Listen

discern intent, to target the right message

……recognize a shopper from a browser

………. gauge opinion and sentiment

………… understand what people are saying

measuring *information* … what is “news”?

“The In   nes Gleick, 2011

why did they do this?

so that *you* read the story!

“dog bites man” – not news

“man bites dog” – interesting!

why?

The New York Times

Europe

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

AFRICA AMERICAS ASIA PACIFIC **EUROPE** MIDDLE EAST

At British Inquiry, Murdoch Apologizes Over Scandal

By ALAN COWELL
Published: April 26, 2012

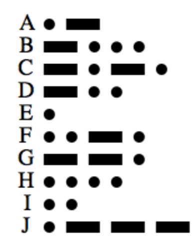
LONDON — After a day of testimony at a British judicial inquiry over his ties, friendships and disputes with British politicians, Rupert Murdoch returned to the witness stand on Thursday, saying he apologized for failing to take measures to avert the hacking scandal that has convulsed his media outpost here.

FACEBOOK
TWITTER
GOOGLE+
E-MAIL
SHARE

Claude Shannon (1948): *information* is related to surprise

a message informing us of an event that has probability p conveys

$-\log_2 p$ bits of *information* $-\log .5 = 1$



a, in, the, ..
information
miscellaneous

“It from bit” John Wheeler, 1990

when we pick up a newspaper, we are looking for maximum information, so more `surprising’ events make for better news!

in passing, you glance at some ads, and the paper makes money!

information and online advertising

when to place and ad, and *where* to place an ad?

what if the interesting news is on the sports page?

communication along a noisy channel (Shannon):

mutual information

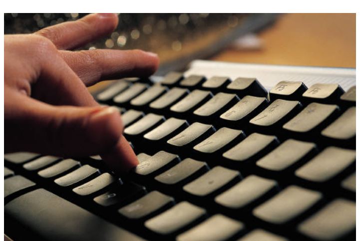
transmitted signal =
sequence of messages

channel

received signal =
sequence of messages

intent, attention

clicks, queries, content



advertising model

AdSense, keywords and mutual information

advertisers bid for keywords in Google's online auction

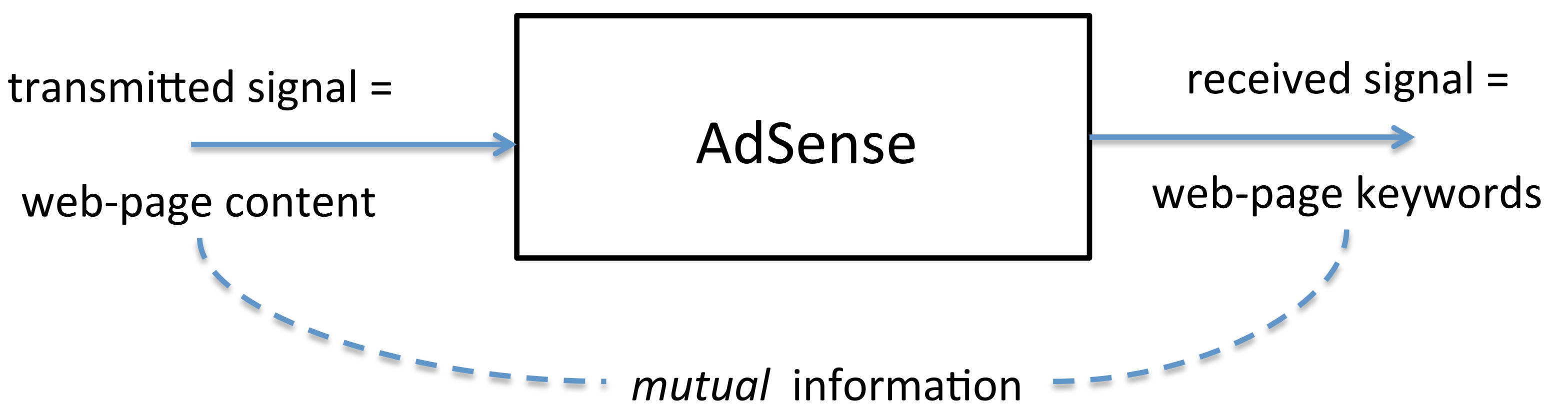
highest bidders' ads placed against matching searches

➢ increases *mutual information* between ad $s and sales..

