Bayesian statistics and Information Theory

Lecture 3: Information theory

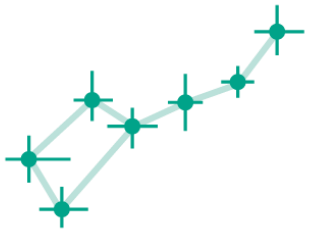
... a.k.a. *how much is there to be learned in my data anyway?*

Florent Leclercq

www.florent-leclercq.eu

Imperial Centre for Inference and Cosmology

Imperial College London



May 28th, 2019

ICIC Imperial Centre for Inference & Cosmology

Imperial College London

Outline: Lecture 3

- Bayesian model comparison
  - Nested models and the Savage-Dickey density ratio
  - Bayesian model selection as a decision analysis
  - Bayesian model averaging
  - (Dangers of) model selection with insufficient summary statistics

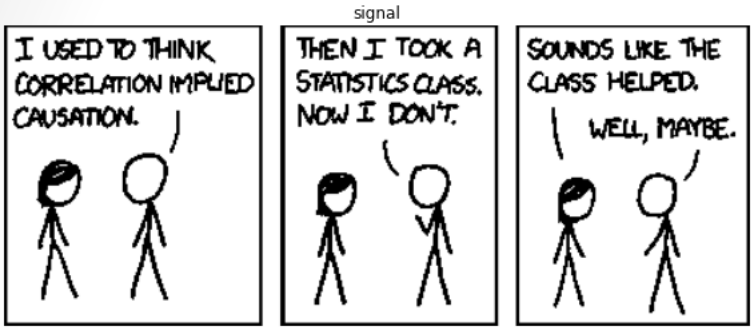
- Information theory
  - The noisy binary symmetric channel
  - Low-density parity check codes
  - Measures of entropy and information
  - Information-theoretic experimental design
  - Supervised machine learning basics

Bayesian model comparison

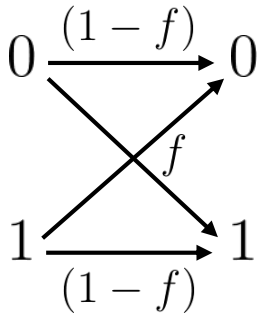
The noisy binary symmetric channel

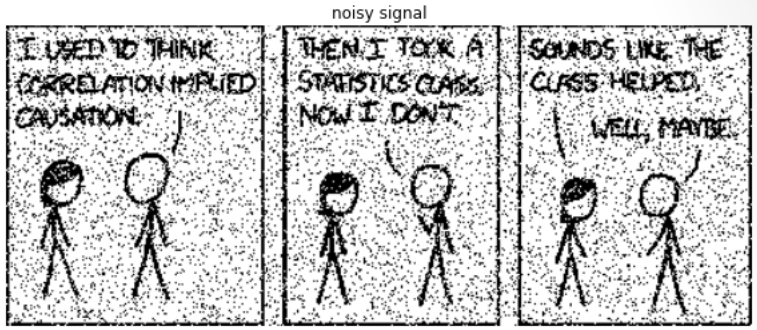
Notebook 13: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/IT_noisy_binary_channel.ipynb

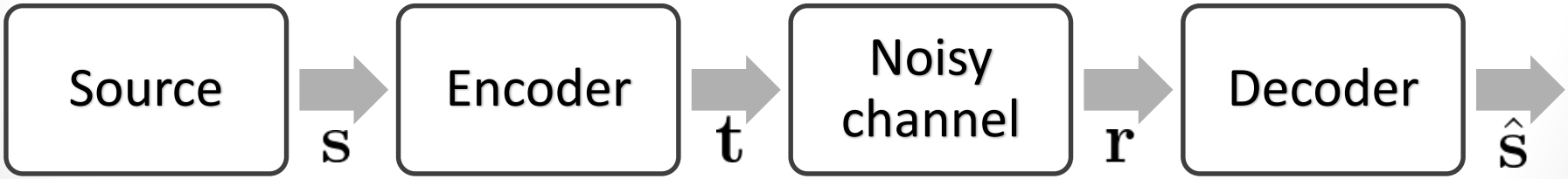
The noisy binary symmetric channel



https://xkcd.com/552/







Rate of information transfer: $R = \frac{\#\mathbf{s}}{\#\mathbf{t}} = \frac{K}{N}$

