Elements of Probability Theory

Machine Learning 2019

Michael Wand, Jürgen Schmidhuber, Cesare Alippi

TAs: Robert Csordas, Krsto Prorokovic, Xingdong Zou, Francesco Faccio, Louis Kirsch

based on slides by Jan Unkelbach

Introduction

- So far we have covered neural networks in detail
  - during application (forward phase), we can imagine them as dynamic computers
  - what exactly they compute is learned during training, by gradient descent

Introduction

- So far we have covered neural networks in detail
  - during application (forward phase), we can imagine them as dynamic computers
  - what exactly they compute is learned during training, by gradient descent
- We now will have a look at probability theory
  - a fundament of machine learning and AI
  - important to understand many algorithms
  - important to understand the *outcome* of your experiments (statistical testing!!)

Introduction

- So far we have covered neural networks in detail
  - during application (forward phase), we can imagine them as dynamic computers
  - what exactly they compute is learned during training, by gradient descent
- We now will have a look at probability theory
  - a fundament of machine learning and AI
  - important to understand many algorithms
  - important to understand the *outcome* of your experiments (statistical testing!)

- This is intended as a recap lesson!
  - If you find that you did not understand parts of this lecture, please have a look at a good tutorial
  - Here is a reasonable one, with exercises: http://homepages.inf.ed.ac.uk/sgwater/teaching/general/probability.pdf

Roadmap

In the following two lectures, we want to revisit

- elementary notions of probability
- random variables
- discrete and continuous probability measures
- conditional probabilities and Bayes' theorem

Why Probability Calculus?

Some things are certain:

- a piece of rock falls to the ground if we drop it
- use classical physics for description

but many things are uncertain:

- stock market
- rolling dice

and are subject to a probabilistic description

Why Probability Calculus for ML?

We aim at building artificial systems which make good decisions in an uncertain environment

- build a backgammon (or chess…) computer that makes good moves against an unknown opponent despite not knowing the following moves
- build robots which perform well in difficult environments despite having limited information about their surroundings
- build a handwriting recognition system that gets most of it right despite large variations in people's handwriting
- ***we want to reason in an uncertain world, and we want our machines to be able to do so as well***

Why Probability Calculus for ML?

USI/SUPSI
Istituto Dalle Molle di studi sull'intelligenza artificiale
IDSIA

We train systems where uncertainty is inherent

- some tasks (including training a neural network) do not have an exact analytic solution
  - approximation required
- some methods *require* randomness (neural network initialization)
- train a neural network with a small amount of training *samples*

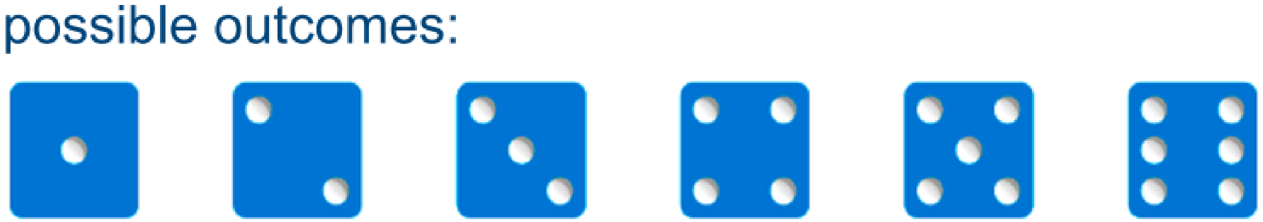
- often: build systems which can estimate how well they are performing!

→ most of AI / machine learning is in some way based on randomness
probabilistic descriptions necessary

