CSC310 – Information Theory — Sam Roweis

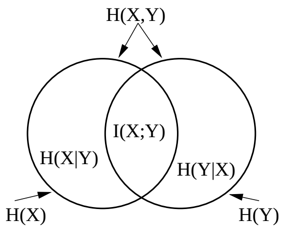
LECTURE 24:

SUMMARY/REVIEW

December 4, 2006

ENTROPY & MUTUAL INFORMATION

- The *mutual information* between the input and the output:
  $I(X;Y) \;=\; H(X) + H(Y) - H(X,Y)$
- Mutual information is meant to represent the amount of information that is being communicated from sender to receiver.



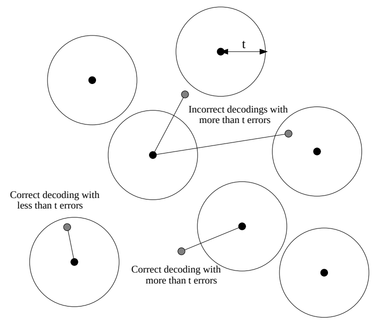
- Joint and Conditional Entropies:
  $$H(X,Y)=\sum_{i=1}^{r}\sum_{j=1}^{s} p(x=a_i, y=b_j)\log(1/p(x=a_i, y=b_j))$$
- Cross-Entropy (Kullback-Leibler Divergence)
  $$KL[p\|q] = \sum_i p_i \log p_i/q_i; \quad I(X;\,Y) = KL[p(x,y)\|p(x)p(y)]$$

CHANNELS & CAPACITY

- Channels: Input/Output alphabets; Transition Probabilities
- Memoryless (Independent), Synchronized
- Examples: BSC, BEC, Z
- Capacity == maximal average mutual information between input symbol and output symbol that can be obtained with any choice of input distribution. (Achieved when the channel is driven by its "resonant input distribution".)
- The capacity is the rate at which data can be sent through the channel with vanishingly small probability of error.

CODES, MESSAGES & RATES

- N identical useages of a channel form the N^{th} extension of the channel.
- A *code* is a subset of all possible length N sequences of input symbols. The elements of the subset are the *codewords*.
- If we use 2^K codewords, we can create a mapping between K-bit *messages* and the N-bit codewords.
- The *rate* of a code $R = K/N$ is the fractional effective speed at which message bits get transmitted across the channel.
- Minimum distance, maximum error tolerance.



- Encoding: map from K bits to one of 2^K codewords.
- Decoding: MAP, ML, minimum-distance
- When the sender transmits a codeword in $\mathcal{C}$, the receiver might (in general) see any output block, $b_{j_1} \cdots b_{j_N} \in \mathcal{A}_Y^N$.

  The receiver can try to *decode* this output in order to recover the codeword that was sent.

  The optimal method of decoding is to choose a codeword, $w \in \mathcal{C}$, which maximizes

$$P(w \mid b_{j_1} \cdots b_{j_N}) = \frac{P(w)\, P(b_{j_1} \cdots b_{j_N} \mid w)}{P(b_{j_1} \cdots b_{j_N})}$$

  In case of a tie, we pick one of the best w arbitrarily.

  If $P(w)$ is the same for all $w \in \mathcal{C}$, this scheme is equivalent to choosing w to maximize the "likelihood", $P(b_{j_1} \cdots b_{j_N} \mid w)$.

- Shannon's noisy coding theorem states that:

  For any channel with capacity C, any desired error probability, $\epsilon > 0$, and any transmission rate, $R < C$, there exists a code with some length N having rate at least R such that the probability of error when decoding this code by maximum likelihood is less than ϵ.
- In other words: We can transmit at a rate arbitrarily close to the channel capacity with arbitrarily small probability of error.
- The converse is also true: We *cannot* transmit with arbitrarily small error probability at a rate greater than the channel capacity.
- We can always chose to transmit beyond the capacity, but not with vanishly small error – our best possible error rate will still be finite.
- Proof by random choice of codes.

- Problems with large block lengths.
- Using aritmetic mod 2, we could represent a code using either a set of basis functions or a set of constraint (check) equations, represented a generator matrix G or a parity check matrix H.
- Every linear combination basis vectors is a valid codeword & all valid codewords are spanned by the basis; similarly all valid codewords satisfy every check equation & any bitstring which satisfies all equations is a valid codeword.
- The rows of the generator matrix form a basis for the subspace of valid codes; we could encode a source message $\mathbf{s}$ into its transmission $\mathbf{t}$ by simple matrix multiplication: $\mathbf{t} = \mathbf{s}G$.
- The rows of the parity check matrix H form a basis for the complement of the code subspace and represent check equations that must be satisfied by every valid codeword.

