A CRASH COURSE ON SHANNON'S MUTUAL INFORMATION FOR CATEGORICAL DATA ANALYSIS

Matan Gavish

1. Motivation: A problem from Information Retrieval

Here a general form of a data set from the field of information retrieval. A corpus of documents (scientific papers, say) contains documents already labeled into topics (e.g physics, bio, math), a list of keyword, and a count matrix M where $M_{x,y}$ is the number of appearances (appropriately normalized to correct for differences in document length) of word x in any document of topic y. You can find a few famous datasets of this sort here
`http://archive.ics.uci.edu/ml/datasets.html?format=&task=&att=&area=&numAtt=&numIns=&type=text&sort=attUp&view=list`
or here `http://techtc.cs.technion.ac.il/`
or make it yourself: try matlab tool `http://scgroup.hpclab.ceid.upatras.gr/faculty/stratis/Papers/tmgHPCLAB-SCG-1-1-05.pdf`. The Matrix M have something on the order of 10^4 rows (keywords) and 10 columns (topics). It's tall - too tall - and thin.

Typical Questions we can ask about such a dataset:

(1) Do words contain information on the topic association of documents? Or, how well can we predict the topic of a new document from just its words content?
(2) 10^4 rows is too much for anything we'll want to do with this data. How can we map (or cluster) words into a smaller set "concepts", or meta-words, such that the concepts will contain most of the above mentioned information?

Why is (2) important? If we had such a mapping, given a keyword query we could recommend, or search by, closely related words. We could also present a small table summarizing the data instead of a $10^4 \times 10$ one.

Date: January 7, 2011.

Since M is nothing more than a two-way table of counts, the classical theory of categorical data analysis and in particular a loglinear regression model for the two-way table can (in principle, at least - don't try this at home) answer question (1). However, it's not immediately clear how to use this theory for answering question (2).

In this short note, we will present one possible answer using the prism of Shannon's Information Theory. For example, through this prism, the answer to question (1) will turn out to be $I(X;Y) = H(X) - H(X|Y) = D(p(x,y)||p(x)p(y)) = \sum_{x,y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$ where $p(x,y)$ is $M_{x,y}$ normalized to become the empirical joint distribution of words and topics. Let's start by making sense of the symbols I (mutual information) H (entropy) and D (relative entropy).

2. Shannon's Entropy

Below, RV stands for random variable. All RVs will have finite sample space. The RV X will have sample space $\Omega = \{1, \ldots, |\Omega|\}$, and probability function $p(x)$. The probability function corresponds to a probability vector $\left(p_1, \ldots, p_{|\Omega|}\right)$.

All the content of §2 and §3 appeared in Shannon's original, world-changing, highly recommended paper [Shannon 1948] which started Information Theory (Note that his proofs in this paper are not very rigorous by mathematical standards). The explanation below is quoting [CT] almost verbatim.

Consider RVs X_1 and X_2on $\{0,1,2\}$ with probability vectors $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ and $\left(\frac{1}{2}, 0, \frac{1}{2}\right)$. Which of them "has more uncertainty"?

How to measure the degree of uncertainty? clearly this question ignores the actual values of the RV and looks only at the underlying probability vector. That is, a function quantifying degree of uncertainty of probability vectors of length n using real numbers is a function $H_n : \Delta^{n-1} \to \mathbb{R}$ (Δ^{n-1} is the set of probability vectors with n entries). It's natural to expect a few properties from such sequence of functions, e.g. that they should be continuous. It turns out that most sets of properties you'll suggest will force that

$$H_n(p_1 \ldots, p_n) \propto -\sum p_i \log(p_i) \ .$$

See [Shannon 1948, Csiszár 2008] and [CT, pp. 53 problem 2.46] for such axiomatic characterizations. The proportion constant is just a choice of units. For Shannon's entropy, whose

first interpretation is "the only reasonable measure of uncertainty of a RV with finitely many values", we thus have

Definition 2.1.

$$H(p_1 \ldots, p_n) = -\sum_{i=1}^{n} p_i \log(p_i) ,$$

where here and below the base of the logarithm is 2, so that Entropy is measured in Bits.

This definition for a single RV X gives $H(X) = -\sum_{x \in X} p(x) \log p(x)$, for jointly distributed RVs, $H(X,Y) = -\sum_{x \in X, y \in Y} p(x,y) \log p(x,y)$ and so on for $H(X_1 \ldots X_n)$. We define the conditional entropy of Y given X naturally to be $H(Y|X) = \sum_{x \in X} p(x) H(Y|X = x)$.