The Basics

Random Experiments

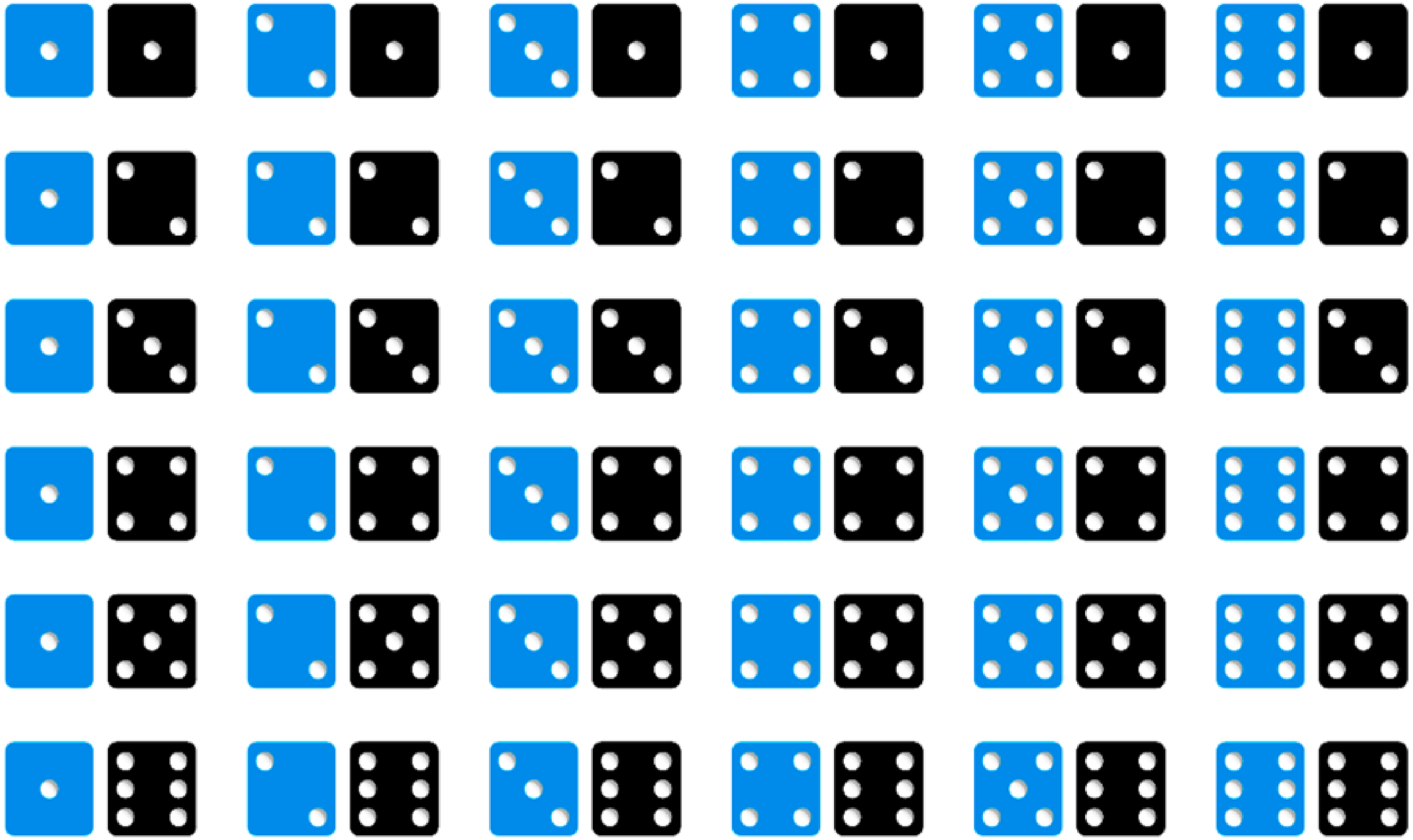
- Consider the prototypical random experiment: let's roll a die!



- The set of all possible outcomes is called the *sample space S*.

Random Experiments

- And if we roll two dice?
- Sample space



Events

- An *Event* is a subset of possible outcomes
- Example events for rolling a dice twice
  - having a sum of 10:
    $A_1 = \{ (4\ 6), (5\ 5), (6\ 4) \}$
  - getting at least one three, and a sum of at least 8:
    $A_2 = \{ (3\ 5), (3\ 6), (5\ 3), (6\ 3) \}$
- The elements of the sample space are called *simple events, e.g.*
  $A_{simple} = \{ (1\ 2) \}$

Events

- The *union* of two events A_1 and A_2 is the event consisting of all events that are either in A_1 or A_2 or both: $A_1 \cup A_2$
- The *intersection* of two events A_1 and A_2 is the event consisting of all events that are in both A_1 or A_2: $A_1 \cap A_2$
- Two events are *mutually exclusive* if they have no outcomes in common, i.e. $A_1 \cap A_2 = \emptyset$
- The complement $\neg A$ of an event A is the set of all outcomes in S that are not in A.

Events

- A *partition* of an event A is a set of events

  $\{A_1, A_2, \ldots, A_n\}$

  with the following properties:
  - all pairs A_i, A_j are mutually exclusive, i.e. $A_i \cap A_j = \emptyset$
  - the union of all A_i is the event A: $A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n = A$

Introduction of Probability

- A probability measure assigns a number to each possible event A, with the following properties:
  - $P(A) \geq 0$
  - $P(S) = 1$
  - for every partition of A,
    $P(A_1) + P(A_2) + \ldots + P(A_n) = P(A)$
- If the sample space is finite (or countable…), one can fully describe the probability measure by giving the probabilities of the simple events.

