CS340 Machine learning
Information theory

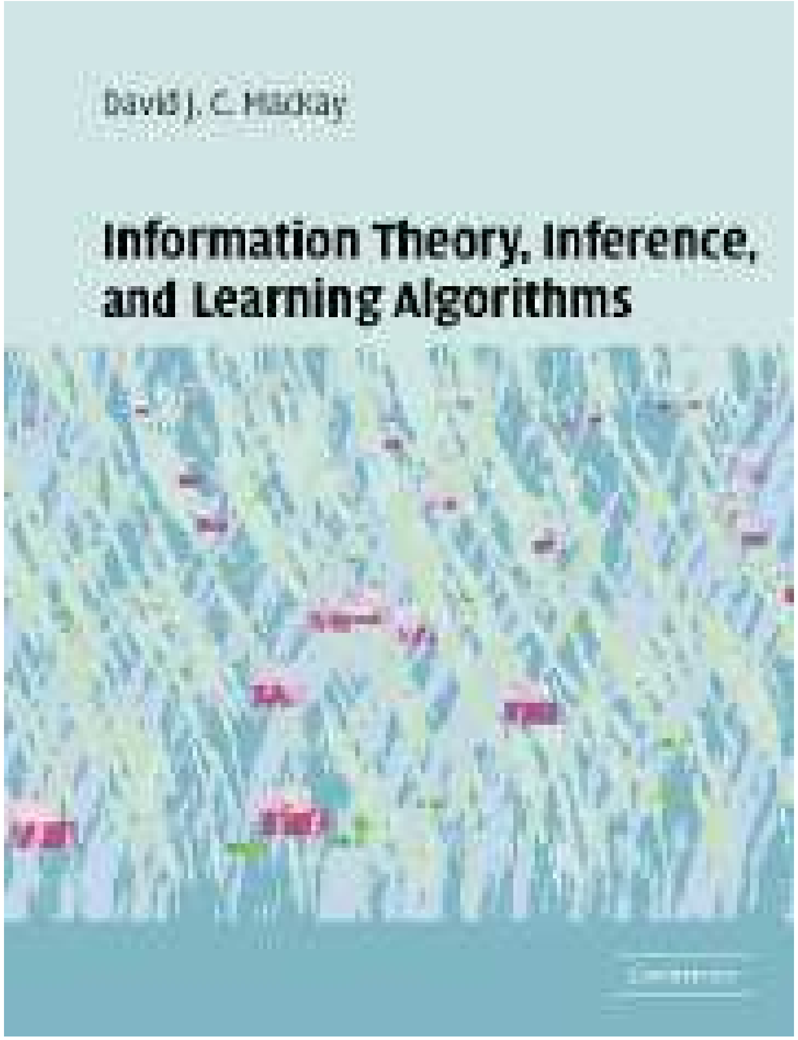
Announcements

- If you did not get email, contact hoytak@cs.ubc.ca
- Newsgroup ubc.courses.cpsc.340
- Hw1 due wed – bring hardcopy to start of class
- Added knnClassify.m, normalize.m
- Add/drop deadline tomorrow

Information theory

- Data compression (source coding)
  - More frequent events should have shorter encodings
- Error correction (channel coding)
  - Should be able to infer encoded event even is message is corrupted by noise
- Both tasks require building probabilistic models of data sources, p(x), and hence are related to machine learning
- Lower bounds on coding length and channel capacity depend on our uncertainty about p(x), defined in terms of *entropy*