Google's AdSense places ads in *other* web-pages as well

which keyword-bids should get ad-space on a page?

(`inverse-search': pages to keywords vs. query words to pages)



➢ how to maximize the mutual information?

TF-IDF

clearly, a word like `the' conveys much less about the content of a page on computer science than say `Turing'

rarer words make better keywords

IDF = inverse document frequency of word w = $\log_2 \frac{N}{N_w}$

(N total documents, with N_w containing w)

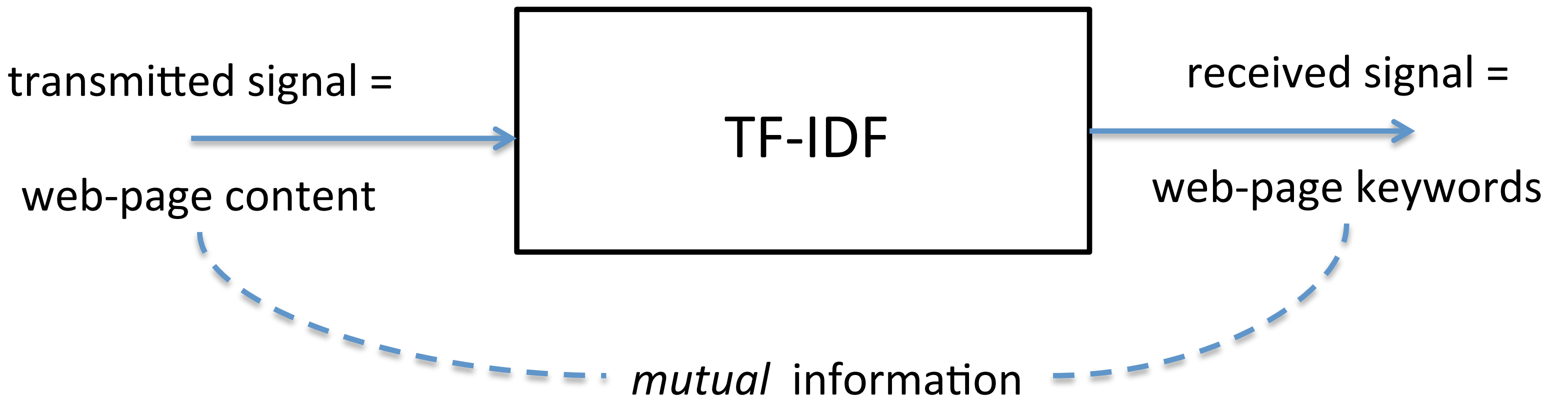
a document that contains `Turing' 15 times is more likely about computer science than one with 2 occurrences

more frequent words make better keywords

if n_w^d = frequency of w in document d

TF-IDF = term-frequency x IDF = $n_w^d \log_2 \frac{N}{N_w}$

TF-IDF and mutual information



TF-IDF was invented as a *heuristic* technique

However it has been shown that the mutual information

between *all-pages* and *all-words* is prop. to $\sum_d \sum_w n_w^d \log_2 \frac{N}{N_w}$

"An information-theoretic perspective of TF-IDF measures", Kiko Aizawa, Journal of Information Processing and Management, Volume 39 (1), 2003

keyword summarization: TF-IDF + web

TF – from text
where to get IDF?

