NOTES FROM INDUSTRY

Mutual Information: Prediction as Imitation

Douglas Hamilton · Follow

Published in Towards Data Science

12 min read · May 7, 2021

This is the second in a series of articles about Information Theory and its relationship to data driven enterprises and strategy. While there will be some equations in each section, they can largely be ignored for those less interested in the details and more in the implications. The first article on Entropy can be found here.

After defeating the combined armies of Norway and Ireland in Act I of Macbeth, the titular character comes across 3 witches. The Witches then make three prophecies, Macbeth will be

1. Thane of Glamis, his current title

2. Thane of Cawdor, the title of his recently defeated foe

3. And finally King of Scotland

Later on they would deliver two more prophecies with more dire connotations. Their 4th prophecies is that no man (of woman born) will slay Macbeth and their 5th that he will be safe until the nearby woods encroach on a nearby hill. To the extent you can spoil one of the most well read pieces in western lit, Macbeth is eventually slain by Macduff after Macduff's armies use camouflage made from the nearby woods to take the nearby hill[1]. Like many other oracles from Delphi to Omaha the witches' prophecies came true thanks to a goodly mix of luck, inference and sufficient vagary that makes it hard to be wrong.

Of course these days' executives tend to care less about how their competitors were born and have very little time for arborous riddles. Instead they want actionable information, clearly stated that doesn't require 5 acts and a full cast of characters to play out. Often times, particularly in operational environments, they want answers before data can be carefully collected and models rigorously tested.

Mutual Information provides both a formal as well as a notional way to approach the problem of prediction by reimaging inference as a game of optimal imitation. It give us a way to both formally and instinctively evaluate the value of information we see and understand how to cut through the noise to prevent information overload. Here we will discuss how to use information to make effective predications and the surprising ancillary utility of this approach.

Hidden Information — Three Flavors of Prediction

Back to the witches and their first 3 prophecies. Their 3rd prophecy, that Macbeth would be king, is the kind of thing most of us think of when we say prediction. The event they describe is occurring in the future, and is by no means a given. If it comes to pass it is proof of some special knowledge or foresight that others do not possess. It's the type of soothsaying practiced by basic cable weathermen, Wall Street analysts and collect call psychics.

For our purposes we need a more generalized definition of prediction, one that captures a multitude of scenarios. When a computer is shown an image of a dog and correctly labels it as a dog that is a type of prediction. Likewise when an investigator attempts to put together the root cause of a factory accident that too is a type of prediction. The thread that ties all these together is that they are trying to understand information that is otherwise hidden to them. Rather than strictly trying to understand the near term future a better definition of prediction, then, is something like: ***Prediction is the task of uncovering hidden information from expertise, intuition and observation.*** As we will continue to do throughout this series let's satisfy our gut that this makes sense. Again, predictions from the Scottish Play:

1. Thane of Glamis: Past — recovering information or investigating

2. Thane of Cawdor: Present — inferring information and coactivity

3. King of Scotland: Future — proper prediction

It is easy for us to see the 3rd prophecy as prediction. Let's check the first. If you randomly picked anyone on the planet at any time in history (e.g. you have no special information whatsoever) it is very unlikely that they would be an 11th century Scotsman, and even less likely that they would a noble from Glamis. In order to correctly classify someone as having that peerage you would need at least some special information, and even with that information you could be mistaken.

By the same token the witches 2nd prophecy also is a prediction. Most people most of the time are not the Thane of Cawdor, and while the title is in the process of being bestowed upon Macbeth the witches do not have direct knowledge of this. Through some mysticism they infer it is occurring though, again uncovering hidden information.

Finally, let's check if this definition of hidden information tracks with our intuition. Per our intuition, predictions are usually forward looking statements. Applying the same logic as before we see that most people, most of the time are not and will never be King of Scotland. Once again, correctly classifying someone as the future King of Scotland would require either considerable special information or considerable luck.

Through these three examples we have described a consistent process of prediction. First have special information or insight. Second, consider the importance and implication of that information. Finally make a falsifiable statement about an event that may be wrong. This fits very nicely with our definition above.

Now a new problem emerges. The world is full of information, most information is not useful. Knowing the weather in Peoria is unlikely to give you insight into royal lineage. Ideally you want to rely on information that is relevant, reliable and high quality and ignore information that is not. How do we measure the quality of information at our disposal?

