Information Theory

Final Overview

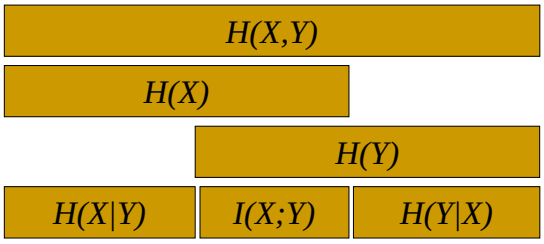
The Biggest Points of the Course

- "Information" is sent and/or stored
- This involves noise and/or compression
- One must infer the original information from the noisy/compressed information
- There are theoretical limits to communication rate/compression ratio
- There are procedures for coding/compressing/inferring to try and approach these limits

Math Basics

- Forward probability problems
- Backward probability problems
  - Bayes rule
- Eigenvalues and eigenvectors (concept)

Entropy basics



Shannon's Theorems

- Shannon's Theorems show that
  - There are arbitrarily good communication schemes and compression schemes, bounded by H(X) for the noise involved
  - Very good, linear communication/lossy compression schemes exist (e.g., block codes)
  - Huffman coding is a practical symbol coding implementation of lossless compression
    - Be able to do the procedure
  - For better performance, codes that consider context (Arithmetic, Limpel-Ziv) achieve better efficiency
    - Be able to explain the concepts

Step one of Bayesian Inference

- **Model fitting**: for a number of parameterized hypotheses, we determine the best parameters for the given data:

$$P(\mathbf{w} \mid D, M) = \frac{P(D \mid \mathbf{w}, M) P(w \mid M)}{P(D \mid M)}$$

$$\text{posterior for parameters of model} = \frac{(\text{likelihood of parameters of model})(\text{prior for parameters of model})}{(\text{likelihood of model})}$$

- MacKay calls the likelihood of the model the *evidence*

Step two of Bayesian Inference

- **Model Comparison:** for all the models to which you have fit the data, select the best one

$$P(M \mid D) = \frac{P(D \mid M)P(M)}{P(D)}$$

- *The key quantity here is the "evidence" P(D|M)*

The Capacity of a Channel Q

- Is defined as the maximum information we can convey about *x* by reading *y*
- We can accomplish this by picking the coding

$$C(Q) = \max_{P_x, P_y} I(X;Y)$$

Clustering

- Soft and Hard K-means
  - Know the procedures
- Soft K-means as maximum likelihood reasoning
  - Know the concept

Independent Component Analysis

- ICA relies on the assumption of
  - Statistically Independent underlying signals
  - That are non-Gaussian
  - zero mean and fixed variance
- The algorithm involves
  - minimizing mutual information between signals
  - which leads to maximizing non-gaussinaity
  - which leads to minimizing negentropy
  - which is approximated
  - which results in a NN-like update algorithm

Gaussian Channel

- Real channels send information as the amplitudes of orthonormal basis functions
- This transmission is limited by power and bandwidth
- Looking at the discrete time Gaussian Channel, the power limited channel capacity is defined in terms of SNR
- We can relate this back to the continuous time channel

Modelling Probability Distributions (and sampling)

- The principle of maximum entropy can be utilized to determine appropriate distributions given testable facts
- Monte Carlo sampling methods are often necessary to generate samples and evaluate functions of random variables
- Some of the best of these are Markov Chain Monte Carlo methods
- These include the Metropolis Method
- From which the Hamiltonian Method and Simulated Annealing can be derived