Statistical Information Theory course information

Pedro A.M. Mediano

February 2023

Welcome to Statistical Information Theory! This document contains useful information for students considering the course, and will act as a permanent reference during the course itself.

1 Module aims

What does fitting a machine learning model have to do with copying a file over a computer network? The answer, you may be surprised to hear, is quite a lot. The fundamental principles of information theory were originally devised to study optimal communication over information channels, but have since found widespread applications across areas of data science and computing. This course will introduce the basics of information theory as devised by Claude Shannon in 1948, and then delve into its deep connections with statistics and machine learning. Through mathematical and computational exercises, the course presents information as a unifying theory linking computing, statistics, and geometry, and provides crucial theoretical background for students wishing to pursue a career in data science or machine learning.

1.1 Who is this course for?

The course is recommended for students wishing to build a strong background in foundational machine learning, especially for those intending to pursue further research in machine learning or statistics. It is also recommended for students interested in diversifying their knowledge of statistics, and provides a theoretical foundation for data science applications. The course works especially well in conjunction with Mathematics for Machine Learning.

Conversely, this course is not recommended for students that are not interested in statistics, data science, or machine learning.

There are no strict requirements, but background knowledge in probability and statistics is strongly recommended. Chapters 3, 6, and 7 of Deisenroth's *Mathematics for Machine Learning* are also strongly recommended as prior reading (see full reading list below).

2 Module logistics

The material will be taught through lectures and consolidated in both computer- and paper-based tutorial exercises, designed to reinforce your understanding of the material. In-person teaching (both lectures and tutorials) will take place twice a week:

- **Mondays, 2-4pm**, in room 414 in the Roderic Hill building.
- **Thursdays, 2-4pm**, in room 144 in the Huxley building.

Online discussions about module content will be hosted on EdStem. Please use EdStem instead of email for any questions about the module.

There will be one mid-term assessment in the form of an in-class paper-based assessed coursework, worth 20% of the module marks. The remaining 80% of the marks will be assessed in an exam in the final week of term.

3 Module content

3.1 Learning outcomes

Upon successful completion of this module you will be able to:

- Explain the key properties of information metrics in terms of communication principles.
- Calculate these metrics in common probability distributions.
- Compare different metrics and coding schemes on real-world data.
- Explain the mathematical connections between data transmission and model fitting.
- Design principled data analysis plans with appropriate statistical criteria.

3.2 Syllabus

The course is divided in three parts, with approximately equal weight.

Part I: Data compression and source coding

The course will begin with a revision of the basic properties of probability distributions, and focus on the fundamental quantity of information theory: entropy. We will cover:

- Entropy and information content
- Statistical interpretation of entropy
- Source coding theorem and Huffman codes

Part II: Data transmission and channel coding

This part considers extensions of entropy to two or more variables, and studies in depth another key quantity in information theory: mutual information. We will cover:

- Mutual information and its properties
- Channel coding theorem and Hamming codes

Part III: Advanced topics in statistics

This final part of the course builds on top of the other two to describe deep connections with statistics, probabilistic inference, and other branches of mathematics. We will cover:

- Model comparison and information geometry
- Maximum entropy models
- Origin of the bias-variance trade-off

4 Reading list

The two main textbooks the course will follow are:

- MacKay, D. (2003). *Information Theory, Inference and Learning Algorithms.* Cambridge University Press. This book is kindly provided free of charge by Cambridge University Press:

  https://www.inference.org.uk/mackay/itila/book.html

- Cover, T., & Thomas, J. (2006). *Elements of Information Theory.* John Wiley & Sons. This book is available from Imperial's library.

Other books that are recommended, but not required, are:

- Bishop, C. (2006). *Pattern Recognition and Machine Learning.* Springer.
- Deisenroth, M. *et al.* (2021). *Mathematics for Machine Learning.* Cambridge University Press.
- Boyd, S., & Vandenberghe, L. (2004). *Convex Optimization.* Cambridge University Press.