Practical Prelude to Machine Learning for Sport

Kyle D. Peterson [1]

[1] Sports Science Department, University of Iowa, Iowa City, IA, USA

Technical note | Statistics | Machine Learning

Headline

Practitioners in today's sport science scene are often becoming awash in high-dimensional data sets. From an athlete monitoring perspective, teams may exploit several mediums concurrently; from force platforms and hydration to wellness, GPS and more – each contributing numerous, if not hundreds, of variables. How are practitioners supposed to select markers that meaningfully influence their objective? The purpose of this document is to showcase a realistic methodology for extracting predictive variables from a soup full of options by employing elementary machine learning principles.

Discussion

From a signaling viewpoint, the world is a noisy place – in order to make sense of anything, we have to be meticulous with our attention. Thankfully, humans have become rather proficient at filtering out irrelevant, background signals. Consider the sport of tennis, for instance. In order to return an opponent's shot, an athlete makes an enormous array of intricate calculations while inundated by multifarious sensory signals. We take for granted the amount of subconscious differential calculus the nervous system is capable of calculating (to forecast ball trajectory, speed, spin, etc) while simultaneously accounting for the myriad of trivial impeding signals (climate, playing surface, spectators, etc). This is a testament of just how efficient we are at sifting importance out of extremely noisy data.

Certainly a blank-slate computer given a continuous stream of high-dimensional data would face an arduous task deciphering which signals meaningfully impact a given objective. Fortunately, there are statistical and computational tools well-equipped to identify patterns in noisy, real-world data, which will be addressed below.

The problem with r.

When given a large data set comprised of thousands of candidate predictors, a tempting tactic is to tabulate pairwise correlations into what is commonly coined a correlation matrix. This conventional data dredging attempts to unmask meaningful relations between variables of interest, and conversely, alert the practitioner of fruitless markers. We generally characterize a pairwise relationship as a cloud of points scattered astride a trend line; more dispersion indicating "noisier" data, a weaker relationship, and vice versa.

However, there is a major catch – Pearson's correlation (r) can only seek linear relations! Since r simply compares each individual data point against the overall mean (1), it is restricted to solely consider straight lines, and therefore, is unable to detect non-linear relationships. This heavy a priori assumption vastly curtails its practicality when examining macroscopic descriptions (heart rate variability, sRPE, etc) of complex, non-linear systems.

Refer to Figure 1 for simulated examples. Before any calculations are made, one could argue that the amoeboid-shaped Figure 1A presents a weak dependency in contrast to the distinct functional relationship in Figure 1B. Such a non-linear relationship in Figure 1B could be witnessed between a physiological marker (e.g., parasympathetic tone, x-axis) against a performance outcome (race times, y-axis), with optimal performance manifesting when physiological marker near the median (suggesting autonomic balance (2)). However, when measuring associative strength via Pearson r, Figure 1B is close to zero - hinting there is nothing worth delving into. This is because the linear trend, indicated by the horizontal red line, has a minute gradient compared to Figure 1A. Yet, the relationship between x and y in Figure 1B is clearly non-random, making x a potentially useful predictor of y. How do machines identify this? Luckily there are other associative measures calculated through different mechanics that are not bound by the statistical straightjacket of linearity. Let's take a look at a couple of them.

Distance correlation.

Distance correlation moderately coincides with Pearson's r, but is derived using a rather disparate notion. The method replaces our everyday concept of covariance and standard deviation with distance analogies. That is, as opposed to measuring how two variables co-vary in the distance from their respective means, distance correlation measures how two variables co-vary in the distance from *all other points* (3). This frees up the opportunity to capture non-linear dependencies between variables.

A cheesy, yet illustrative, metaphor to grasp the concept of distance correlation is to imagine a fleet of rubber duckies floating on the surface of a pond. If there is no prevailing wind, each ducky will drift in random course. Under a prevailing wind, the duckies will tend to drift in the same direction dependent upon the strength of the wind. Uncorrelated variables can be thought of as duckies drifting without a prevailing wind, whereas correlated variables can be thought of drifting under the influence of wind. In this metaphor, distance correlation uses the distances between the duckies to infer the strength of the prevailing wind, a tendency to "follow" each other. If we then allow prevailing winds to vary at different points of the pond we bring a notion of non-linearity into the analogy.

Comparable to coefficient of determination (R^2), distance correlation ranges between zero and one – zero implying independence, one indicating perfect relationship. Referring back to Figure 1, distance correlation indeed detects the present non-linear dependency, with a moderate value of 0.43 instead of the paltry r = -0.04. At first glance this discrepancy may alarm some practitioners, but the amplified relational strength is arguably more realistic with given data points. However, in the case when a true linear relationship exists between two variables, distance correlation will measure synonymously with r (3). For practitioners or researchers who fancy precision of estimates, confidence intervals can be established via bootstrap resampling. Interested readers are welcome to vet accompanied R script to explore distance correlation, as well as bootstrap, with their own data sets.