<u>web!</u>

The course is about building `web-intelligence' applications exploiting big data sources arising social media, mobile devices and sensors, using new big-data platforms based on the 'map-reduce' parallel programming paradigm. The course is being offered ..

word	hits	IDF	TF	TF-IDF
the	25 B	50 / 25 = 2	2	2
course	2 B	50 / 2 = 25	2	9.2
media	7 B	50 / 7 = 7	1	2.8
map-reduce	0.2 B	50 / .2 = 250	1	7.9
web-intelligence	0.3 B	50 / .3 = 166	1	7.3

so the top keywords can be easily *computed*

what about choosing among these for a good *title?* …

language and *information*

transmitted signal =

`meaning'

language

received signal =

spoken or written *words*

grammatical
correctness: Chomsky

mutual information?

language is highly *redundant*: 75% redundancy in English: Shannon

"the lamp was on the d..." – you can easily guess what's next

language tries to maintain `uniform information density'

"Speaking Rationally: Uniform Information Density as an Optimal Strategy for Language Production", Frank A, Jaeger TF, 30th Annual Meeting of the Cognitive Science Society 2008

language and *statistics*

imagine yourself at a party -

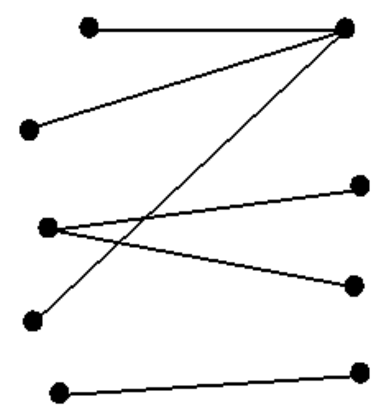
- snippets of conversation; which ones catch your interest?

a `web intelligence' program tapping Twitter, Facebook or Gmail

- what are people talking about; who have similar interests …

"similar documents have similar TF-IDF keywords" ??

- e.g. 'river' , 'bank' , 'account', 'boat', 'sand', 'deposit', …
- *semantics* of a word-use depend on context … *computable* ?
- do similar keywords co-occur in the same document?
- what if we *iterate* … in the bi-partite graph:



➢ latent semantics / topic models / … vision

is semantics – i.e., meaning, just statistics?

what about <u>intent</u>?

machine learning: surfing or shopping?

keywords: *flower*, *red*, *gift*, *cheap*;

- should ads be shown or not? - *are you a surfer or a shopper?*

machine learning is all about learning from past <u>data</u>

- past behavior of many *many* searchers using these keywords:

R	F	G	C	Buy?
n	n	y	y	y
y	n	n	y	y
y	y	y	n	n
y	y	y	n	y
y	y	y	n	n
y	y	y	y	n
…..				
……				

prediction using conditional probability

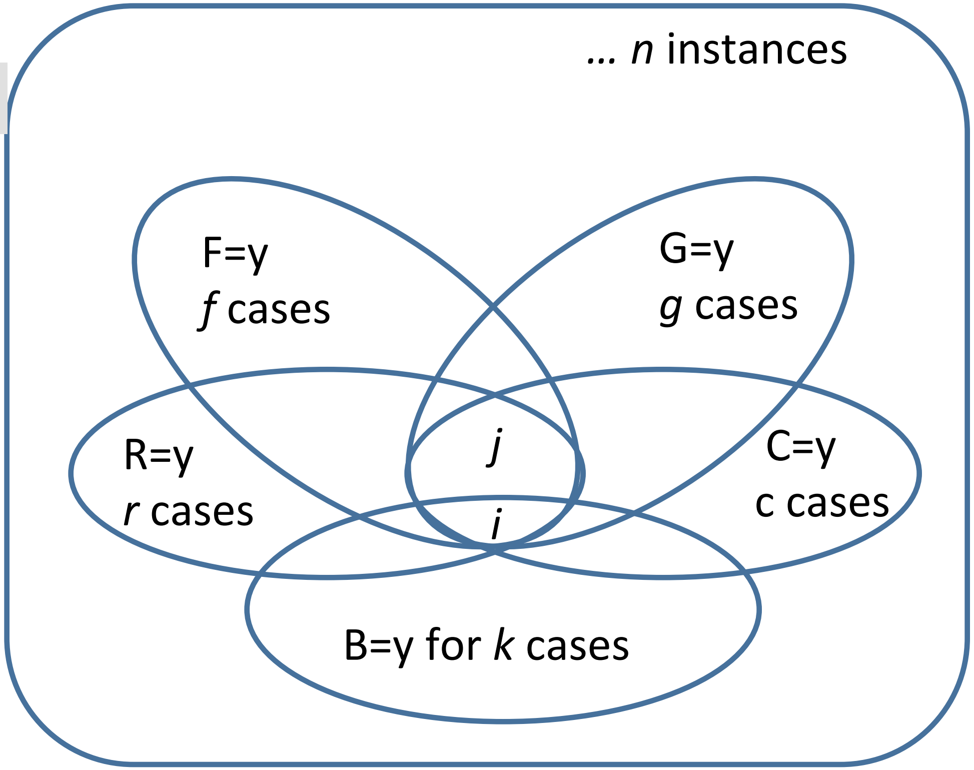
we want to determine *P*(B), given R, F, G, C

