Foundations of Machine Learning: Week 1: Probability

Professor Christopher Yau
christopher.yau@manchester.ac.uk

October 21, 2020

Foundations of Machine Learning:
Week 1: Probability
Professor Christopher Yau
christopher.yau@manchester.ac.uk
October 21, 2020

What you will learn

Lecture 1
- Terminology of probability
- Rules of probability

Lecture 2
- Conditional probability

Lecture 3
- Total (marginal) probability
- Bayes' Theorem

Example Sheet: Probability

Lecture 4
- Random Variables
- Discrete distributions

Example Sheet: Random Variables, Distributions and Expectations

Lecture 5
- Bayesian Networks
- Directed Acyclic Graphs
- Probabilistic Graphical Models

Lecture 6
- Bayesian Networks

Example Sheet: Probabilistic Graphical Models

What you will learn
Lecture 1
Terminology of probability
Rules of probability
Lecture 2
Conditional probability
Lecture 3
Total (marginal) probability
Bayes' Theorem
Example Sheet: Probability
Lecture 4
Random Variables
Discrete distributions
Example Sheet: Random Variables, Distributions and Expectations
Lecture 5
Bayesian Networks
Directed Acyclic Graphs
Probabilistic Graphical Models
Lecture 6
Bayesian Networks
Example Sheet: Probabilistic Graphical Models

In this first week of the course, we will be examining foundational concepts in probability theory.
We emphasise developing a strong mathematical language for describing probabilistic experiments and problems.
The goal is to ensure that you have the right fundamentals for further study in probability theory.
There will be six short video lectures and three accompanying examples sheets with exemplar problems to consider. There are also online quizzes on Blackboard with short true/false questions to test your understanding of the material.

What is probability?

Both of these are probability questions:

1. What is the probability of a head after a coin flip?
2. What is the probability that you will receive a phone call from an old friend tomorrow?

Answering these questions requires different notions of probability ...

What is probability?
Both of these are probability questions:
1. What is the probability of a head after a coin flip?
2. What is the probability that you will receive a phone call from an old friend tomorrow?
Answering these questions requires different notions of probability ...

In this section of the course, we will be examining probability and its use in machine learning.
Before we do this, let us first consider what "probability" is by looking at these two probabilistic statements.
Both of these are valid probability statements involving *uncertainty*.
We do not know the outcome of the coin flip until it happens.
We do not know if a phone call will occur until tomorrow.
But a response to these two requires different paradigms regarding how we quantify the uncertainty.

Two schools of probability

Frequency-based, objective, defined through observation.

Examples: Coin flipping, rolling a die, bus arrivals, etc.

Broadly, a frequency-based interpretation of probability is best utilised when one can (potentially) repeat an experiment.

Belief-based, subjective, defined through combinations of beliefs and observations (Bayesian).

Examples: Image denoising, audio-to-text translation, etc.

Broadly, a belief-based interpretation of probability is best utilised when problems cannot be framed in terms of being able to repeat experiments.

Two schools of probability

Frequency-based, objective, defined through observation.
Examples: Coin flipping, rolling a die, bus arrivals, etc.
Broadly, a frequency-based interpretation of probability is best utilised when one can (potentially) repeat an experiment.

Belief-based, subjective, defined through combinations of beliefs and observations (Bayesian).
Examples: Image denoising, audio-to-text translation, etc.
Broadly, a belief-based interpretation of probability is best utilised when problems cannot be framed in terms of being able to repeat experiments.

There are two major schools of probabilistic reasoning: the frequentist framework which is the most common within Statistics, and the Bayesian framework which is popular within machine learning.

In a frequentist framework, probability is defined in terms of the *frequency* with which outcomes occur in an experiment. This makes it easy to assign probability to experiments that can actually be done.

When assessing the performance of machine algorithms, you might use frequentist statistics because you might repeatedly run an algorithm on different datasets and record the different results to make a summary performance statement.

In a Bayesian framework, probability is defined in terms of degrees of *belief*. It can be subjective but it means we can assign probabilities to events that would not be so simple to do under a frequentist framework making it amenable to a range of machine learning applications where we want to assign uncertainty to different quantities such as the value of a neural network parameter, or a coefficient in a regression model.