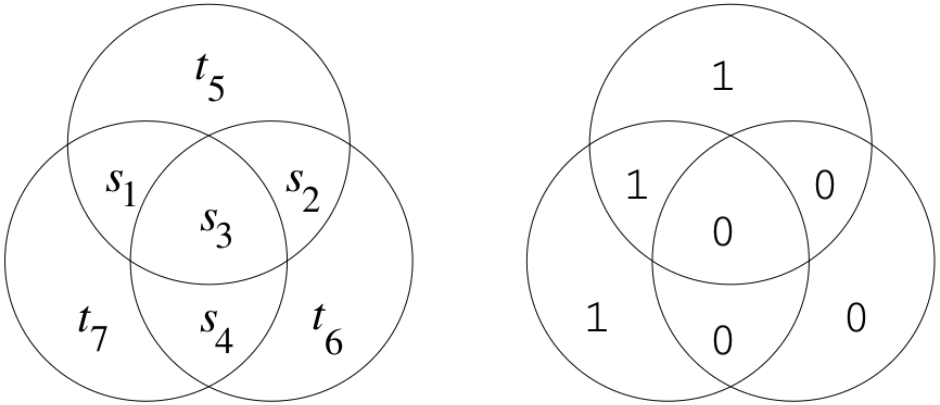
The R3 code

s	0	0	1	0	1	1	0
t	0 0 0	0 0 0	1 1 1	0 0 0	1 1 1	1 1 1	0 0 0
n	0 0 0	0 0 1	0 0 0	0 0 0	1 0 1	0 0 0	0 0 0
r	0 0 0	0 0 1	1 1 1	0 0 0	0 1 0	1 1 1	0 0 0
$\hat{\mathbf{s}}$	0	0	1	0	0	1	0
corrected errors		$\star$					
undetected errors					$\star$		

MacKay 2003, chap. 1

The (7,4) Hamming code: Encoder

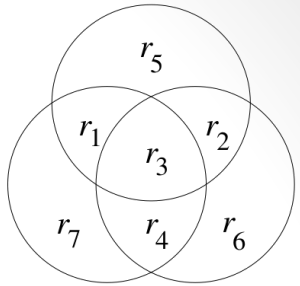
- Introducing the concept of **parity-checks**:



$$\mathbf{t} = \mathbf{G}\mathbf{s} \qquad \mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

MacKay 2003, chap. 1

The (7,4) Hamming code: Decoder



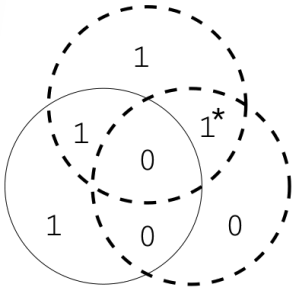
- Introducing the concept of **syndrome**:
- Error correction:

syndrome

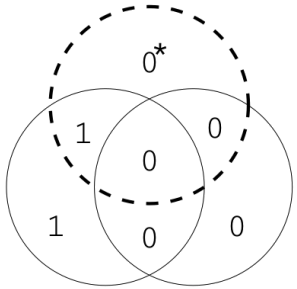
$$\mathbf{z} = \mathbf{Hr}$$

$$\mathbf{H} = \left[\begin{array}{cccc|ccc} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right]$$