in other words, *P*(B|R,F,G,C) – *conditional* probability

R	F	G	C	B
y	y	y	y	y
n	y	y	y	y
n	n	y	y	y
n	n	n	y	y
……				
y	y	y	y	n
n	y	y	y	n
n	n	y	y	n
……				

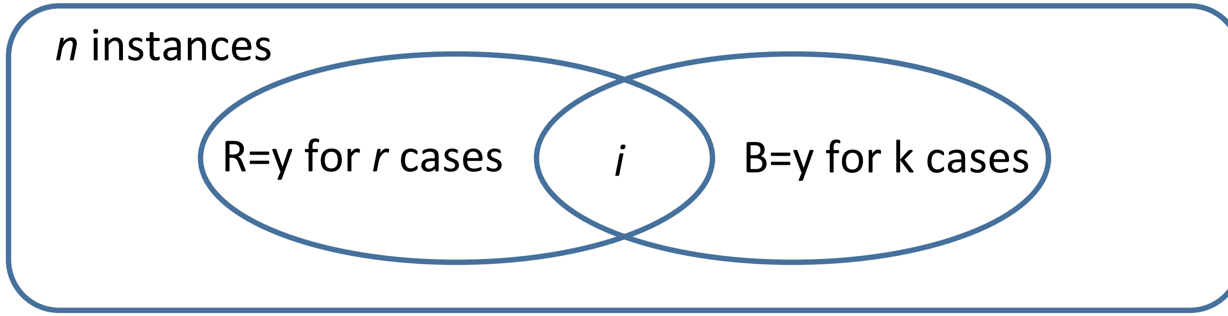
(i/n)*(n/|R⋁F⋁G⋁C|)

(j/n)*(n/|R⋁F⋁G⋁C|)



sets, frequencies and Bayes rule

#	R	B
1	y	y
2	n	n
3	y	n



probability $p(\mathrm{B}|\mathrm{R}) = \mathrm{i/r}$

probability $p(\mathrm{R}) = \mathrm{r/n}$

probability $p(\mathrm{R\ and\ B}) = \mathrm{i/n} = (\mathrm{i/r}) * (\mathrm{r/n})$

so $\quad p(\mathrm{B,R}) = p(\mathrm{B}|\mathrm{R})\, p(\mathrm{R})$

this is Bayes rule:

$$P(\mathrm{B,R}) = P(\mathrm{B}|\mathrm{R})\, P(\mathrm{R}) = P(\mathrm{R}|\mathrm{B})\, P(\mathrm{B})\ [= (\mathrm{i/k})*(\mathrm{k/n})]$$

independence

statistics of *R* do not depend on *C* and vice versa

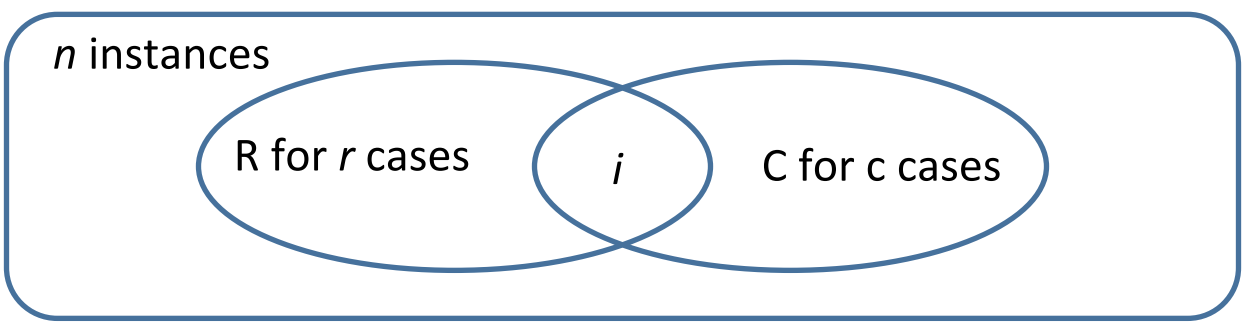
$P(\mathrm{R}) = \mathrm{r/n}$, $P(\mathrm{C}) = \mathrm{c/n}$

