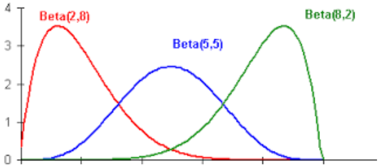
- To facilitate analysis, assume prior has **Beta distribution** $B(\alpha_h, \alpha_t)$

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

- Then

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

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.

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

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

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 $\frac{m_h}{m_h+m_t}$,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.
- In all case, both MLE and Bayesian estimations can used, although the latter is harder mathematically.
- In this course, we will focus on MLE.