Exercise 2.2. Some properties (prove yourself or see [CT]).

(1) $H(X) \geq 0$
(2) $H(X,Y) = H(X) + H(Y|X)$ (aka the chain rule. See if you can locate this fact in Shannon's paper [Shannon 1948])
(3) $\mathrm{H}(X|Y) \leq H(X)$ with equality iff they are independent
(4) $H(X_1 \ldots X_n) \leq \sum_{i=1}^{n} H(X_i)$ with equality iff they are mutually independent

(Some of these may be nontrivial. Convince yourself that all these are natural properties to expect from a measure of uncertainly.)

Exercise 2.3. Plot $H(X)$ for $X \sim Bernoulli(p)$ (a p-coin flip) and $X \sim Uniformon\ \Omega$.

Exercise 2.4. What can you say about the entropy of a simple random walk as function of the step number?

2.1. **Interpretations of H.** The interpretation above - a measure of uncertainly - is too vague to be useful.

Interpretation 1:

we play a game where a value is drawn from X and we are to discover it by asking yes/no questions. We decide on an asking strategy in advance. Then $H(X)$ is the smallest minimal value of expected number of questions needed.

For example, let X have probability vector $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$. It would make the most sense to ask "is it 1?" first, and if the answer is no, to ask "is it 2?". The expected number of question we will

need is $\frac{1}{2}\cdot 1+\frac{1}{2}\cdot 2=\frac{3}{2}$. You can check that $H(X)=\frac{3}{2}$ bits, and try to find a strategy that required less than $\frac{3}{2}$ questions in expectations. In general the expected number of questions is at most 1 bit from $H(X)$ (see [CT, section 5.4 pp. 112]).

Interpretation 2: (aka the "operational meaning of entropy", and immediately leads to Shannon's Noiseless Coding Theorem for the iid case)

$H(X)$ is the asymptotic (as $n\to\infty$) value of $\frac{\text{expected number of bits required to describe}\,(X_1\ldots X_n)}{n}$. Equivalently, it is $\log$(number of typical values of $(X_1\ldots X_n)$), as we now explain.

Let $X_1,X_2,\ldots\sim X$ be iid RVs. Fix n and consider the RV

$$-\frac{1}{n}\log p\left(X_1,\ldots,X_n\right).$$

To understand what's going on, recall that for every function of n real variables $f(z_1,\ldots,z_n)$ we can plug $X_1\ldots X_n$ in and get a random variable defined on the sample space $\Omega^n=\Omega\times\ldots\times\Omega$. So we take the function $f(z_1,\ldots,z_n)$ to be $p(z_1\ldots z_n)$, the joint probability function of $X_1\ldots X_n$. It's self-referential and strange, but complete legal. Now, by the weak law of large numbers, as $n\to\infty$,

$$-\frac{1}{n}\log p\left(X_1,\ldots,X_n\right)=-\frac{1}{n}\sum_{i=1}^{n}\log p\left(X_i\right)\xrightarrow{P}-\mathbb{E}\left(\log p\left(X\right)\right)=H\left(X\right).$$

This convergence in probability means that for all $\varepsilon,\delta>0$ we can find n such that

$$\mathbb{P}\left(\left|-\frac{1}{n}\log p\left(X_1,\ldots,X_n\right)-H\left(X\right)\right|<\varepsilon\right)>1-\delta,$$

or equivalently, for all strings $(x_1,\ldots,x_n)\in\Omega^n$, expect for an exception set of strings in Ω^n of probability $<\delta$,

$$2^{-n(H(X)+\varepsilon)}<p\left(x_1,\ldots,x_n\right)<2^{-n(H(X)-\varepsilon)}.$$

Note that $p\left(X_1,\ldots,X_n\right)$ is a RV while $p\left(x_1,\ldots,x_n\right)$ is a number - the probability mass function of $(X_1\ldots X_n)$ evaluated at $(x_1,\ldots,x_n)\in\Omega^n$.

The proof of this is trivial, but the meaning is striking and not at all trivial. except for a small exceptional set, all strings in Ω^n have approximately the same probability. In other words, on the typical set of strings (the set where the above holds), the probability is almost uniform. Thus the number of typical strings is approximately $2^{nH(X)}$. We need $n\cdot H\left(X\right)$ bits to enumerate a set of this size, namely $H(X)$ bits per draw from X.