Imitation & Mutual Information

Consider the following game: flip a coin, whatever face it lands on flip a second coin until it lands on the same face, record the results. It's not an interesting game. In fact this game has the same exact entropy as the original flip: 1 bit.

$$H\_ \xrightarrow[flip]{} HT \xrightarrow[flip]{} HH$$

$$T\_ \xrightarrow[flip]{} TT$$

Some examples from the coin flip game

We're going to change the game a little bit. Instead of just recording the result we are now going to try to predict the outcome. Without any information there is a 50/50 chance of getting a HH or TT outcome. Here's our new game:

1. Flip the first coin and hide the result

2. You may then know any one piece of information in the universe past or present,

3. Guess the final state of the game

4. Flip the second coin until it matches the first

Our new goal is to find some data that informs us about the final outcome. The solution here is obvious, uncover the first flip and you will know the game's final state. Put another way the final flip is a perfect imitation of the first. Alternatively we could flip a coin outside the game, this would tell us nothing about the final state of the game and provide no information.

In this flipping game we want to make an observation the shares maximum Mutual Information with the end state. Mutual information tells us how much one event reduces the uncertainty in another. It is the degree to which an observed signal mimics –even imperfectly — a separate signal. The second flip is a perfect facsimile of the first. A coin outside the game is independent and therefore provides no new information. Knowing someone is an 11th century Scot does not guarantee they will be king, but it gives some small amount of information you would otherwise have not had. Mutual Information is given by:

$$I(X;Y) = \sum_{X,Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \quad eqn\ 4$$

Mutual Information I(X;Y); p(x, y) is the joint probability of members of X and Y while p(x) and p(y) are their individual probabilities

Admittedly this one is a little more intimidating than the previous articles equations, but no less intuitive. The equation states that Mutual information (I) of two events (X and Y) is proportional to the probability of each set of outcomes p(x, y) for X and Y together and their independent probabilities p(x) and p(y). Just like with entropy this is a formalized way to state something we already know: the more interlinked two events are the more information they tell us about each other. If we consider the two coins in the flipping game they share a mutual information of 1 bit because they are entirely dependent. Knowing the result of the first flip removes 1 bit of uncertainty from a game with 1 bit of entropy. The coin from outside the game though offers 0 bits of Mutual Information and knowing that is irrelevant to the game's outcome.

Strategy Analysis

Informative imitation is not always as obvious as our coin flipping game. Consider the following trading strategies.

Returns for two strategies on the S&P 500 form April 2016 — April 2021

While the two time series here do not look like exact mimics they do hold maximum Mutual Information about each other.[2]

These two strategies will be regulars in our toys for understanding information so it's worthwhile to understand what they are. The blue strategy is a simple Momentum. A simple momentum strategy looks at what the market did yesterday and assumes it will do the same thing today. If the market went up yesterday then

you buy today, if it went down sell. Simple mean reversion — in orange — is the opposite: if the market went up assume its overvalued and sell today, if it went down assume its undervalued now so buy. Knowing what position one strategy takes perfectly informs you about the position of the other.

Momentum and mean reversion are interesting strategies from an information theoretic perspective for another reason as well. Recall from the previous section we defined a minimum information strategy called Buy and Hold. This strategy merely buys and holds (hence the name) a security[3].

Same strategies benchmarked against buy and hold, a lower information strategy

While momentum and mean reversion are perfect predictors of each other they also share minimum Mutual Information with Buy and Hold, our default strategy to beat. This trio, a minimum information approach and its informational adversaries form a basis for benchmarking strategies more broadly. New, more exotic approaches should improve upon these simple strategies and can be described as having a momentum or mean reversion bias. Since both mean reversion and momentum go through periods of over and under performance the ideal exotic strategy will be biased toward neither.

A quick aside on strategy. We will spend a great deal of time in this series discussing portfolio construction and trading strategies as they relate to equity investment. The reason for this is, primarily, the availability of financial data. From both an information, notional and statistical[4] perspective there is very little difference between making decisions about how and when to move in and out of a position in

the S&P 500 and how and when to expose a set of people to an advertising campaign. There is very little difference between dividing a pile of money among a portfolio of securities to capture returns (portfolio construction) and delegating a pile of tasks across personnel or machines to hit KPIs.

Optimal Guessing

Any strategy is a perfect predictor of Buy and Hold, but no optimal guessing strategy can be designed from buy and hold to predict either the actions a momentum or mean reversion strategy. In this series 'guessing' will be distinguished from 'predicting' in a way similar to how we distinguish between the two in common parlance. Guessing is haphazard and quickly done, predicting is more considerate and more time consuming. More specifically guessing is a type of prediction that has no exigent information about an event beside at most the probabilities of outcomes. The two types of guessing we will tend to discuss are

· Pure guessing: a strategy is that randomly chooses outcome from all possible outcomes

· Optimal guessing: a strategy that chooses an outcome and is correct as often as possible without careful study of exigent information

We have encountered Pure Guessing before. In the dice rolling game in the previous section we used a pure guessing strategy to correctly guess the outcome for both a fair and unfair dice 16.7% of the time. We have also encountered an optimal guessing strategy. US Large Cap Equities tend to produce positive returns, or at least have for the last ~120+ years. If that is all the information you have an optimal guessing strategy is to buy the market and walk away.