Info theory & ML



CUP, 2003, freely available online on David Mackay's website

Entropy

- Consider a discrete random variable X $\in$ {1,…,K}
- Suppose we observe event X=k. The info content of this event is related to its surprise factor

$$h(k) = \log_2 1/p(X=k) = -\log_2 p(X=k)$$

- The entropy of distrib p is the average info content

$$H(X) = -\sum_{k=1}^{K} p(X=k) \log_2 p(X=k)$$

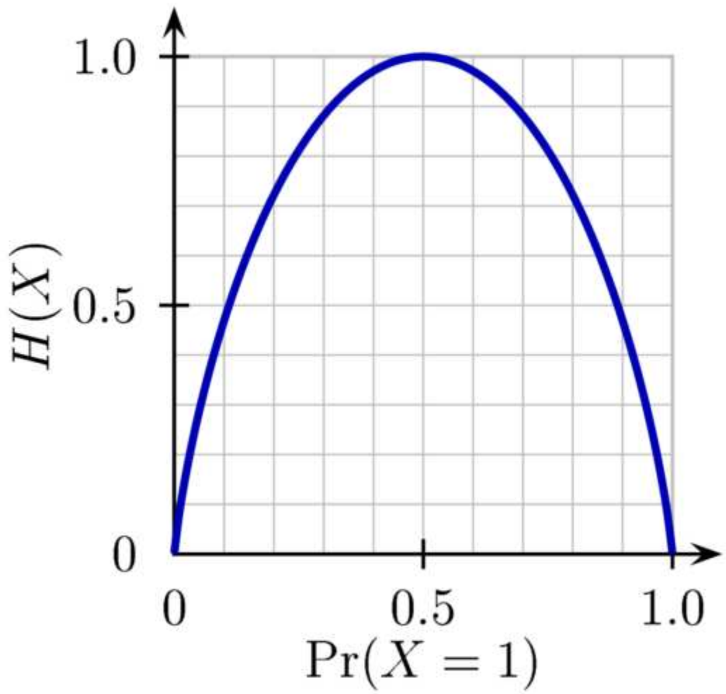
- Max entropy = uniform, min entropy = delta fn

