Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung

Thang Vu

Math for ML: Probabilities

Contents

Questionnaire

1

Questionnaire: Linear Algebra & Calculus

- Please go to Ilias
- Open the questionnaire 'Algebra & Calculus'
- Note that the questionnaire is anonymous, meaning we only receive the final statistics and responses, not the identities of the individuals who submitted them.

Introduction

Introduction to Probability

Probability is the study of uncertainty, often interpreted as:

- The frequency of events occurring over repeated trials
- A degree of belief about the likelihood of an event

In machine learning, probability is used to model uncertainty in different parts:

- Data
- Model parameters
- Predictions

An Example

Throwing a dice: we want to calculate how likely we will get 6.

- All possible outcomes: 6
- There is only one side of a dice with 6 dots
- **Probability** ($P(A)$) = $1/6 \approx 0,167$

An Example

Throwing a dice: we want to calculate how likely a **double** comes up when throwing 2 dice.

- All possible outcomes of an experiment: 36 including (1,1), (1,2), (1,3), (1,4), ..., (6,3), (6,4), (6,5), (6,6), known as **Sample space** (Ω)
- 6 sample points: (1,1), (2,2), (3,3), (4,4), (5,5) and (6,6), known as **Event space** (A): a set of events, subsets of Ω:
- **Probability** ($P(A)$) = $6/36 \approx 0,167$

Constructing a Probability Space

Probability spaces consist of three components:

- **Sample space** (Ω): All possible outcomes of an experiment.
- **Event space** (A): The event space is the space of potential results of the experiment. A set of events, a subset of Ω
- **Probability** (P): With each event A, we associate a number $P(A)$ that measures the probability or degree of belief that the event will occur. $P(A)$ is called the probability of A.

Another Example

Throwing a dice: we want to calculate how likely the **sum** is equal to 7 when throwing 2 dice.

- **Sample space** (Ω): 36 including (1,1), (1,2), (1,3), (1,4), ..., (6,3), (6,4), (6,5), (6,6)
- **Event space** (A): (1,6), (2,5), (3,4), (4,3), (5,2) and (6,1); thus 6 sample points
- **Probability** $P(A) = 6/36 = 0,167$

Probability and Random Variables in ML

Three central concepts for probability in ML:

- **Probability space**: Foundation for defining probabilities on events
- **Random variables**: It takes an outcome and returns a particular quantity of interest - a value in a target space.
- **Probability distribution**: a function that measures the probability that a particular outcome (or set of outcomes) will occur, foundation to ML tasks, such as classification and regression

This framework underpins the probabilistic reasoning used to handle uncertainty in ML.

An Example

Throwing a dice: we want to calculate how likely the **sum** is equal to 7 when throwing 2 dice.

- **Sample space** (Ω): 36 including (1,1), (1,2), (1,3), (1,4), ..., (6,3), (6,4), (6,5), (6,6)
- **Event space** (A): (1,6), (2,5), (3,4), (4,3), (5,2) and (6,2); thus 6 sample points
- Now the **random variable** is $X = 7$
- **Probability**:
  $P(X = 7) = P(1,6) + P(2,5) + P(3,4) + P(4,3) + P(5,2) + P(6,1) =$
  $1/36 + 1/36 + 1/36 + 1/36 + 1/36 + 1/36 = 6/36 = 0,167$

Random Variables

A **random variable** maps outcomes in the sample space to values of interest:

- Discrete: Maps to finite or countable outcomes
- Continuous: Maps to real values, often representing measurements like height or temperature

Random Variables

The random variable X assigns a particular quantity of interest (the **target space** $\mathcal{T}$) to an outcome of an experiment, $X : \Omega \to \mathcal{T}$

A probability distribution describes the probabilities associated with each possible value of the random variable.

An Example

Consider a statistical experiment where we model a funfair game consisting of drawing two coins from a bag (with replacement). There are coins from USA (denoted as \$) and UK (denoted as £) in the bag, and since we draw two coins from the bag, there are four outcomes in total. The state space or sample space Ω of this experiment is then (\$, \$), (\$, £), (£, \$), (£, £). Let us assume that the composition of the bag of coins is such that a draw returns at random a \$ with probability 0.3.