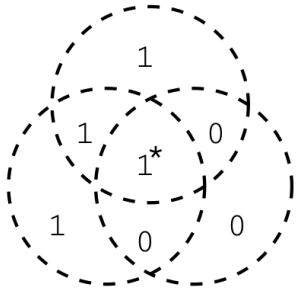
parity-checks identity



$\mathbf{z} = 110$



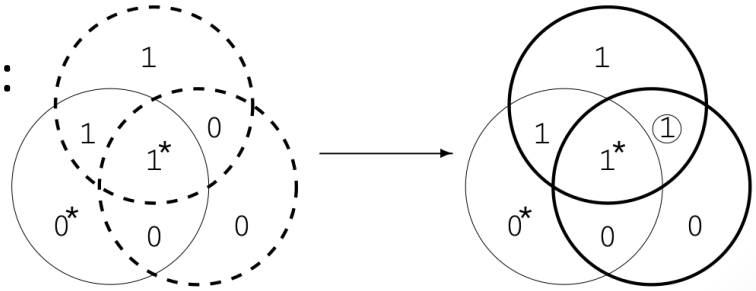
$\mathbf{z} = 100$



$\mathbf{z} = 111$

Syndrome $\mathbf{z}$	000	001	010	011	100	101	110	111
Unflip this bit	None	r_7	r_6	r_4	r_5	r_1	r_2	r_3

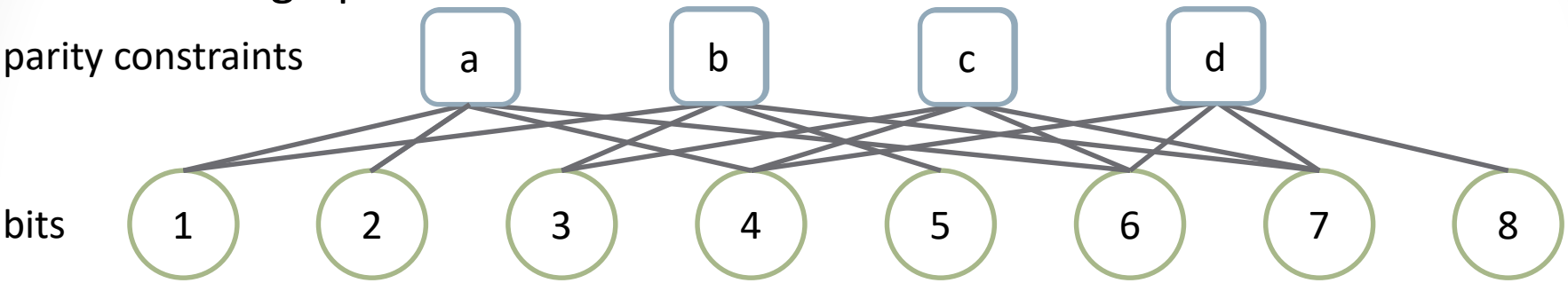
- Unsuccessful error correction:



MacKay 2003, chap. 1

Low-density parity check (LDPC) codes

- Tanner graph:



- *N* bits, *M=(1-R)N* parity-check constraints
  - 2^{RN} possible "words": the ***dictionary***
  - A ***sparse*** parity-check matrix
- Decoding LDPC codes: the general theory borrows from **statistical physics**: Ising spins in interaction and the BP mean field approximation (Bethe-Peierls – Belief Propagation).

MacKay 2003, chap. 47

Measures of entropy and information

Measures of entropy and information

- **Information content**: $I[X] \equiv -\sum_{x \in \mathcal{X}} \log_2 p(x)$
- **Entropy**:

$$H[X] \equiv \langle I[X] \rangle_{p(X)} = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

- **Conditional entropy**:

$$H[X|Y] \equiv \langle H[X|Y=y] \rangle_{p(Y)} = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log_2 \left(\frac{p(y)}{p(x,y)} \right)$$

- Properties:
  - chain rule: $H[X|Y] = H[X,Y] - H[Y]$
  - "Bayes's theorem": $H[X|Y] + H[Y] = H[Y|X] + H[X]$

