Unsupervised Learning and Minimum Entropy Codes

Patrik Hoyer

Sept. 28th, 2000

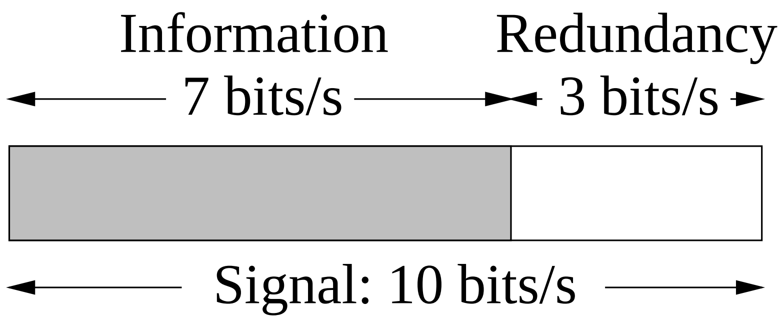
Based on: 'Unsupervised Learning' (Barlow 1989) and
'Finding Minimum Entropy Codes' (Barlow et al. 1989).

What is unsupervised learning?

- Unsupervised learning is:
  - Finding the structure in a set of unlabeled data.
  - Extracting useful information from a large dataset.
  - Making sense of a large body of signals.
  - "Data mining"

Information vs. redundancy

- Signal can be divided into information and redundancy:



- Goal of unsupervised learning:
  - Long-term: Find the structure of the signal (identify redundancy)
  - Short-term: Use the structure to extract information from the signal

A simple example of redundancy

- Signal:

  $\{1, 2, 3, 4, 1, 2, 3, 4, 5, 6, 7, 1, 2, 1, 2, 3, 4, 5, 6\}$

- Structure: Each number x is followed by 1 or $(x + 1)$

- Information:

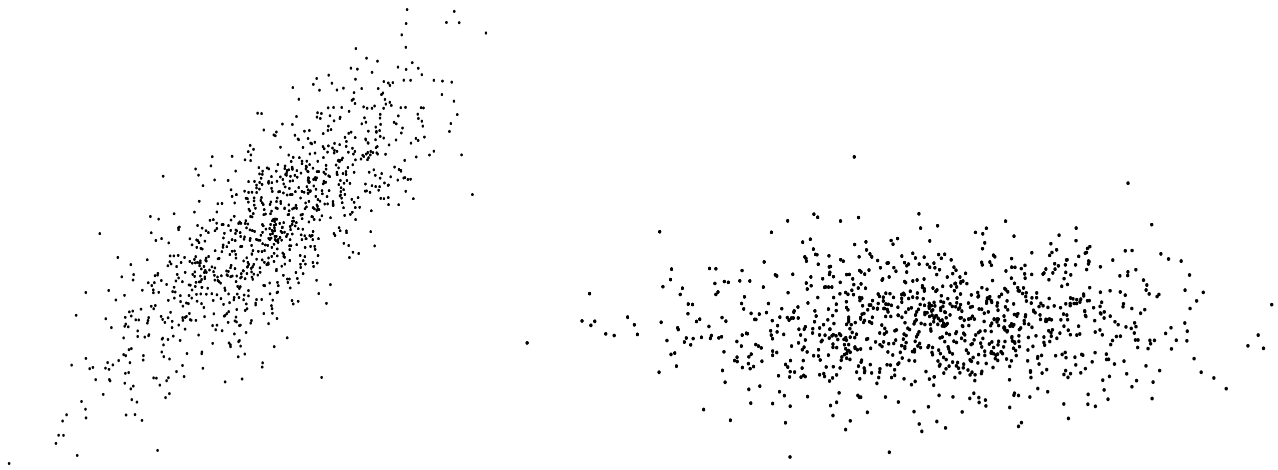
  $\{1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0\}$

- Additional structure

  – Interval length distribution?
  – Dependencies between interval lengths?
  – Etc.

Example #2: 2D Gaussian

- Input: samples from a 2D Gaussian
- Structure to be found: covariance matrix
- Extracting information: decorrelate signals



Minimum Entropy Coding

- Redundancy can be present in individual inputs and/or in dependencies between inputs.
- Goal: reversibly transform the inputs so that the outputs are as independent as possible (=ICA).
- In these papers, the discussion is restricted to binary representations of discrete symbols.
- How? (Discussed later in the talk.)
- Why? (Discussed next.)

Why independence?

- Component redundancy is easier to handle than dependencies between components!
  - Signal processing: compression, denoising etc. can be done *componentwise*!
  - Biological networks: sparse activations conserves energy.
- Independence useful when seeking to learn new associations (next slide).
- Gives an explicit representation of the data in terms of 'causes'.

Detecting new associations

- Consider classical conditioning: conditioned stimulus C, unconditioned stimulus U.
- To learn the association between U and C, we must estimate $P(U, C)$ and compare it to $P(U, C|\text{independent}) = P(U)P(C)$.
- Calculating this latter value is in general difficult, unless we have U and C represented independently!

Other factors affecting the utility of representations

- Generalization performance (how does the learned representation extend to unseen data?)
- Biological significance (markings of prey, predators, mates etc.).
- Reinforced stimuli.

Finding minimum entropy codes: Framework

- A finite number of states are to be coded with binary codes.
- Each state j occurs with probability A_j.
- We can assign any binary code to any state.
- Try to find the code where bits are independent of each other.

Symbol entropy vs. sum of bit entropies

- The symbol entropy is $E(A) = -\sum A_j \log A_j$.
- The sum of bit entropies using a code $\mathbf{b}$ is $e(A, \mathbf{b}) = -\sum p_i \log p_i - \sum q_i \log q_i$, where $p_i = 1 - q_i$ is the probability of a certain bit being 1 given the probability distribution A.
- Always $E(A) \leq e(A, \mathbf{b})$, with equality only when bits are independent.
- Thus: We can search for independence by minimizing the sum of bit entropies!

Finding the code #1

- Barlow et al. (1989) found no general method short of searching through all possible codes (impractical).
- One approach (*Minimum bit probability*): minimize the expected number of active units $= \sum p_i$.
- This can be done by ranking the states according to probability and assigning all '0' to most probable, states with only one '1' to the next probable ones, and so on...
- However, this is not the same as minimizing sum of bit entropies!

Finding the code #2

- A different approach (*Binary Sequence Coding*): States are ranked and given codes in binary sequence.
- Gives a factorial code when state probabilities form a geometric progression: $A_j = Kx^j$.
- This happens to be nearly the case for ASCII data!

Results for text

Code	$e(A, \mathbf{b})$	Red. (%)
Minimum bit probability	4.50	3.7
Binary sequence	4.37	0.7
ASCII	5.42	24.9
Random-1	6.89	58.8
Random-2	6.77	56.0
Random-3	6.74	55.3

(Character entropy $E(A) = 4.34$ bits.)

Subsection coding

- When the number of states (equiv. number of inputs) increases, these methods become impractical as they require estimating frequencies of all possible states!
- This can be alleviated by only considering a subset of the inputs at any one time, and recoding them individually.

Conclusions

- Barlow proposes factorial coding (ICA) as a suitable goal for unsupervised learning.
- In the second paper, some methods are presented for approximate *binary* factorial coding of *discrete* data.
- Although these algorithms have not been as widely applied as recent ICA algorithms, the motivation of ICA for unsupervised learning presented here is as relevant as ever.