$$0 \leq H(X) \leq \log_2 K$$

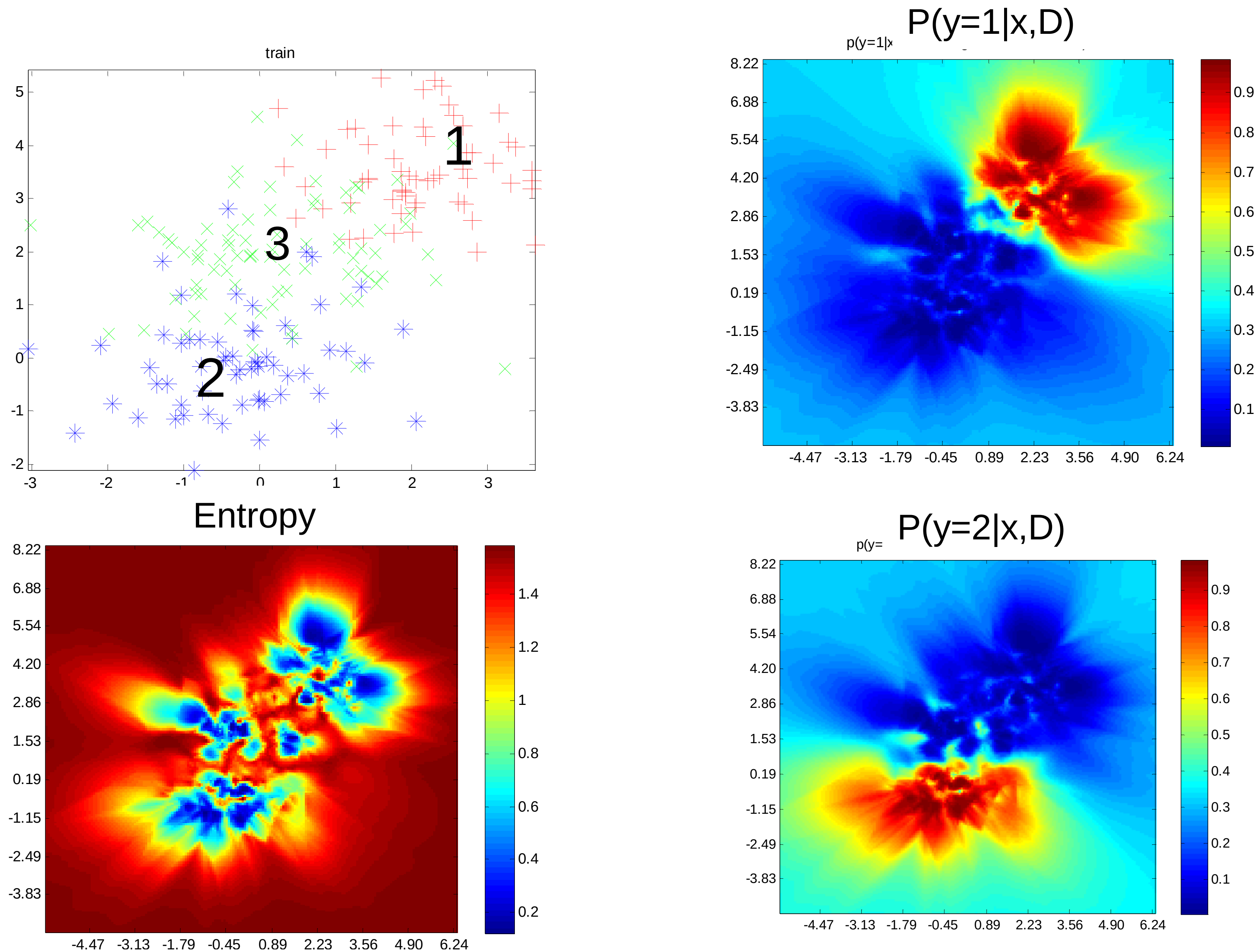
Binary entropy function

- Suppose X $\in$ {0,1}, p(X=1)=θ, p(X=0)=1- θ
- We say X ~ Bernoulli(θ)

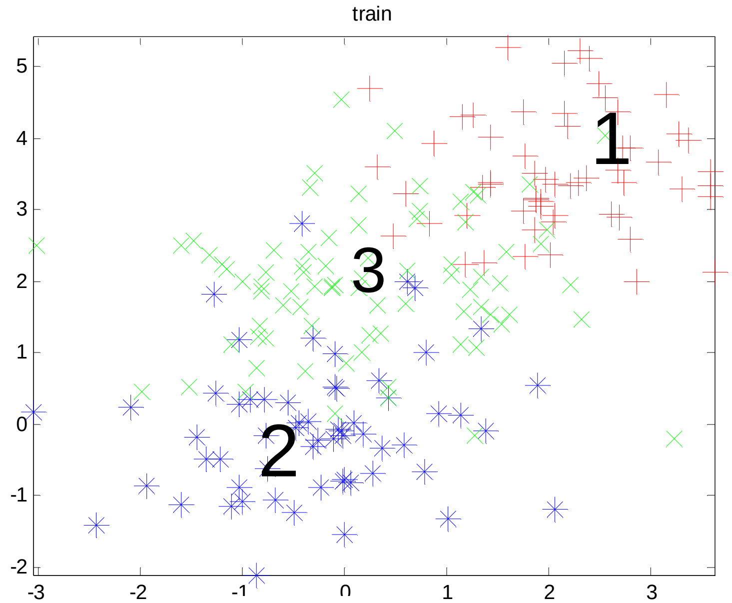
$$\begin{aligned} H(X) = H(\theta) &= -[p(X=1)\log_2 p(X=1) + p(X=0)\log_2 p(X=0)] \\ &= -[\theta \log_2 \theta + (1-\theta)\log_2(1-\theta)] \end{aligned}$$