Probabilities

USI/SUPSI
Istituto Dalle Molle di studi sull'intelligenza artificiale
IDSIA

- Example: rolling a fair die once

  $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$

- Example: The event of getting a result ≥ 5

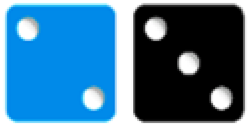
- $P(\{5, 6\}) = P(5) + P(6) = 1/3$

  because the events 5 and 6 are mutually exclusive!

Probabilities

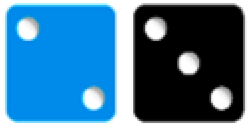
Istituto Dalle Molle di studi sull'intelligenza artificiale
IDSIA
USI/SUPSI

- More examples: We roll two dice again
  - each simple event has probability 1/36
    because there are 36 simple events which are equally likely
  - Event A: First die shows a 5
  - Event B: Second die shows a 3
  - Event C: The sum of both dice is 10



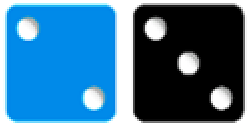
Probabilities

- More examples: We roll two dice again
  - each simple event has probability 1/36
    because there are 36 simple events which are equally likely
  - Event A: First die shows a 5
  - Event B: Second die shows a 3
  - Event C: The sum of both dice is 10



- Clearly, P(A) = P(B) = 1/6 and P(C) = 3/36 = 1/12
  because C = *{ (4 6), (5 5), (6 4) }*, and we can just count

Probabilities



- More examples: We roll two dice again
  - each simple event has probability 1/36
    because there are 36 simple events which are equally likely
  - Event A: First die shows a 5
  - Event B: Second die shows a 3
  - Event C: The sum of both dice is 10
- Clearly, P(A) = P(B) = 1/6 and P(C) = 3/36 = 1/12
  because C = *{ (4 6), (5 5), (6 4) }*, and we can just count
- What is the probability of A ∩ B, i.e. that *both* A and B happen?

Independence

- Event A: First die shows a 5; $P(A) = 1/6$
- Event B: Second die shows a 3; $P(B) = 1/6$
- Event C: The sum of both dice is 10; $P(C) = 1/12$

Clearly, $P(A \cap B) = 1/36$, and we observe that $P(A \cap B) = P(A) \cdot P(B)$

Events with this property are called *independent:* The presence or absence of event A has no influence on event B.

Independence

- Event A: First die shows a 5; $P(A) = 1/6$
- Event B: Second die shows a 3; $P(B) = 1/6$
- Event C: The sum of both dice is 10; $P(C) = 1/12$

What is the probability of $A \cap C$ or $B \cap C$?

Independence

- Event A: First die shows a 5; $P(A) = 1/6$
- Event B: Second die shows a 3; $P(B) = 1/6$
- Event C: The sum of both dice is 10; $P(C) = 1/12$

What is the probability of $A \cap C$ or $B \cap C$?

$P(A \cap C) = 1/36 \neq P(A) \cdot P(C)$

$P(B \cap C) = 0 \neq P(B) \cdot P(C)$

so we see that neither A and C nor B and C are independent

Exclusive events

- Event A: First die shows a 5; $P(A) = 1/6$
- Event B: Second die shows a 3; $P(B) = 1/6$
- Event C: The sum of both dice is 10; $P(C) = 1/12$

And what about the *joint* events $A \cup C$ and $B \cup C$ (i.e. any of the two events happens)?

Exclusive events

- Event A: First die shows a 5; P(A) = 1/6
- Event B: Second die shows a 3; P(B) = 1/6
- Event C: The sum of both dice is 10; P(C) = 1/12

And what about the *joint* events A ∪ C and B ∪ C (i.e. any of the two events happens)?

$$P(A \cup C) = P(\{(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),(6,4),(4,6)\}) = \frac{8}{36} = \frac{2}{9} \neq P(A) + P(C)$$

$$P(B \cup C) = P(\{(1,3),(2,3),(3,3),(4,3),(5,3),(6,3),(6,4),(5,5),(4,6)\}) = \frac{9}{36} = \frac{1}{4} = P(B) + P(C)$$

Remember: Add probabilities only if the events are exclusive!

Conditional probabilities

- The *conditional probability* of B given A is defined as

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

- This is the probability of B if we assume that A is true.
- Exercise: if A and B are independent, show that P(B|A) = P(B).

Conditional Probabilities: Example

Let us look at the two dice again.