Reviews: Crooks 2016; Leclercq *et al.* 2016, arXiv:1606.06758 (Appendix A)

Measures of entropy and information

- **Mutual information**: $I[X:Y] \equiv \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right)$
- Properties:
  - $I[X:Y] \geq 0$
  - $I[X:X] = H[X]$
  - $I[X:Y] = H[X] - H[X|Y]$
    $= H[Y] - H[Y|X]$
    $= H[X] + H[Y] - H[X,Y]$
    $= H[X,Y] - H[X|Y] - H[Y|X]$
  - Consequence: $H[X|Y] \leq H[X]$
- **Relative entropy/Kullback-Leibler divergence/Information gain**: $D_{\mathrm{KL}}[p||q] \equiv \sum_{x\in\mathcal{X}} p(x)\log_2\left(\frac{p(x)}{q(x)}\right)$
- Properties:
  - Gibbs's inequality: $D_{\mathrm{KL}}[p||q] \geq 0$
  - Relation to mutual information:

$$I[X:Y] = D_{\mathrm{KL}}[p(x,y)||p(x)p(y)] = \langle D_{\mathrm{KL}}[p(x|y)||p(x)]\rangle_{p(Y)}$$

Reviews: Crooks 2016; Leclercq *et al.* 2016, arXiv:1606.06758 (Appendix A)

Measures of entropy and information

- **Symmetrisation procedures**:
  - Jeffreys symmetrisation: $C_s[A:B] = \frac{1}{2}C_a[A||B] + \frac{1}{2}C_a[B||A]$
  - Jensen symmetrisation:

$$C_s[A:B] = \frac{1}{2}C_a[A||M] + \frac{1}{2}C_a[B||M] \quad \text{with} \quad M = \frac{A+B}{2}$$

- **Jeffreys divergence**: $D_{\mathrm{J}}[p:q] = \frac{1}{2}D_{\mathrm{KL}}[p||q] + \frac{1}{2}D_{\mathrm{KL}}[q||p]$
- **Jensen-Shannon divergence**:

$$D_{\mathrm{JS}}[p:q] = \frac{1}{2}D_{\mathrm{KL}}[p||r] + \frac{1}{2}D_{\mathrm{KL}}[q||r] \quad \text{with} \quad r \equiv \frac{p+q}{2}$$

- Properties:
  - $D_{\mathrm{JS}}[p:q] = H[r] - \frac{1}{2}H[p] - \frac{1}{2}H[q]$
  - $0 \leq D_{\mathrm{JS}}[p:q] \leq 1$
  - $D_{\mathrm{JS}}[p:q] = I[m:z] \quad \text{with} \quad m \equiv z\,p + (1-z)\,q \qquad z = \begin{cases} 0 & (p = 1/2) \\ 1 & (p = 1/2) \end{cases}$

Reviews: Crooks 2016; Leclercq *et al.* 2016, arXiv:1606.06758 (Appendix A)

Information-theoretic experimental design

Information-theoretic experimental design

(expected) data

experimental design

- **General problem**: maximize $U(\xi) = \langle U(d,\xi)\rangle_{p(d|\xi)} = \int p(d|\xi)U(d,\xi)\,\mathrm{d}d$

- **1- Parameter inference utility functions**:
  - Maximal information gain:

$$U(d,\xi) \equiv D_{\mathrm{KL}}[p(\theta|d,\xi)||p(\theta|\xi)] \qquad U(\xi) = I[\theta : d|\xi]$$

  - A-optimality:

$$U_{\mathrm{A}}(d,\xi) \equiv \frac{1}{\mathrm{tr}(\mathrm{cov}(\theta|d,\xi)^{-1})}$$

  - D-optimality:

$$U_{\mathrm{D}}(d,\xi) \equiv \det(\mathrm{cov}(\theta|d,\xi))$$

Review: Leclercq *et al*. 2016, arXiv:1606.06758 (Appendix B)