This fact is known in some communities as the Asymptotic Equipartition Property (AEP). The most general case, where we replace convergence in probability with convergence almost surely, and consider $X_1, X_2, \ldots$ that is an ergodic stationary process rather than just an iid sequence, is called the Shannon-McMillan-Breiman Theorem, and is proved using Birkhoff's Ergodic Theorem rather than by the weak law of large numbers.

To get an intuitive sense about the probability of the exceptional set when n is large, consider the following Gedanken experiment. The air in the room where you're reading this is approximately an ideal gas. Draw an imaginary line in the middle of the room. Any molecule flips a fair coin to decide on which half of the room it's going to be. The locations (this half or that half) of all the molecules are a draw from $(X_1 \ldots X_n)$ where $n \approx 10^{23}$, the order of magnitude of Avogadro's number. The typical strings correspond to configurations where approximately half of the air molecules are here and half are there. The exceptional strings correspond to significant difference in pressure between the two room halves, including the configuration of almost vacuum in the half you're now. in. If were never concerned about the possibility of suffocation in this room simply because too many air molecules decide to visit the other half at the same time, you're estimating that the exceptional set has very small probability. And luckily, that's true.

3. Shannon's Mutual Information and Relative Entropy

3.1. **Relative Entropy.** Suppose again that we're playing the question gave from §2. We think that the unknown value that we have to discover is drawn from some distribution $q(x)$, and decide on a question strategy accordingly, but actually it's drawn a possibly different distribution $p(x)$. What's the expected number of questions we will need?

Exercise 3.1. Convince yourself that the expected number of questions we will need $\sum p(x) \left\lceil \log \frac{1}{q(x)} \right\rceil$ questions in expectation and that this number is between $H(p)+D\left(p||q\right)$ and $H(p)+D\left(p||q\right)+1$, where -

Definition 3.2. For probability functions p and q, the Relative Entropy (or K-L Divergence) is defined as

$$D\left(p||q\right) = \sum p(x) \log \frac{p(x)}{q(x)}\,.$$

The first interpretation we can give $D\left(p||q\right)$ is, then, that is the "penalty" for preparing a code for q when the real distribution of the source is p. See [CT, thm 5.4.3 pp.115].

Many results in the mathematical sciences, including some properties of H that we have already mentioned, can be traces back to Jensen's inequality (which states that "a convex function is convex") through the following property of D :

Proposition 3.3. *For any distributions p, q, $D(p||q) \geq 0$ and equality holds iff $p \equiv q$.*

D is not a metric (why?) but is a very useful notion of "distance" between distributions.

Exercise 3.4. Show that $D(p||u) = \log|\Omega| - H(X)$ where u is the uniform distribution. Conclude that always $H(X) \leq \log|\Omega|$.

D can be related to Fisher Information (which we will not need or mention again here). See [CT, exercise 11.7 pp.401].

3.2. **Mutual Information.** Suppose that (X, Y) are jointly distribution RVs. Recall that $H(X|Y) \leq H(X)$. How much uncertainty was lost by conditioning on Y ?

Let X, Y be jointly distributed RVs. Their mutual information is defined as

$$I(X;Y) = H(X) - H(X|Y) \ .$$

So the first interpretation of I is the reduction in uncertainty in X due to the knowledge of Y .

Exercise 3.5. Show that $I(X;Y) \geq 0$ with equality iff X and Y are independent and that $I(X;X) = H(X)$.

Exercise 3.6. Show that $I(X;Y) = H(X) + H(Y) - H(X,Y) = D(p(x,y)||p(x)p(y))$ where $p(x), p(y)$ are the marginals. Conclude that $I(X;Y)$ is symmetric in X, Y.