The event we are interested in is the total number of times the repeated draw returns \$. Let us define a random variable X that maps the sample space Ω to $\mathcal{T}$, which denotes the number of times we draw \$ out of the bag. We can see from the preceding sample space we can get zero \$, one \$, or two \$s, and therefore $\mathcal{T} = \{0, 1, 2\}$. The random variable X (a function or lookup table) can be represented as a table like the following:

$$X((\$, \$)) = 2 \tag{6.1}$$
$$X((\$, £)) = 1 \tag{6.2}$$
$$X((£, \$)) = 1 \tag{6.3}$$
$$X((£, £)) = 0\,. \tag{6.4}$$

Figure: Example from `https://mml-book.github.io/book/mml-book.pdf`

An Example

Since we return the first coin we draw before drawing the second, this implies that the two draws are independent of each other, which we will discuss in Section 6.4.5. Note that there are two experimental outcomes, which map to the same event, where only one of the draws returns \$. Therefore, the probability mass function (Section 6.2.1) of X is given by

$$
\begin{aligned}
P(X=2) &= P((\$,\$)) \\
&= P(\$) \cdot P(\$) \\
&= 0.3 \cdot 0.3 = 0.09 && (6.5) \\
P(X=1) &= P((\$,£) \cup (£,\$)) \\
&= P((\$,£)) + P((£,\$)) \\
&= 0.3 \cdot (1-0.3) + (1-0.3) \cdot 0.3 = 0.42 && (6.6) \\
P(X=0) &= P((£,£)) \\
&= P(£) \cdot P(£) \\
&= (1-0.3) \cdot (1-0.3) = 0.49\,. && (6.7)
\end{aligned}
$$

Figure: Example `https://mml-book.github.io/book/mml-book.pdf`

Bayesian vs. Frequentist Interpretations

Probability interpretations in machine learning:

- **Bayesian**: Probability as a degree of belief, updating with new evidence via Bayes' theorem
- **Frequentist**: Probability as the long-run frequency of events in repeated experiments

Bayesian approaches allow incorporating prior knowledge, while frequentist approaches rely solely on the observed data.

Automated Reasoning with Probability

Probability generalizes logical reasoning,

- Extends binary true/false values to continuous plausibilities
- Allows updating of beliefs based on observed events, as in daily decisions (e.g., predicting traffic delays)

These ideas form the foundation of probabilistic models used in ML for prediction and decision-making.

Statistics vs. Probability in ML

- Probability theory and statistics are often introduced together, but they focus on different aspects of uncertainty.
- Using probability, we can model a process where the underlying uncertainty is represented by random variables, and we use probability rules to determine what happens.
- In statistics, we observe that something has occurred and attempt to understand the underlying process that accounts for the observations.
- In this sense, machine learning is similar to statistics in its goal of building a model that accurately represents the process that generated the data. We can apply the rules of probability to find a “best-fitting” model for some data.

Live Voting

Discrete and Continuous Probabilities

3

Discrete Probabilities – I

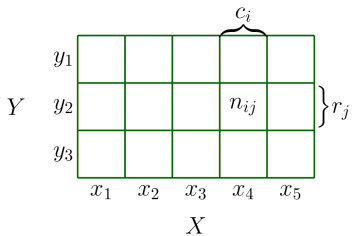
Discrete Target Space

- When the target space T is discrete, we can specify the probability that a random variable X takes a particular value $x \in T$, denoted as $P(X = x)$
- The expression $P(X = x)$ for a discrete random variable X is known as the **probability mass function**.

Discrete Probabilities – I

Discrete Target Space Representation:

- When the target space is discrete, the probability distribution of multiple random variables can be visualized as filling out a multidimensional array of numbers.
- The target space of the joint probability is the Cartesian product of the target spaces of each random variable. Note that if set $A = \{1, 2\}$ and set $B = \{a, b\}$, their Cartesian product $A \times B$ is: $\{(1, a), (1, b), (2, a), (2, b)\}$



Discrete Probabilities – I

Definition of Joint Probability

For discrete random variables X and Y, the joint probability is defined as:

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

where:

- n_{ij} is the number of events where $X = x_i$ and $Y = y_j$.
- N is the total number of events.