As you may have guessed by its inclusion in this article optimal guessing is intimately linked to mutual information. Let's think about our dice again. As a reminder a fair dice has a 1/6 chance of coming up any face; the unfair one has a 90% chance of '1' and 2% for each other face. A pure guessing strategy produces the same results for each, but it seems like there should be a better strategy for the unfair dice.

From the mutual information equation (eqn 4) make X the outcome of a dice roll and Y the outcome of our guess. An optimal guessing strategy will maximally align the true outcome with the guess. Another way to say this, an optimal guess

maximizes the mutual information between the roll and the guess. A strategy that does for the unfair dice is to always guess '1.' This is a pretty good imitation of the actual dice rolls we see and you'll be right 90% of the time.

Optimal guessing is crucial for quick decisions. This is the domain of real experts. Real experts use their experience to know the uninformed probabilities as well as heuristics to capture information from anecdotal observation. How to check if an expert actually has expertise is a problem for later. Optimal guessing is also crucial for low information decisions. Venture Capital and Angel Investors have very little real information about early round firms, but if they can identify a subset of types of firms that on average produce returns it makes sense to invest in all of them[5].

Information Overload & Information Pruning

One quirk form the optimal guessing strategy above, a strategy of 'always guess 1' has the same mutual information with the unfair dice as a strategy of 'always guess 2.' An information first approach says that, while 2 comes up rarely, the information it provides can be used in a better way. If we built a second rule, for example, that said 'whenever we guess 2, change it to 1' we'd perform very well. Said another way, **mutual information tells us when the data is good but the model is bad**.

If mutual information can tell us when data is good, it can also tell us when data is bad. Conjecture: you should ignore bad data. Incidentally it can also be used to tell us if data is redundant. A strategy of 'always guess 1' gives no information that a strategy of 'always guess 2' doesn't give. A simple momentum strategy gives no information a simple mean reversion strategy doesn't give. We can ignore one of those 2 strategies and be no worse off.

In the real world we seldom find perfect information substitutes though, nor do we find completely irrelevant information. Even the rain in LA can be lightly correlated with correlated with trading behavior[6]. If you combine this with our insights about the power of information, that more information always reduces uncertainty, it's no surprise that operations managers receive 50 page reports of charts and figures every morning. 50 pages that are impossible for a person to make sense of. We need a way to limit information intake to match our bandwidth to consume it.

A traditional statistical approach to pruning all these charts into something more readable is principal component analysis (PCA). PCA combines all your charts into just handful, which is the goal. Unfortunately the handful of charts are synthetically

generated admixtures of your initial 50 and have no intuitive relationship to the factory floor. A much better approach would be to remove redundant and irrelevant information and only keep the relevant and unique charts.

Again we have a formal and intuitive mutual information based solution. Formally we want to find a set of predictors with maximum mutual information with a target while also being minimally redundant with each other. Less formally we want to find signals that mimic KPIs while not mimicking each other. This is often called a Maximum Relevance Minimum Redundancy (mRMR)[7] approach to feature selection. It's not hard to write or find an mRMR algorithm to prune a morning report. It can also be done with a few rules of thumbs: do my charts look like my target right side up or upside down (relevance), does each chart look different or do a bunch of them look the same (redundancy), can I fit all the charts on 1 page per KPI (bandwidth preservation). With this strategy it's possible to reduce massive information overload to an easily consumable 5 page report.

The Basics

With entropy and Mutual information firmly established we're now in possession of the building blocks of information theory. We have seen a number of toys, allegories and anecdotes that point to the unique power and flexibility of this approach. In the next article we will use our building blocks to establish a number of Axioms and Principals that enable us to approach data driven decisions and information risk in more complex way.

[1] Macbeth, by Shakespeare…do I really need to cite this?

[2] S&P data retrieved from yahoo finance

[3] Think a stock or bond

[4] For more about the grim/infuriating statistics of Financial Markets see Taleb's *Black Swan* or Mandelbrot's *Misbehavior of Markets*

[5] At the time of writing this Andeerssen Horowitz has more than 250 active members of its portfolio (5/6/2021, https://a16z.com/portfolio/)