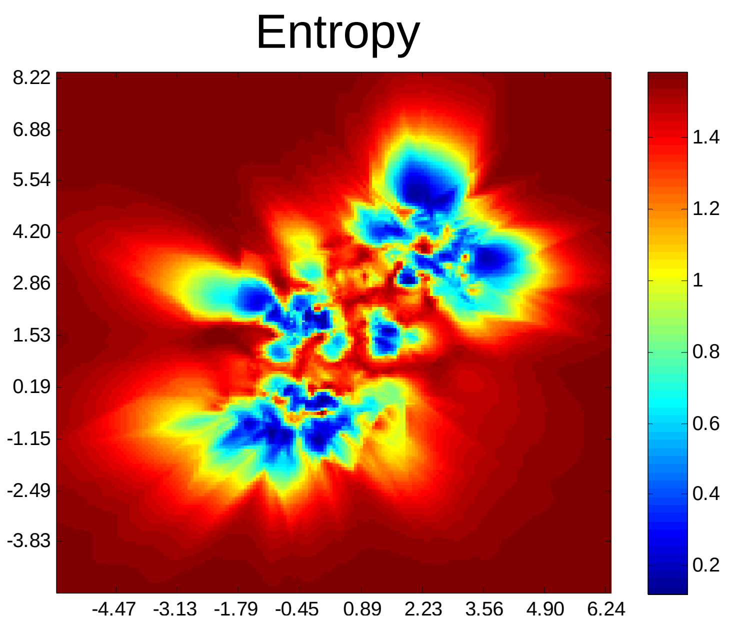


Entropy of p(y|x,D) for kNN



Active learning





- Suppose we can request the label y for any location (feature vector) x.
- A natural (myopic) criterion is to pick the one that minimizes our predictive uncertainty

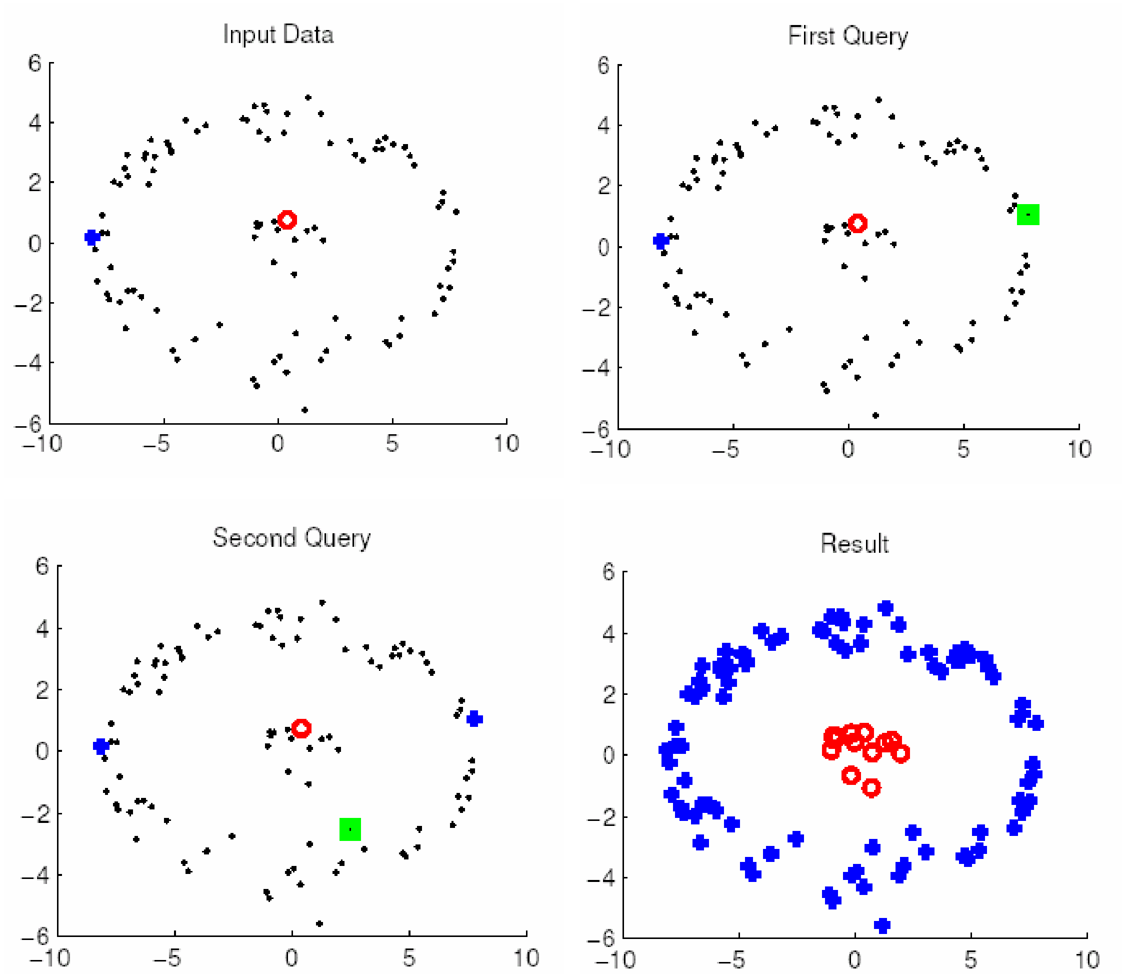
$$x^* = \arg\min_{x \in \mathcal{X}} H(p(y|x, D))$$