$P(\mathrm{R|C}) = \mathrm{i/c}$, $P(\mathrm{C|R}) = \mathrm{i/r}$

R and B are independent if and only if

$$\mathrm{i/c} = \mathrm{r/n} \quad \equiv \quad \mathrm{i/r} = \mathrm{c/n}$$

$$\text{or} \quad P(\mathrm{R|C}) = P(\mathrm{R}) \quad \equiv \quad P(\mathrm{C|R}) = P(\mathrm{C})$$



“naïve” Bayesian classifier

assumption – R and C are independent *given* B

$P(B|R,C) * P(R,C) = P(R,C|B) * P(B)$ (Bayes rule)

$= \underline{P(R|C,B) * P(C|B)} * P(B)$ (Bayes rule)

$= \underline{P(R|B)} * P(C|B) * P(B)$ (independence)

so, given values r and c for R and C

compute:

$$\frac{p(r|B=y) * p(c|B=y) * p(B=y)}{p(r|B=n) * p(c|B=n) * p(B=n)}$$

choose B=y if this is > α (usually 1), and B=n otherwise

'NBC' works the same for N features

for example, 4 features R, F, G, C …, and in general
N features, $X_1 \ldots X_N$, taking values $x_1 \ldots x_N$
compute the *likelihood ratio*

$$L = \prod_{i=1}^{N} \frac{p(x_i | B=y)}{p(x_i | B=n)} * \frac{p(B=y)}{p(B=n)}$$

and choose B=y if $L > \alpha$ and B=n otherwise

normally we take logarithms to make multiplications into additions, so you would frequently hear the term "*log-likelihood*"

sentiment analysis via machine learning

100s of millions of Tweets per day:

can listen to "the voice of the consumer" like never before

sentiment – brand / competitive position … +/- counts

count		Sentiment
2000	I really **like** this course and am learning a **lot**	positive
800	I really **hate** this course and think it is a **waste** of time	negative
200	The course is really too **simple** and quite a **bore**	negative
3000	The course is **simple**, fun and *very* **easy** to follow	positive
1000	I'm **enjoying** this course a **lot** and learning something too	positive
400	I would **enjoy** myself a **lot** *if* I did *not* have to be in this course	negative
600	I did *not* **enjoy** this course enough	negative

smoothing

$p(+) = 6000/8000 = .75$; $p(-) = 2000/8000 = .25$

$p(\text{like}|+) = 2000/6000 = .33$; $p(\text{enjoy}|+) = .16$; …. $\underline{p(\text{hate}|+) = 1/6000 = .0002}$ …

$p(\text{hate}|-) = 800/2000 = .4$; $p(\text{bore}|-) = .1$; $p(\text{like}|-) = 1/2000 = .0001$;

also … $\underline{p(\text{enjoy}|-) = 1000/2000 = .5}$! and while $p(\text{lot}|+) = .5$, $\underline{p(\text{lot}|-) = .4}$!