Discrete Probabilities – II

Interpretation:

- The joint probability $P(X = x_i, Y = y_j)$ represents the probability of the intersection of both events $X = x_i$ and $Y = y_j$.
- Mathematically, this is expressed as:

$$P(X = x_i, Y = y_j) = P(X = x_i \cap Y = y_j)$$

Discrete Marginal Probabilities

Marginal Probability:

- The marginal probability that X takes the value x, irrespective of the value of random variable Y, is written as $p(x)$.
- We write $X \sim p(x)$ to denote that the random variable X is distributed according to $p(x)$.

How to compute the discrete marginal probabilities:

$$p(x_i) = P(X = x_i) = \frac{c_i}{N}$$

where:

- c_i is the number of events where $X = x_i$, indepedent from Y.
- N is the total number of events.

Discrete Conditional Probabilities

Conditional Probability:

- If we consider only the instances where $X = x$, then the fraction of those instances for which $Y = y$ is written as the conditional probability $p(y \mid x)$.

How to compute the discrete conditional probabilities:

$$p(y_j \mid x_i) = P(Y = y_j \mid X = x_i) = \frac{n_{ij}}{c_i}$$

where:

- n_{ij} is is the number of events where $X = x_i$ and $Y = y_j$.
- c_i is the number of events where $X = x_i$, indepedent from Y.

Discrete Marginal and Conditional Probabilities

In short,

- Marginal probability $p(x)$ represents the probability distribution of X alone.
- Conditional probability $p(y \mid x)$ represents the probability of $Y = y$ given that $X = x$.

An Example

Example 6.2
Consider two random variables X and Y, where X has five possible states and Y has three possible states, as shown in Figure 6.2. We denote by n_{ij} the number of events with state $X = x_i$ and $Y = y_j$, and denote by N the total number of events. The value c_i is the sum of the individual frequencies for the ith column, that is, $c_i = \sum_{j=1}^{3} n_{ij}$. Similarly, the value r_j is the row sum, that is, $r_j = \sum_{i=1}^{5} n_{ij}$. Using these definitions, we can compactly express the distribution of X and Y.

Figure: Example from `https://mml-book.github.io/book/mml-book.pdf`

An Example

The probability distribution of each random variable, the marginal probability, can be seen as the sum over a row or column

$$P(X = x_i) = \frac{c_i}{N} = \frac{\sum_{j=1}^{3} n_{ij}}{N} \tag{6.10}$$

and

$$P(Y = y_j) = \frac{r_j}{N} = \frac{\sum_{i=1}^{5} n_{ij}}{N}, \tag{6.11}$$

where c_i and r_j are the ith column and jth row of the probability table, respectively. By convention, for discrete random variables with a finite number of events, we assume that probabilties sum up to one, that is,

$$\sum_{i=1}^{5} P(X = x_i) = 1 \quad \text{and} \quad \sum_{j=1}^{3} P(Y = y_j) = 1\,. \tag{6.12}$$

The conditional probability is the fraction of a row or column in a par-

Figure: Example from `https://mml-book.github.io/book/mml-book.pdf`

An Example

ticular cell. For example, the conditional probability of Y given X is

$$P(Y = y_j \,|\, X = x_i) = \frac{n_{ij}}{c_i}\,, \tag{6.13}$$

and the conditional probability of X given Y is

$$P(X = x_i \,|\, Y = y_j) = \frac{n_{ij}}{r_j}\,. \tag{6.14}$$

Figure: Example from `https://mml-book.github.io/book/mml-book.pdf`

Continuous Probabilities – I

Definition (Probability Density Function)

A function $f : \mathbb{R}^D \to \mathbb{R}$ is called a probability density function (pdf) if:

- $\forall x \in \mathbb{R}^D : f(x) \geq 0$
- Its integral over $\mathbb{R}^D$ exists and

$$\int_{\mathbb{R}^D} f(x)\, dx = 1$$

- For probability mass functions (pmf) of discrete random variables, the integral is replaced with a sum.