- Implementing this in practice may be quite difficult, depending on the size of the X space, and the form of the probabilistic model p(y|x)

Will cover later

Active learning with Gaussian Processes

If we assume the yi labels are correlated with their nearest neighbors,
we can propagate information and rapidly classify all the points



Nando de Freitas

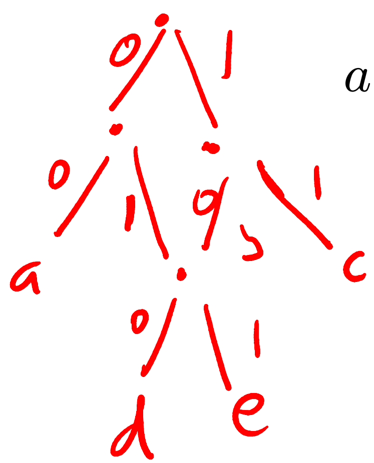
Will cover later

Entropy & source coding theorem

- Shannon proved that **the minimum number of bits needed to encode an RV with distribution p is H(p)**
- Example: X in {a,b,c,d,e} with distribution

$$p(a) = 0.25, p(b) = 0.25, p(c) = 0.2, p(d) = 0.15, p(e) = 0.15$$

- Assign short codewords (00,10,11) to common events (a,b,c) and long codewords (010,011) to rare events in a prefix-free way



$$a \to 00, b \to 10, c \to 11, d \to 010, e \to 011$$

$$001011010 \to 00, 10, 11, 010 \to abcd$$

Build tree bottom up – Huffman code

Not on exam

Example cont'd

- Example: X in {a,b,c,d,e} with distribution

$$p(a) = 0.25, p(b) = 0.25, p(c) = 0.2, p(d) = 0.15, p(e) = 0.15$$

$$a \rightarrow 00, b \rightarrow 10, c \rightarrow 11, d \rightarrow 010, e \rightarrow 011$$

- Average number of bits needed by this code

$$0.25 * 2 + 0.25 * 2 + 0.2 * 2 + 0.15 * 3 + 0.15 * 3 = 2.30$$

- Entropy of distribution: H = 2.2855
- To get closer to lower bound, encode blocks of symbols at once (arithmetic coding)

Not on exam

Joint entropy

- The joint entropy of 2 RV's is defined as

$$H(X,Y) = -\sum_{x,y} p(x,y) \log_2 p(x,y)$$

- If X and Y are independent

$$X \perp Y \iff p(X,Y) = p(X)p(Y)$$

then our uncertainty is maximal (since X does not inform us about Y or vice versa)

$$X \perp Y \iff H(X,Y) = H(X) + H(Y)$$

- In general, considering events jointly reduces our uncertainty $H(X,Y) \leq H(X) + H(Y)$ (non trivial proof – see later)
- and our joint uncertainty is >= marginal uncertainty

$$H(X,Y) \geq H(X) \geq H(Y) \geq 0$$

When is H(X,Y) = H(X)?