Bayesian sentiment analysis (cont.)

positive likelihoods	negative likelihoods
p(like|+) = .33	*p*(like|-) = .0001
p(lot|+) = .5	*p*(lot|-) = .4
p(hate|+) = .0002	*p*(hate|-) = .4
p(waste|+) = .0002	*p*(waste|-) = .4
p(simple|+) = .5	*p*(simple|-) = .1
p(easy|+) = .5	*p*(easy|-) = .0001
p(enjoy|+) = .16	*p*(enjoy|-) = .1

now faced with a *new* tweet:

I really **like** this **simple** course a **lot**

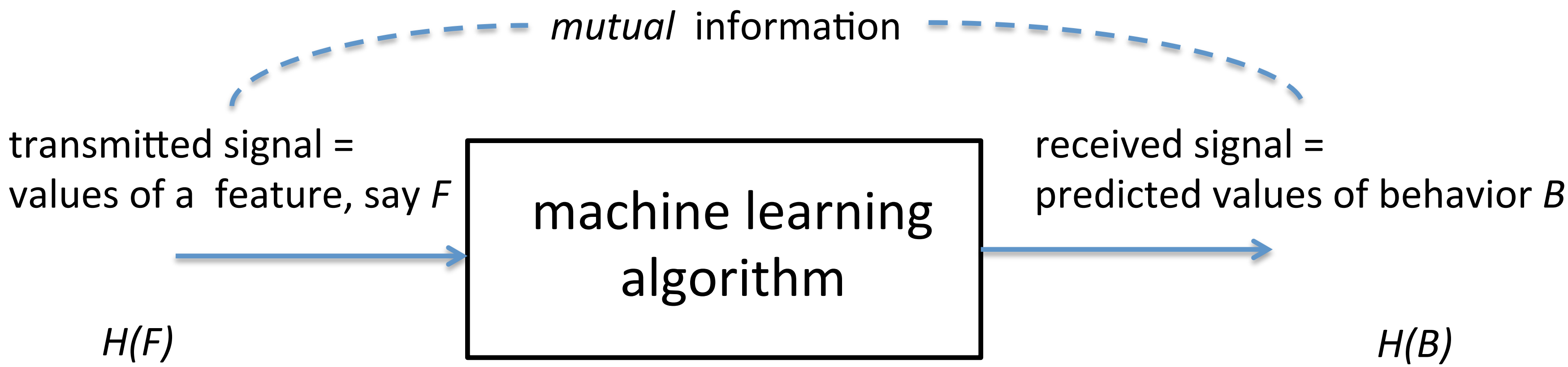
compute the *likelihood ratio:*

$$L = \frac{p(like \mid +)p(lot \mid +)[1-p(hate \mid +)][1-p(waste \mid +)]p(simple \mid +)[1-p(easy \mid +)][1-p(enjoy \mid +)]p(+)}{p(like \mid -)p(lot \mid -)[1-p(hate \mid -)][1-p(waste \mid -)]p(simple \mid -)[1-p(easy \mid -)][1-p(enjoy \mid -)]p(-)}$$

we get $L = \frac{.026}{.00005}$ >> 1 so the system labels this tweet as \`positive'

all words considered, even *absent* ones

machine learning & mutual information



mutual information between *F* and *B* is <u>defined</u> as

$$I(F,B) \equiv \sum_{f,b} p(f,b) \log \frac{p(f,b)}{p(f)p(b)}$$

H(F) + H(B) - H(F,B)

notice first that if a feature and behavior are *independent*, $p(f,b) = p(f)p(b)$ and $I(F,B) = 0$ … looks right

mutual information example

count		Sentiment
2000	I really **like** this course and am learning a **lot**	positive
800	I really **hate** this course and think it is a **waste** of time	negative
200	The course is really too **simple** and quite a **bore**	negative
3000	The course is **simple**, fun and *very* **easy** to follow	positive
1000	I'm **enjoying** this course a **lot** and learning something too	positive
400	I would **enjoy** myself a **lot** *if* I did *not* have to be in this course	negative
600	I did *not* **enjoy** this course enough	negative

p(+)=.75; *p*(-)=.25; *p*(hate)=800/8000; *p*(~hate)=7200/8000;

p(hate,+)=1/8000; *p*(~hate,+)=6000/8000; *p*(~hate,-)=1200/8000; *p*(hate,-)=.1;

$$I(H,S)=p(hate,+)\log\frac{p(hate,+)}{p(hate)p(+)}+p(\neg hate,+)\log\frac{p(\neg hate,+)}{p(\neg hate)p(+)}+p(hate,-)\log\frac{p(hate,-)}{p(hate)p(-)}+p(\neg hate,-)\log\frac{p(\neg hate,-)}{p(\neg hate)p(-)}$$