Continuous Probabilities – II

Association with Random Variables:

- A probability density function is any function f that is non-negative and integrates to one.
- We associate a random variable X with this function f by:

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

  where $a, b \in \mathbb{R}$ and $x \in \mathbb{R}$ are outcomes of the continuous random variable X.
- For states $x \in \mathbb{R}^D$, this is extended by considering vectors in $\mathbb{R}^D$.
- This association is called the law or distribution of the random variable X.

Cumulative Distribution Function

Definition (Cumulative Distribution Function)

The cumulative distribution function (cdf) of a multivariate real-valued random variable X with states $x \in \mathbb{R}^D$ is given by:

$$F_X(x) = P(X_1 \leq x_1, \ldots, X_D \leq x_D)$$

where $X = [X_1, \ldots, X_D]^\top$, $x = [x_1, \ldots, x_D]^\top$, and the right-hand side represents the probability that each random variable X_i takes a value less than or equal to x_i.

Discrete vs. Continuous Probabilities

Type	"Point probability"	"Interval probability"
Discrete	$P(X = x)$ Probability mass function	Not applicable
Continuous	$p(x)$ Probability density function	$P(X \leqslant x)$ Cumulative distribution function

Figure: Example from `https://mml-book.github.io/book/mml-book.pdf`

Live Voting

Operations on Probability Distributions

4

The Sum Rule in Probability

The Sum Rule:

- The sum rule allows us to compute the marginal probability $p(x)$ from the joint probability $p(x, y)$.

For Discrete Random Variables:

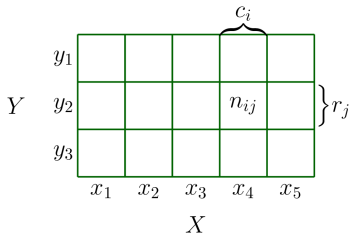
$$p(x) = \sum_{y \in Y} p(x, y)$$

For Continuous Random Variables:

$$p(x) = \int_{Y} p(x, y)\, dy$$

where Y represents the states of the target space of the random variable Y.

Discrete Marginal Probabilities and the Sum Rule



How to compute the discrete marginal probabilities with the sum rule

$$P(X = x_4) = P(X = x_4, Y = y_1) + P(X = x_4, Y = y_2) + P(X = x_4, Y = y_3) \quad (1)$$

$$= \frac{n_{41}}{N} + \frac{n_{42}}{N} + \frac{n_{43}}{N} = \frac{c_4}{N} \quad (2)$$

The Product Rule in Probability

The Product Rule:

- The product rule relates the joint distribution to the conditional distribution.

Mathematical Formulation:

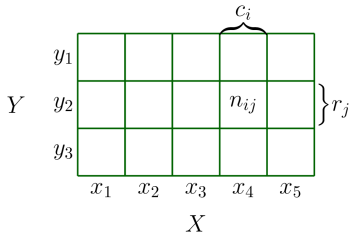
$$p(x, y) = p(y \mid x)\, p(x)$$

Interpretation:

- Every joint distribution of two random variables can be factorized (written as a product) of:
  - The marginal distribution of the first random variable $p(x)$.
  - The conditional distribution of the second random variable given the first $p(y \mid x)$.
- Since the ordering of random variables is arbitrary in $p(x, y)$, the product rule also implies:

$$p(x, y) = p(x \mid y)\, p(y)$$

Discrete Joint Probabilities and the Product Rule



How to compute the discrete marginal probabilities with the sum rule

$$P(X = x_4, Y = y_2) = P(Y = y_2 \mid X = X_4)P(X = x_4) \tag{3}$$

$$= \frac{n_{42}}{c_4} \cdot \frac{c_4}{N} = \frac{n_{42}}{N} \tag{4}$$

Bayes' Theorem

Bayes' Theorem (also known as Bayes' Rule or Bayes' Law)