Example

- Let X(n) be the event that n is even, and Y(n) be the event that n is prime, for n ∈ {1,…,8}

n	1	2	3	4	5	6	7	8
X	0	1	0	1	0	1	0	1
Y	0	1	1	0	1	0	1	0

- The joint distribution = normalized counts

		Y: 0	Y: 1
X	0	1/8	3/8
	1	3/8	1/8

$$H(X,Y) = -[\frac{1}{8}\log_2\frac{1}{8} + \frac{3}{8}\log_2\frac{3}{8} + \frac{3}{8}\log_2\frac{3}{8} + \frac{1}{8}\log_2\frac{1}{8}] = 1.8113$$

What is H(X) + H(Y)?

Example cont'd

- The joint and marginal distributions are

		Y		
		0	1	p(X)
X	0	1/8	3/8	4/8
	1	3/8	1/8	4/8
p(Y)		4/8	4/8	

- Hence H(X)=H(Y)=1, so

$$H(X,Y) = 1.8113 < H(X) + H(Y) = 2$$

Conditional entropy

- H(Y|X) is expected uncertainty in Y after seeing X

$$
\begin{aligned}
H(Y|X) &\stackrel{\text{def}}{=} \sum_x p(x) H(Y|X=x) \\
&= -\sum_x p(x) \sum_y p(y|x) \log p(y|x) \\
&= -\sum_{x,y} p(x,y) \log p(y|x) \\
&= -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)} \\
&= -\sum_{x,y} p(x,y) \log p(x,y) - \sum_x p(x) \log \frac{1}{p(x)} \\
&= H(X,Y) - H(X)
\end{aligned}
$$

When is H(Y|X)=0? When is H(Y|X) = H(Y)?

Information never hurts

- Conditioning on data always decreases (or at least, never increases) our uncertainty, *on average*

$$
\begin{aligned}
H(X,Y) &\leq H(Y) + H(X) \text{ from before} \\
H(Y|X) &= H(X,Y) - H(X) \text{ from above} \\
&\leq H(Y) + H(X) - H(X) \\
&\leq H(Y)
\end{aligned}
$$

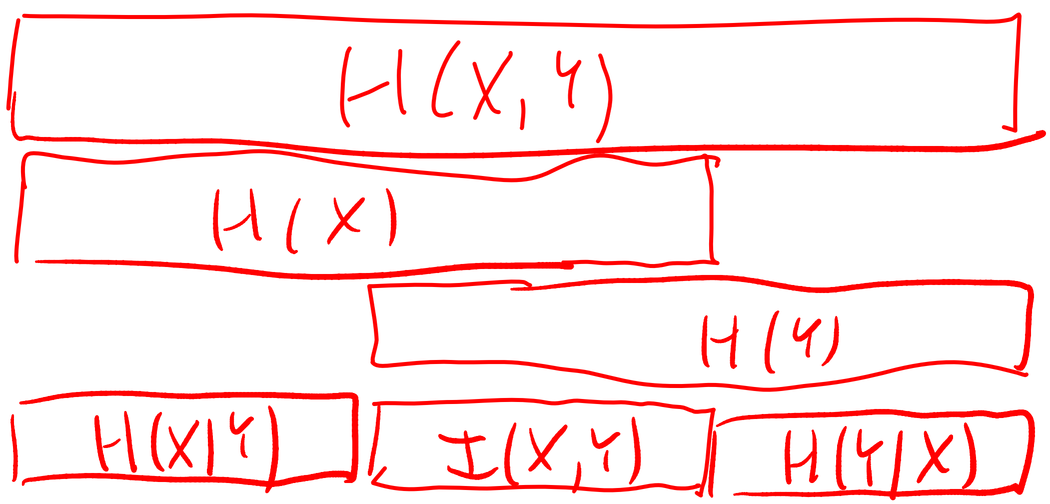
Mutual information

- I(X,Y) is how much our uncertainty about Y decreases when we observe X

$$
\begin{aligned}
I(X,Y) &\stackrel{\text{def}}{=} \sum_y \sum_x p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= -H(X,Y) + H(X) + H(Y) \\
&= H(X) - H(X|Y) = H(Y) - H(Y|X)
\end{aligned}
$$