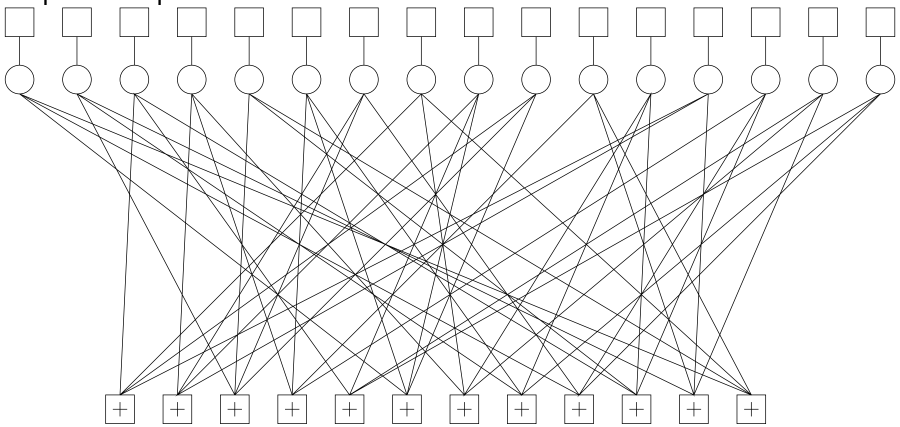
- For linear codes, minimum distance = minimum codeword weight.
- The maximum number of errors we can guarantee to correct is the half the minimum distance (minus one).
- Perfect Packing, Sphere Packing Bound, Gilbert Varshamov Bound
- For each positive integer c, there is a binary Hamming code of length $N = 2^c - 1$ and dimension $K = N - c$. These codes all have minimum distance 3, and hence can correct any single error.
- They are also perfect, since

$$2^N/(1+N) \;=\; 2^{2^c-1}/(1+2^c-1) \;=\; 2^{2^c-1-c} \;=\; 2^K$$

  which is the number of codewords.
- Hamming Codes have a very simple decoding procedure
- Syndrome Decoding for general linear codes

Low Density Parity Check Codes 8

- Sparse Graphs



- Message Passing Decoding for BEC – hard decisions
- Message Passing Decoding for BSC – probabilities

Product Codes, Reed-Solomon 9

- A *product code* is formed from two other codes $\mathcal{C}_1$, of length N_1, and $\mathcal{C}_2$, of length N_2. The product code has length $N_1 N_2$.
- We can visualize the $N_1 N_2$ symbols of the product code as a 2D array with N_1 columns and N_2 rows.
- Definition of a product code:
  An array is a codeword of the product code if and only if
  - all its rows are codewords of $\mathcal{C}_1$
  - all its columns are codewords of $\mathcal{C}_2$
- Product Codes can correct bursts of errors

0	1	?	0	0	1	0	0	0	?	?	?	?	?	?	0	0	0	0	1	0	0	0	1	?	0	0

- Erasure Channels - packets get lost, not corrupted
- Reed-Solomon codes are based on polynomial interpolation

LT & Digital Fountain Codes 10

- The idea of LT codes is that the sender is a fountain that produces an endless supply of encoded packets. (Hence these codes and their variants are often called digital fountain codes.)
- Say the original source file has a size of Kl bits, and each packet contains l encoded bits. Anyone who wishes to receive the complete file holds a bucket under the fountain and collects packets until they have collected a little more than K (in practice, this is usually around 5%). They can then almost certainly (with prob. $1-\delta$) recover the original file exactly.
- LT codes are *rateless* in the sense that the number of encoded packets that can be generated from the source message is potentially limitless and the number of encoded packets generated can be determined on the fly.
- LT codes also have fantastically small encoding and decoding complexities.

Lossy Compression & Rate Distortion 11

- Many kinds of data — such as images and audio signals — contain "noise" and other information that is not really of interest. Preserving such useless information seems wasteful.
- **A common approach:** *Lossy compression*, for which decompressing a compressed file gives you something *close* to the original, but not necessarily exactly the original.
- We should be able to compress to a smaller file size if we don't have to reproduce the original exactly.
- JPEG Lossy Image Compression.