we get *I*(*HATE*,*S*) = .22

p(+)=.75; *p*(-)=.25; *p*(course)=8000/8000; *p*(~course)=1/8000;

p(course,+)=.75; *p*(~course,+)=1/8000; *p*(~course,-)=1/8000; *p*(course,-)=.25;

we get *I*(*COURSE*,*S*) = .0003

mutual information example

count		Sentiment
2000	I really **like** this course and am learning a **lot**	positive
800	I really **hate** this course and think it is a **waste** of time	negative
200	The course is really too **simple** and quite a **bore**	negative
3000	The course is **simple**, fun and *very* **easy** to follow	positive
1000	I'm **enjoying** myself a **lot** and learning something too	positive
400	I would **enjoy** myself a **lot** *if* I did *not* have to be here	negative
600	I did *not* **enjoy** this course enough	negative

p(+)=.75; *p*(-)=.25; *p*(hate)=800/8000; *p*(~hate)=7200/8000;

p(hate,+)=1/8000; *p*(~hate,+)=6000/8000; *p*(~hate,-)=1200/8000; *p*(hate,-)=.1;

$$I(H,S) = p(hate,+)\log\frac{p(hate,+)}{p(hate)p(+)} + p(\neg hate,+)\log\frac{p(\neg hate,+)}{p(\neg hate)p(+)} + p(hate,-)\log\frac{p(hate,-)}{p(hate)p(-)} + p(\neg hate,-)\log\frac{p(\neg hate,-)}{p(\neg hate)p(-)}$$

we get *I*(*HATE*,*S*) = .22

p(+)=.75; *p*(-)=.25; *p*(course)=6600/8000; *p*(~course)=1400/8000;

p(course,+)=5/8; *p*(~course,+)=1000/8000; *p*(~course,-)=400/8000; *p*(course,-)=16/80

we get *I*(*COURSE*,*S*) = .008

features: which ones, how many …?

choosing features – use those with highest MI …

costly to compute exhaustively

proxies – IDF; iteratively - AdaBoost, etc…

are more features always good?

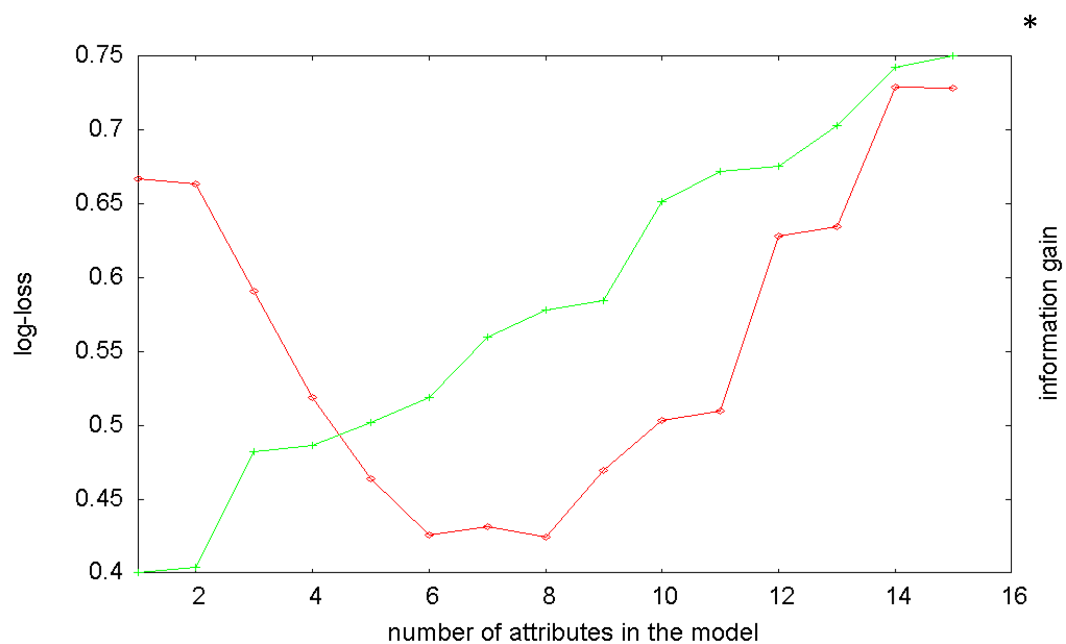
as we add features:

- NBC first improves
- then degrades! why?
- wrong features? no ..

redundant features

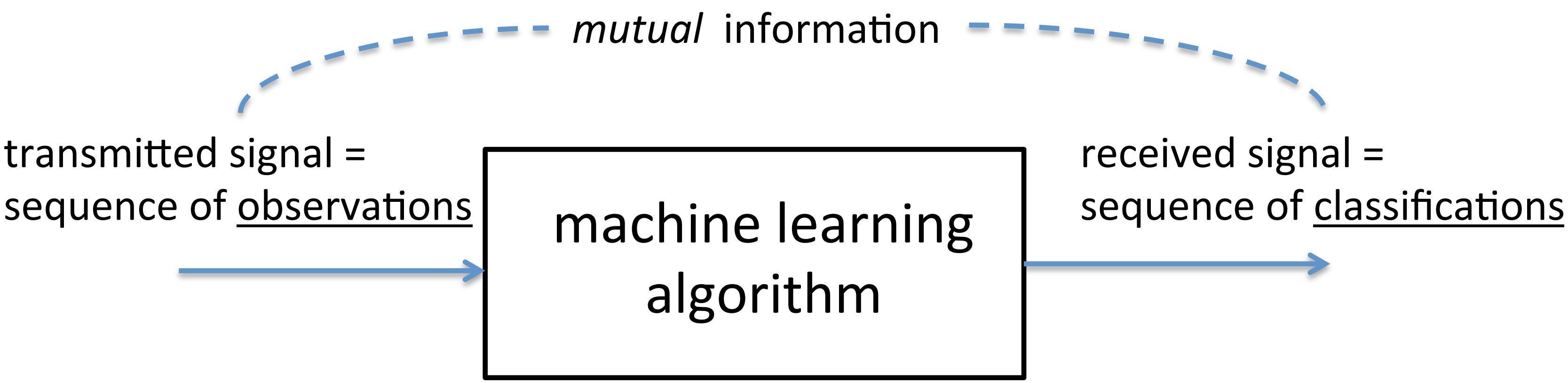
$$I(f_i, f_j) \neq \varepsilon$$

confuses NBC that assumes independent features!



*Aleks Jakulin

learning and information *theory*



Shannon defined *capacity* for communications channels:

> *"maximum mutual information between sender and receiver per second"*

what about machine learning?

> "… complexity of Bayesian learning using information theory and the VC dimension", Haussler, Kearns and Schapire, J. Machine Learning, 1994

`right' Bayesian classifier will eventually learn any concept

… how fast? … it depends on the concept itself – 'VC' dimension"

opinion mining vs sentiment analysis

100s of millions of Tweets per day:

can listen to "the voice of the consumer" like never before

sentiment – brand / competitive position … +/- counts

but: what are consumers saying / complaining about?

"book me on an American flight to New York ; I hate English food"

what does the word 'American' mean? nationality or airline?

"I only eat Kellogs cereals" vs. "only I eat Kellogs cereals"

what can you say about this home's breakfast stockpile?

"took the new car on a terrible, bumpy road, it did well though"

is this family happy with their new car?

Bayesian learning using a `bag-of-words' – is it enough?

- 'natural language processing' and 'information extraction'

recap of Listen

‘mutual information’ – M.I.

statistics of language in terms of M.I.

- keyword summarization using TF-IDF

communication & learning in terms of M.I.

- naive Bayes classifier

limits of machine-learning

- information-theoretic => feature selection
  - *suspicions about the ‘bag of words’ approach*
- more importantly – *where do features come from?*

NEXT: excursion into big-data technology

- *using* it for indexing, page-rank, TF-IDF, NBC/MI …