- Hence

$$H(X,Y) = H(X|Y) + H(Y|X) + I(X,Y)$$



Mackay 9.1

Mutual information

- MI captures dependence between RVs in the following sense:

$$I(X,Y) \geq 0 \quad \text{and } I(X,Y) = 0 \iff X \perp Y$$

- If $X \perp Y \Rightarrow I(X,Y)=0$ is easy to show; $I(X,Y)=0 \Rightarrow X \perp Y$ is harder.

- This is more general than a correlation coefficient, $\rho \in [-1,1]$ which is only captures linear dependence

- For MI, we have

$$0 \leq I(X,Y) \leq H(X) \leq \log_2 K$$

When is $I(X,Y) = H(X)$?

Example

- Recall the even/ prime example with joint, marginal and conditional distributions

X \ Y	0	1	p(X)
0	1/8	3/8	4/8
1	3/8	1/8	4/8
p(Y)	4/8	4/8	

p(Y|X)		
	1/4	3/4
	3/4	1/4

- Hence

$$H(Y|X) = -[\frac{1}{8}\log_2\frac{1}{4} + \frac{3}{8}\log_2\frac{3}{4} + \frac{3}{8}\log_2\frac{3}{4} + \frac{1}{8}\log_2\frac{1}{4}] = 0.8113$$

$$I(X,Y) = H(Y) - H(Y|X) = 1 - 0.8113 = 0.1887$$

cond = normalize(joint) = joint ./ repmat(sum(joint,2), 1, Y)

Relative entropy (KL divergence)

- The Kullback-Leibler (KL) divergence is defined as

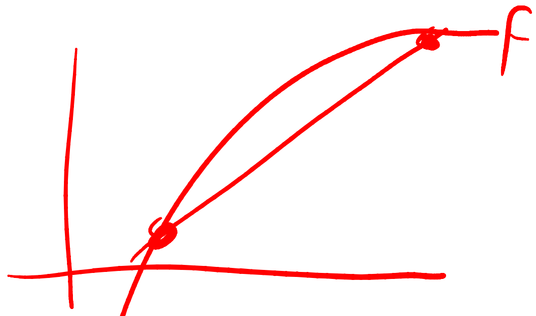
$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = -H(p) - \sum_x p(x) \log q(x)$$

- where $\sum_x$ p(x) log q(x) is the cross entropy
- KL is the average number of *extra* bits needed to encode the data if we think the distribution is q, but it is actually p.
- KL is not symmetric and hence not a distance.
- However, KL(p,q) >= 0 with equality iff p=q.

Jensen's inequality

- A concave function is one which lies above any chord

$$f(\lambda x_1 + (1-\lambda)x_2) \geq \lambda f(x_1) + (1-\lambda) f(x_2)$$



- Jensen: for any concave f,

$$E[f(X)] \leq f(E[X]) \quad \sum_x p(x) f(x) \leq f(\sum_x p(x))$$

- Proof by induction: set

$$\lambda = p(x=1), \;\; 1-\lambda = \sum_{x=2}^{K} p(x)$$

Not on exam

Proof that KL >= 0

- Let f(u)= log 1/u be a concave fn, and u(x)=p(x)/q(x)

$$\begin{aligned} D(p||q) &= E[f(q(x)/p(x))] \\ &\geq f\left(\sum_x p(x)\frac{q(x)}{p(x)}\right) \\ &= \log(\frac{1}{\sum_x q(x)}) = 0 \end{aligned}$$

- Hence

$$I(X,Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y)) \geq 0$$

and

$$H(X) + H(Y) = I(X,Y) + H(X,Y) \geq H(X,Y)$$

Not on exam