- Event A: First die shows a 5; P(A) = 1/6
- Event C: The sum of both dice is 10; P(C) = 1/12
- We had computed: P(A ∩ C) = 1/36

$$P(C|A) = \frac{P(C \cap A)}{P(A)} = \frac{1/36}{1/6} = \frac{1}{6}$$

$$P(A|C) = \frac{P(C \cap A)}{P(C)} = \frac{1/36}{1/12} = \frac{1}{3}$$

Exercise: verify that by counting!

- Note that P(A|C) is *different* from P(C|A)!

Recap: Rules of Computation

Istituto Dalle Molle di studi sull'intelligenza artificiale
USI/SUPSI
IDSIA

These are the major rules you should remember:

- Probabilities are nonnegative and sum to 1
- Assuming two events A and B,
  - $P(A \cap B) = P(A) \cdot P(B)$ if and only if the events are independent
  - $P(A \cup B) = P(A) + P(B)$ if and only if the events are exclusive
- Conditional probability:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \text{ and thus } P(B|A) \cdot P(A) = P(B \cap A)$$

Random Variables

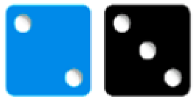
Random Variables

We use the following definition

- A *Random Variable* (RV) X assigns numbers (or vectors) to events

$$X: S \to \mathbb{R} \text{ or } X: S \to \mathbb{R}^N$$

- We thus get an induced probability distribution on the space $\mathbb{R}$ or $\mathbb{R}^N$
- Requires to get some mathematical details right, we'll just skip that

Example:  -> $(2,3) \in \mathbb{R}^2$

Another example: Map the outcome of a throw of two dice to the *sum*

- possible values: 2 … 12, so we lose some information
- outcomes are not equally likely any more

Random Variables

We can describe a random variable by giving its probabilities on the value space.

- Example: sum of two dice
  - p(2) = 1/36, p(3) = 2/36, p(4) = 3/36, etc.
  - We say X has the distribution p: $X \sim p$
  - p is nonnegative, and the sum of all its values is 1.
- Example: Y -> { 0,1 }, Y =1 if the first die shows 5
  - Exercise: describe the distribution of Y

Random Variables

- Two random variables are *independent* if their joint distribution factorizes.
  - Simple example: The sample space S is the space of rolls with two dice, as before
  - X: S -> { 0,1 }, X = 1 if the first die shows “five”.
  - Y: S -> { 0,1 }, Y = 1 if the second die shows “three”.
  - Let $X \sim p_X, Y \sim p_Y, (X, Y) \sim p_{X,Y}$

Random Variables

- Two random variables are *independent* if their joint distribution factorizes.
  - Simple example: The sample space S is the space of rolls with two dice, as before
  - X: S -> { 0,1 }, X = 1 if the first die shows “five”.
  - Y: S -> { 0,1 }, Y = 1 if the second die shows “three”.
  - Let $X \sim p_X, Y \sim p_Y, (X,Y) \sim p_{X,Y}$

  We have

  $p_X(0) = 5/6$, $p_X(1) = 1/6$, $p_Y(0) = 5/6$, $p_Y(1) = 1/6$
  $p_{XY}(0,0) = 25/36$, $p_{XY}(1,0) = 5/36$, $p_{XY}(0,1) = 5/36$, $p_{XY}(1,1) = 1/36$ (verify by counting)

  Since $p_X(a) \cdot p_Y(b) = p_{XY}(a,b)$ for *all* possible pairs a,b, X and Y are independent.

Random Variables

- Two random variables are *independent* if their joint distribution factorizes.
  - Simple example: The sample space S is the space of rolls with two dice, as before
  - X: S -> { 0,1 }, X = 1 if the first die shows “five”.
  - Y: S -> { 0,1 }, Y = 1 if the second die shows “three”.
  - Z: S -> { 2, … 12 } is the sum of two dice.
  - W: S -> { 0,1 }, W = 1 if the sum of the dice is even.
  - Let $X \sim p_X, Y \sim p_Y, Z \sim p_Z, (X,Y) \sim p_{X,Y}$ and so on.

  - Exercise: describe the joint distributions. Which random variables are independent?

Random Variables

- The definition of the conditional probability transfers to random variables, e.g. if we have random variables X and Y, we can define

$$P(X = a|Y = b) = \frac{P(X = a \ \wedge Y = b)}{P(Y = b)}$$

and so on (⋀ means “and”).