Have a look at [CT, fig 2.2, pp.22] and make sure you understand the following decomposition of uncertainty of jointly distributed pair (X, Y):

$$H,(X,Y) = H(X|Y) + H(Y|X) + I(X;Y) = H(X) + H(Y) - I(X;Y) \ .$$

So the second interpretation of I is the relative entropy of the joint distribution $p(x, y)$ w.r.t the product of the marginals. Indeed I is smaller the closer the joint distribution $p(x, y)$ is to the distribution of independent RVs with the same marginals. It's an informal reason to say that I measures the amount of information that X holds on Y . There are formal reasons to say so:

The operational meaning of Mutual Information:

To fully grasp this we need to understand the statement of Shannon's Noisy Channel Coding Theorem, which is beyond our scope. Here is an informal explanation about Shannon's model for noisy communication channel, channel capacity, codes and the Noisy Channel Coding Theorem. Hopefully this will convince the reader to understand the statement of the theorem, which, beside being a real intellectual gem, provides the operational meaning of Mutual Information.

Shannon proposed to model a noisy communication channel by a condition probability $p(y|x)$. The space $\Omega_X = \Omega$ is the alphabet of symbols we can send through the channel. The space Ω_Y (which here we assume to be Ω for simplicity) is the alphabet received on the over side of the channel. When sending x, the symbol received is random with distributions $Y|X = x$. Note that no marginal for X (and hence no marginal for Y) is specified, or necessary. A simple example is $\Omega = \{1, 2\}$, $Y|X = 1 \sim (1-p, p)$ and $Y|X = 2 \sim (p, 1-p)$ for some fixed p. In other words, with probability p the symbol in the binary alphabet is flipped and otherwise it is transmitted correctly.

Now suppose there is a set $\{1 \dots k\}$ of messages (different from the alphabet used by the channel) that we wish to transmit reliably through the channel. The way to achieve reliable communication through the channel is to use it multiple times, say n times, and to use the $|\Omega|^n$ possible strings we can transmit *redundantly* so that, even if strings are changed by the channel, we will still be able to decode the message correctly with small probability of error. This is done by partitioning the $|\Omega|^n$ possible strings we can transmit by using the channel n times into k sets, one for each message: $\Omega^n = \biguplus_{i=1}^{k} A_i$. This is called coding the message. To transmit message t, $1 \leq t \leq k$, we choose a specific string in A_t and send it. Due to the noise, the string received the other end of the channel, is not the one transmitted. Suppose it belongs to A_r, $1 \leq r \leq k$. We then decode by announcing that r is the decoded message. The probability of error is the probability that $t \neq r$. If we choose a smart code, we may be able to guarantee that the maximal probability of error is small, and will decrease to 0 as the number of uses of the channel n becomes large. In a nutshell, Shannon's noisy channel coding theorem states:

(1) There exists a number C, called the channel capacity, such that for any $R < C$ it is possible to find a sequence of codes (a code for every n) allowing to transmit $k = 2^{nR}$ messages with n uses of the channel, such that the maximal probability of error is decreasing to zero as $n \to \infty$.

(2) If $R > C$, no such sequence of codes is possible.
(3) $C = \max_{p(x)} I(X;Y)$. (The maximum is taken over all marginals $p(x)$. Fixing $p(x)$ we have a joint distribution $p(x,y)$ and thus it makes sense to write $I(X;Y)$).

Concisely, the maximal number of distinguishable messages that we can transmit with n uses of the channels grows as $2^{n \cdot \max_{p(x)} I(X;Y)}$. This is known as the operational meaning of mutual information.

4. The Data Processing Inequality and Sufficiency

There is a lot to say about connections between statistics and information theory (see [CT, ch. 11]). Here is a connection that is easy to explain, easy to prove, and sheds new light about the classical notion of sufficiency. The below is taken mostly from [CT, section 2.8, pp.34].

Definition 4.1. Jointly distributed RVs (X, Y, Z) are said to form a Markov Chain if Z is independent of X conditional on Y, namely if their joint probability function factors as $p(x,y,z) = p(x)p(y|x)p(z|y)$. When this holds we write $X \to Y \to Z$.

Note that for any X, Y and any function g we have $X \to Y \to g(Y)$.

The Data Processing Inequality [CT, section 2.8, pp.34]. states that if X, Y, Z are RVs such that $X \to Y \to Z$ then $I(X;Y) \geq I(X;Z)$. We parse this as "going downstream in a Markov chain can only reduce mutual information with the origin".

Now consider jointly distributed RVs (Θ, X). We can think about it as a family of conditional distributions $p(x|\theta)$, indexed by θ (we can write $p_\theta(x)$ for $p(x|\theta)$) and a prior on Θ. Let's think about X as the "data" and about Θ as the "parameter" of a family of distributions. A statistic T is any ("deterministic") function of the data, and for any such T , $\Theta \to X \to T(X)$. By the data processing inequality above, $I(\Theta;X) \geq I(\Theta;T(X))$, that is, a statistic can only have less (or equal) mutual information w.r.t the parameter, compared with the data.