Information-theoretic experimental design

- **2- Model selection utility functions**:
  - Maximal Bayes factor:

$$U(\xi) \equiv \mathcal{B}_{12}(\xi) = \frac{p(d|\xi, \mathcal{M}_1)}{p(d|\xi, \mathcal{M}_2)}$$

It is hard to predict Bayes factors… but see Trotta 2007, arXiv:astro-ph/0703063

  - Maximal mutual information between the model indicator and the data:

$$U(\xi) \equiv I[\mathcal{M} : d|\xi]$$

e.g. Cavagnaro *et al.* 2010

  - Maximal Jensen-Shannon divergence between posterior predictive distributions:

$$U(\xi) \equiv D_{\mathrm{JS}}[p_1 : p_2] = I[\mathcal{M} : r|\xi] \quad \text{with} \quad r \equiv \frac{p_1 + p_2}{2}$$

Vanlier *et al.* 2014, FL, Lavaux, Jasche & Wandelt 2016, arXiv:1606.06758

- **3- Utilities for prediction of future observations**:

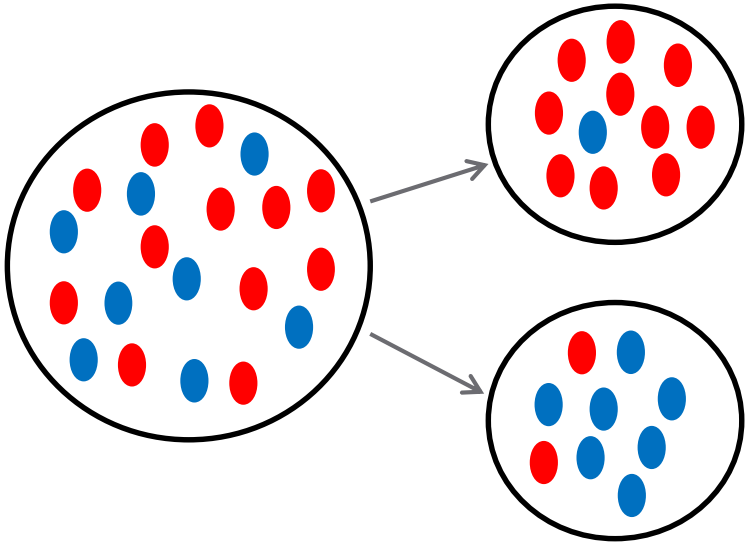
$$U(d, \xi) \equiv D_{\mathrm{KL}}[p(t|d, \xi)||p(t|\xi)] \qquad U(\xi) = I[t : d|\xi] = H[p(t|\xi)] - H[p(t|d, \xi)]$$

i.e. the "supervised machine learning" utility: $U(a) = H[T] - H[T|a]$

Review: Leclercq *et al.* 2016, arXiv:1606.06758 (Appendix B)

Supervised machine learning basics

- How to compute the information gain?



child1 entropy:

$$H = -\frac{10}{11}\log_2\left(\frac{10}{11}\right) - \frac{1}{11}\log_2\left(\frac{1}{11}\right) = 0.4395$$

child2 entropy:

$$H = -\frac{7}{9}\log_2\left(\frac{7}{9}\right) - \frac{2}{9}\log_2\left(\frac{2}{9}\right) = 0.7642$$

parent entropy:

$$H = -\frac{8}{20}\log_2\left(\frac{8}{20}\right) - \frac{12}{20}\log_2\left(\frac{12}{20}\right) = 0.9709$$

weighted average entropy of children:

$$\frac{11}{20} \times 0.4395 + \frac{9}{20} \times 0.7642 = 0.5856$$

information gain for this split: $0.9709 - 0.5856 = 0.3853$ Sh

Supervised machine learning: Titanic example

Notebook 14: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/Machine_Learning_basics.ipynb