Statement:

$$p(x \mid y) = \frac{p(y \mid x)\, p(x)}{p(y)}$$

Annotated Components:

$$\underbrace{p(x \mid y)}_{\text{Posterior}} = \frac{\overbrace{p(y \mid x)}^{\text{Likelihood}} \times \overbrace{p(x)}^{\text{Prior}}}{\underbrace{p(y)}_{\text{Evidence}}}$$

Derivation from the Product Rule

From the Product Rule:

$$\begin{aligned} p(x, y) &= p(x \mid y)\, p(y) \\ &= p(y \mid x)\, p(x) \end{aligned}$$

Derivation:

$$\begin{aligned} & p(x \mid y)\, p(y) = p(y \mid x)\, p(x) \\ \implies & p(x \mid y) = \frac{p(y \mid x)\, p(x)}{p(y)} \end{aligned}$$

Naive Bayes Classifier

Definition:

- The Naive Bayes classifier is a probabilistic model based on Bayes' theorem.
- It assumes **conditional independence** between features given the class label.

Classification Rule:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \; p(y) \prod_{i=1}^{n} p(x_i \mid y)$$

where:

- $\hat{y}$ is the predicted class label.
- $x = (x_1, x_2, \ldots, x_n)$ is the feature vector.
- $p(y)$ is the prior probability of class y.
- $p(x_i \mid y)$ is the likelihood of feature x_i given class y.

Assumption:

- Features X_i are conditionally independent given the class label Y:

$$p(x \mid y) = \prod_{i=1}^{n} p(x_i \mid y)$$

Example: Naive Bayes Classifier – I

Problem: Classify a new fruit as Apple or Orange based on features **Color** and **Size**.

Training Data:

Fruit	Color	Size
Apple	Red	Small
Apple	Red	Medium
Orange	Orange	Medium
Orange	Orange	Large
Apple	Green	Small

Example: Naive Bayes Classifier – II

Objective: Classify a fruit with **Color** = Red and **Size** = Medium.

Naive Bayes Classification:

- **Calculate Prior Probabilities:**

$$P(\text{Apple}) = \frac{3}{5}, \quad P(\text{Orange}) = \frac{2}{5}$$

- **Calculate Likelihoods:**

$$P(\text{Color} = \text{Red} \mid \text{Apple}) = \frac{2}{3}$$

$$P(\text{Size} = \text{Medium} \mid \text{Apple}) = \frac{1}{3}$$

$$P(\text{Color} = \text{Red} \mid \text{Orange}) = 0$$

$$P(\text{Size} = \text{Medium} \mid \text{Orange}) = \frac{1}{2}$$

Example: Naive Bayes Classifier – III

- **Apply Naive Bayes Formula:**

$$P(\text{Apple} \mid \text{Red}, \text{Medium}) \propto P(\text{Apple}) \times P(\text{Red} \mid \text{Apple}) \times P(\text{Medium} \mid \text{Apple})$$

$$P(\text{Orange} \mid \text{Red}, \text{Medium}) \propto P(\text{Orange}) \times P(\text{Red} \mid \text{Orange}) \times P(\text{Medium} \mid \text{Orange})$$

- **Compute Posterior Probabilities:**

$$P(\text{Apple} \mid \text{Red}, \text{Medium}) \propto \frac{3}{5} \times \frac{2}{3} \times \frac{1}{3} = \frac{2}{15}$$

$$P(\text{Orange} \mid \text{Red}, \text{Medium}) \propto \frac{2}{5} \times 0 \times \frac{1}{2} = 0$$

Conclusion: Since $P(\text{Apple} \mid \text{Red}, \text{Medium}) > P(\text{Orange} \mid \text{Red}, \text{Medium})$, we classify the fruit as **Apple**.

References

- Mathematics for Machine Learning. Deisenroth, Marc Peter and Faisal, A. Aldo and Ong, Cheng Soon, 2020.
  https://mml-book.github.io/book/mml-book.pdf