Definition 4.2. A statistic T is called sufficient for θ if $\Theta \to T(X) \to X$ for every distribution of Θ.

Exercise 4.3. Convince yourself that this is equivalent to the classical definition of sufficiency.

Again by the data processing inequality, if T is a sufficient for θ then $I(\Theta;X) \leq I(\Theta;T(X))$, hence $I(\Theta;X) = I(\Theta;T(X))$.

Exercise 4.4. Convince yourself that the converse is true, namely that T is sufficient iff $I(\Theta;X) = I(\Theta;T(X))$ for any distribution of Θ.

We gained a valuable insight on sufficiency: it is equivalent to having the same amount of mutual information with the parameter, as the data. In other words, sufficient statistic preserves mutual information with the parameter. Moreover, we can now quantify sufficiency: if T is any statistic, $0 \leq I(\Theta;T(X)) \leq I(\Theta;X)$. Even if T is not sufficient as per the strict definition, the higher $\max_{p(\theta)} I(\Theta;T(X))$ (which happens to be the channel capacity of the channel defined by $T(X)|\Theta = \theta$), the closer T is to being sufficient.

Another insight to be had from this discussion is - why is minimal sufficient statistic considered a compact representation of the data? Well, if U is a minimal sufficient statistic (namely a function of any other sufficient statistic) for any sufficient statistic T, $X \to T(X) \to U(X)$ and by the data processing inequality, $I(X;T(X)) \geq I(X;U(X))$. So, what's good a minimal sufficient statistic from this viewpoint is that it minimizes the mutual information with the data (hence offering a compact representation) while preserving mutual information with the parameter.

5. Back the documents data - the"Information Bottleneck" Method

The "Information Bottleneck Method" is a framework for finding simplified versions of a RV X that preserve as much as possible mutual information with another RV Y. There are quite a few papers about it. A recommended introduction is [Shamir et al 2010], which also contains the relevant references .

Let us return to the topics-words matrix M from §1. We think about M after normalization as the joint probability function of the RV X , words, and the RV Y, topics. Now, as in the previous section, think about Y as the parameter and X as the data. The desired simplified variable "concepts" will be denoted by T. T can be simply a function of X, which would correspond to hard clustering of the words, but more generally, it can be a RV jointly distributed with X. To say that T is a representation of X that does not depend on Y is equivalent to saying that we are looking for a "concepts" variable T such that $Y \to X \to T$.

We would like the representation T to be simple with respect to X. On the other hand, we want it to hold as much information about the topics Y , namely as sufficient for Y, as possible. These objectives - low complexity and sufficiency - are contradictory, and there will be a trade-off. As in our discussion about sufficiency above, it is natural to measure simplicity w.r.t X by $I(X;T)$ (the lower the better), and to measure sufficiency for Y by $I(Y;T)$ (the higher the better).