Why probability and machine learning?

There are two learning paradigms in machine learning:

1. **Deterministic** approaches where, given the same conditions and inputs, a learning algorithm *always* returns the same output,
2. **Probabilistic** approaches where, given the same conditions and inputs, a learning algorithm returns a *distribution* of outputs.

Probability is critical to probabilistic ML algorithms as it is the mathematical language by which we capture randomness and uncertainties associated with learning problems and build into our algorithms.

Why probability and machine learning?

There are two learning paradigms in machine learning:

1. **Deterministic** approaches where, given the same conditions and inputs, a learning algorithm *always* returns the same output,
2. **Probabilistic** approaches where, given the same conditions and inputs, a learning algorithm returns a *distribution* of outputs.

Probability is critical to probabilistic ML algorithms as it is the mathematical language by which we capture randomness and uncertainties associated with learning problems and build into our algorithms.

Both of these frameworks have utility within machine learning for different purposes but why is probability important?
Broadly, machine learning methods can be divided into two camps.
Deterministic approaches in which algorithms are developed that return exactly one output, for each input and that output is always the same for the same input.
Probabilistic approaches return a *probability distribution* over outputs for each input. Multiple outputs are possible each weighted by a probability.
Deterministic approaches are often simpler to interpret as they give you a straightforward answer for each query. In contrast, probabilistic approaches are more complex to interpret but are more flexible in that one can understand the possible alternatives.

Probabilistic Machine Learning

Given input X, an output Y, a learning function F is:

If f is **deterministic:**

$$X \Rightarrow_F Y$$

A single input X leads to a specific output Y.

If f is a **stochastic process:**

$$X \Rightarrow_F p(Y)$$

A single input X leads to *distribution* of outputs $p(Y)$.

Probabilistic Machine Learning

Given input X, an output Y, a learning function F is:
If f is **deterministic:**
$X \Rightarrow_F Y$
A single input X leads to a specific output Y.
If f is a **stochastic process:**
$X \Rightarrow_F p(Y)$
A single input X leads to *distribution* of outputs $p(Y)$.

Mathematically, we can say that there exists a learning function F which transforms an input X into an output Y.
A deterministic algorithm uses a straightforward deterministic function F whereas a probabilistic algorithm assumes that F represents a *stochastic process* such that each time you give the algorithm the same input X, the output Y may change each time.
However, over many iterations, the frequency of different outputs converges to a certain probability distribution $p(Y)$.

Probability and performance

Probability is also important for understanding machine learning performance.

For example, error analysis for a binary classification algorithm:

		Ground Truth	
		False	True
Machine	False	-	FN
	True	FP	-

What is the probability that the classifier makes a **false positive** (FP) or a **false negative** (FN)?

What if the machine gives us a probabilistic output?

Probability and performance

Probability is also important for understanding machine learning performance.

For example, error analysis for a binary classification algorithm:

What is the probability that the classifier makes a **false positive** (FP) or a **false negative** (FN)?

What if the machine gives us a probabilistic output?

Probability is also important for understanding machine learning performance. Consider, for example, error analysis for a binary classification algorithm so it gives an output FALSE or TRUE. And this classification can either be correct as measured against a ground truth or incorrect.

If the machine says TRUE, but the truth is FALSE, then the error is known as a FALSE POSITIVE.

If the machine says FALSE, but the truth is TRUE, then the error is known as a FALSE NEGATIVE.

The machine will have a certain probability of making each type of error and those probabilities are likely to be different and adjustable depending on the parameters of the model.

What is acceptable for each type of error will depend on the application.

Probability becomes even more important if the classifier is probabilistic and does not give a definite TRUE or FALSE output but a probability associated with these.

Coming next ...

This course will examine the foundational concepts of probability to get you started on your study of these topics.

Coming next ...

This course will examine the foundational concepts of probability to get you started on your study of these topics.

Next we will look at the foundational concepts of probability to get you started on further study of these topics.

For those new to probability, there will be a lot of material, so please use not just this week, but time between now and your examination to reflect on the material and practice with the example problems.