Random Variables

We can now define several standard terms:

- The *expectation* of X is the sum of the possible values of X, weighted with their probabilities
  - $E[X] = \sum_x x \cdot P(X = x)$
  - Example: Expected value when we throw one fair die is 3.5
  - You can also compute $E[f(X)] = \sum_x x \cdot P(X = x)$
- The *variance* of X is the expected squared deviance of X and its expectation:
  - $\mathrm{Var}[X] = E[(X - E[X])^2] = \sum_x (x - E[x])^2 \cdot P(X = x) = E[X^2] - (E[X])^2$
- *The* standard deviation is the square root of the variance:
  - $\mathrm{Std}[X] = \sqrt{\mathrm{Var}[X]}$

A Word about Frequentist Statistics

USI/SUPSI
Istituto Dalle Molle di studi sull'intelligenza artificiale
IDSIA

- Think a final time about the dice.
  - I have got a weighted die from the joke shop.
  - How do you estimate the probability that it shows "6"?

A Word about Frequentist Statistics

- Think a final time about the dice.
  - I have got a weighted die from the joke shop.
  - How do you estimate the probability that it shows “6”?
- In practice: Throw it “many” times and count the fraction of “6”.
- E.g. if we got 25 times “6” in 100 throws, we estimate the probability of the die showing 6 to 1/4.
- Same with the expectation: Throw the die many times and average the outcome.

A Word about Frequentist Statistics

- This is a *frequentist* approach which also gives an intuition on what the expected value is:
  - namely the average that we get when repeating the experiment many times
  - …with each repetition being *independent!*
- If we perform such an experiment, the outcome is probabilistic…
  - thus the estimated probabilities and statistics are *themselves* probabilistic
  - opens up the large field of statistical measures (not right now…)
- Finally, note that the frequentist view fails when we have experiments which are not repeatable.

Continuous Random Variables

From discrete to continuous

- So far, we had *discrete* random variables, i.e. they took values on a discrete space (finite or countable)
- We could give the probability of single values, e.g. $P(X_{die}=5)=1/6$
- Random variables can also take values *continuously*, e.g. on the entire $\mathbb{R}$.
  - Useful when the outcomes are naturally continuous, e.g. physical phenomena (signals…)
  - Will be important when we do statistical tests
  - Allows to use integral calculus
- We give probabilities of (reasonable) *subsets* of $\mathbb{R}$.
  - Each single value occurs with probability zero.

A continuous random variable

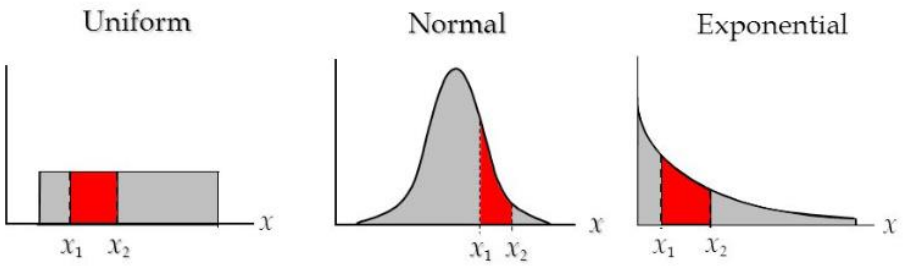
Let's assume X takes all real values. How can we describe X?

- Instead of discrete distribution, use a *density* p

$$p: \mathbb{R} \rightarrow \mathbb{R}^{+}, \int p(x)dx = 1$$

- The probability of x being in the interval x_1 … x_2 is given by

$$P(x \in [x_1, x_2]) = \int_{x_1}^{x_2} p(x)dx$$



- The definition can be generalized to more complex subsets.

A continuous random variable

We define our usual statistical measures as before, just substituting sums with integrals:

$$E[X] = \int x \cdot p(x)dx$$

$$\mathrm{Var}[X] = \int (x - E[x])^2 \cdot p(x)dx = E[X^2] - (E[X])^2$$

$$\mathrm{Std}[X] = \sqrt{\mathrm{Var}[X]}$$

Finally, we define the *cumulative distribution function*

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} p(\xi)d\xi$$

Continuous random variables

Istituto Dalle Molle di studi sull'intelligenza artificiale
USI/SUPSI
IDSIA

This finishes our exposition of basic probability theory. In the next lesson we do Bayes' Theorem and a bit of reasoning with Bayes.

We will occasionally come back to these issues in the future:

- A large class of parametric ML methods estimate parameters of a *distribution* or *density* over the input data.
  - HMMs are probabilistic models
- *Statistical tests* are derived from Bayes' ideas and allow us to quantify how sure we are about our results
- *Information theory* (not covered in this class) yields very fundamental results about our algorithms

Conclusion / Summary

Today you should have revisited

- what is a probabilistic *event,* and what makes events *independent*
- how *probability* is defined, and how to compute elementary probabilities
- what *conditional* probabilities are.

You should also know a bit about

- *random variables*
- their expectation, variance, standard distribution
- distributions and densities.