Both these quantities have limited range. For$I(X;T)$, we have

$$0 \leq I(X;T) = H(X) - H(X|T) \leq H(X)$$

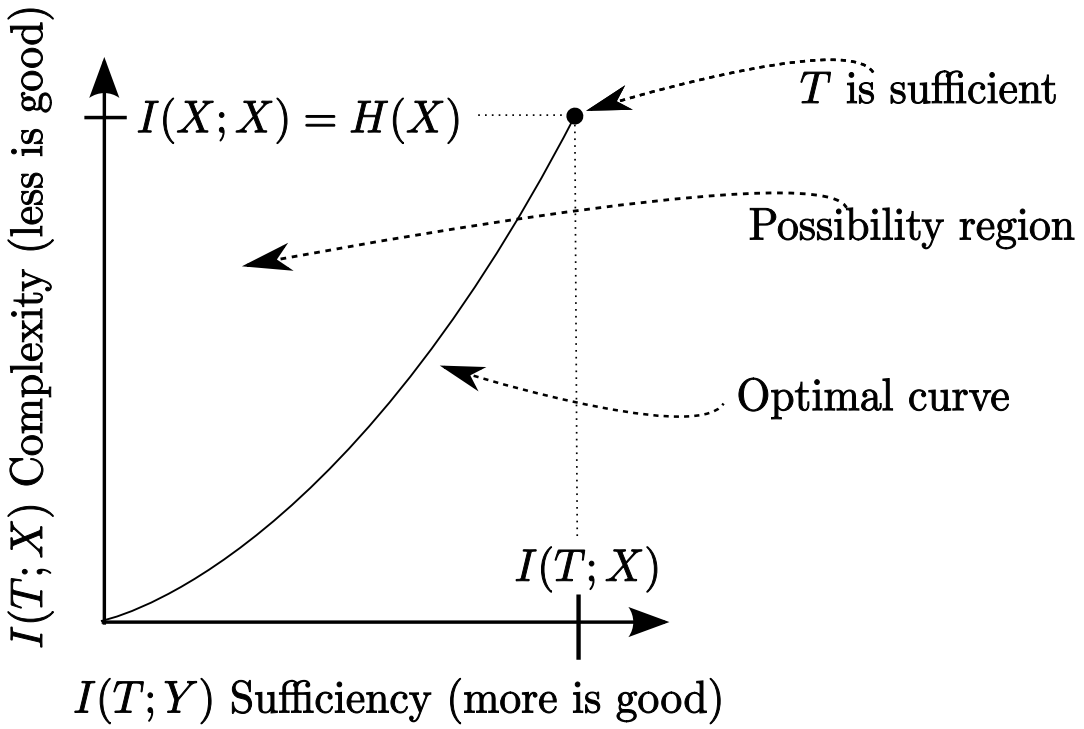
while for $I(Y;T)$ by the data processing inequality we have

$$0 \leq I(Y;T) \leq I(X;T) \ .$$

Looking for a variable T such that $Y \to X \to T$ amounts to looking for a conditional distribution $p(t|x)$, and we need to choose where on the trade off $I(Y;T)$ and $I(X;T)$ we want to be. One way to turn this into a formal problem is - let's find $p(t|x)$ that minimizes

$$\min_{p(t|x)} I(X;T) \text{ s.t. } I(Y:T) \geq c$$

that is, the variable "of least complexity" that at least a specified amount of mutual information with Y. Let $f(c)$ be the above minimal value. In [Gilad-Bachrach et al 2003] it is shown that the possibility region (the area above the graph of c) is convex. A caricature of the $I(Y;T) - I(X;T)$ plane is shown below. The optimal curve is the graph of $c \mapsto f(c)$.



Ignoring the big question of how to solve this minimization problem numerically, suppose that we chose a trade off parameter c and found the optimal distribution $p(t|x)$. We are now in a fully Bayesian setting and can try to answer, for example, question 2 from §1:

"How can we map (or cluster) words into a smaller set concepts, or meta-words, such that the concepts will contain most of the above mentioned information? ".

A trivial way to cluster words is as follows. For each value t that T takes, clustering all the words x with high $p(x|t)$ together.

We can also recommend keywords: for the keyword x , find the most relevant concept t, namely $argmax\, p\,(t|x)$. Other keywords that are related (in the context of this dataset!) are words x' with high $p\,(x'|t)$, and so on. The "Information Bottleneck framework" is a collective name given by the authors to numerous adaptations of this idea to theoretical questions and to applications.

References

[Shannon 1948] Shannon, C., A Mathematical Theory of Communication, The Bell System Technical Journal, Vol. **27**, pp. 379–423, 623–656, July, October, 1948. shannon `http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf`

[CT] Cover, T. M and Thomas, J. A., Elements of Information Theory, Second Edition, John Wiley & Sons, 2006

[Shamir et al 2010] Shamir, O., Sabato, S., Tishbi, N., Learning and Generalization with the Information Bottleneck, Theoretical Computer Science, **411**(29-30), pp.2696-2711, 2010. Learning and Generalization with the Information Bottleneck `http://research.microsoft.com/en-us/um/people/ohadsh/2009_TCS_ShamSabTish.pdf`

[Gilad-Bachrach et al 2003] Gilad-Bachrach, R., Navot, A., and Tishby, N., An information theoretic tradeoff between complexity and accuracy, Proceedings of COLT, Springer (2003) pp. 595–609

[Csiszár 2008] Csiszár, I., Axiomatic Characterizations of Information Measures, Entropy **10**, 2008, pp.261-273; DOI: 10.3390/e10030261. axiomatic `http://www.mdpi.com/1099-4300/10/3/261/pdf`