Although we have a tool which appreciates non-linear relations, are we allowed to conflate such crude summaries with predictive power? This brings us to the fascinating field within mathematics, born from computer science, which constitutes

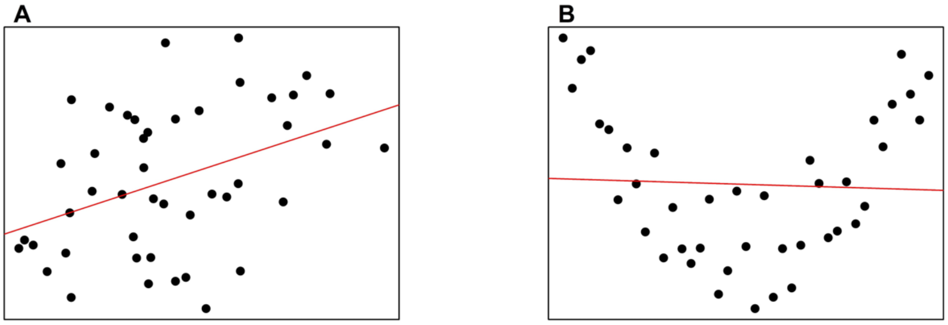

Fig. 1. Artificial data sets exhibiting linear and non-linear relationships. Horizontal red lines signify linear relation. A) r = 0.40; B) r = -0.04.

the core ingredients of popular machine learning constructs used today.

A primer to information theory.

Information theory emerged as a branch of communications engineering to quantify the length of code required to represent varying signals (4). The major concepts needed to solve such problems eventually produced fundamental measures of uncertainty and interdependence between variables, which later lead to major contributions to statistical learning. Although the topic of information is too broad for the current document to capture, the text that follows will swiftly synopsize information theory basics and outline how sport practitioners can leverage information-theoretic learning to recognize a signal blurred by noise.

A key concept to begin with is entropy – the number of bits required to transmit an outcome in the absence of noise (5). More formally, entropy is the logarithm of the number of ways to discriminate between physical states. In statistical mechanics, entropy arises as a measure of uncertainty, perhaps disorganization, in a univariate probability distribution (6). For a gentle demonstration, consider a binomial distribution conveying 90/10% probability of some event occurring. Due to the large imbalance, little uncertainty is present, warranting relatively small entropy (0.47 bits). On the contrary, reckon a second binomial distribution possessing 60/40% probability. More disorganization thus heightens uncertainty, leading to increased entropy (0.97 bits). Overall, entropy is highest (penalized) in the case of equiprobable states and lowest (rewarded) when little uncertainty is present (which is precisely the raison d'être of the logarithmic term to reflect concavity between probability and entropy). However, entropy in isolation is not exclusively useful – let's see what happens when we add another variable.

The distance between two respective distributions is called relative entropy (7). With this in mind, the relative entropy (distance) between the joint distribution $P(x,\ y)$ and the marginal distribution $P(x)P(y)$ is called mutual information (8). In other words, the distance between probability of x and y occurring together to the probability they occur together if independent (Figure 2). Therefore, mutual information is a direct measure of dependency; the expected amount of information a variable provides about another, or the loss of information that arises when falsely assuming independence (9).

To ensure a fair comparison between different distributional schemes, mutual information can be normalized to obtain a modified range between zero and one – this is called Maximal Information Coefficient (MIC) (10). This normalization thus resembles the advantages of statistical effect sizes (emphasizing magnitude of relationship while evading the confounding n). In statistical parlance, MIC (or information-theoretic learning entirely) can be vaguely viewed as nonparametric: distribution-free, assumption-free. This added flexibility should catch the eye of practitioners who wrestle with pesky transformations and/or question the efficacy of linear analyses.

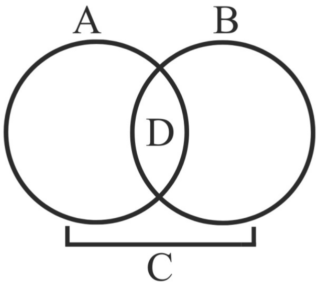

Fig. 2. A = independent marginal distribution P(x); B = independent marginal distribution P(y); C = joint distribution P(x, y); D = mutual information $I(X;Y)$.

Relating back to Figure 1B, this artificial data set produces MIC = 0.48, an effect in the same ballpark as distance correlation (0.43), yet computed from substantially different theory. Alternatively, the predictive power of Figure 1A diminished substantially to MIC = 0.14. Therefore, if these two variables were candidate predictors in a practitioner's data set, MIC would effectively identify the functional relationship within Figure 1B as being superior to Figure 1A, regardless of their linear discrepancies.

Conclusion

Although above techniques may seem rudimentary, the intention is to stimulate thought into a new arena of analysis. Information theory is the building block of some of the most sophisticated predictive models and is important to understand the fundamentals before diving head-first into any subject. An immediately useful application of MIC for practitioners is to scan every candidate predictor against dependent variable and rank in order of pairwise importance. This procedure allows MIC to act as a sieve, drawing attention to meaningfully predictive variables in the absence of statistical rigidity and avoids cherry-picking variables from human-generated biases. (Accompanied R script supplies interested users with a seamless implementation). Such a hands-off approach is rather inductive in nature; data driving hypotheses as opposed to testing data against preconceived hypotheses. Practitioners who

Table 1. Correlational derivatives versus entropy-based MIC from Figure 1.

	Pearson r	**Distance Correlation**	**MIC**
Figure 1A	0.40	0.41	0.14
Figure 1B	-0.04	0.43	0.48

embrace machine learning will ultimately enable the data to speak for itself and may be pleasantly surprised by what they discover.

Practical Applications

- Pearson's correlation only detects linear relations, which is a major limitation when describing non-linear biological systems. Alternatively, distance correlation is sensitive to both linear and non-linear dependencies.
- Information-theoretic measure, entropy, provides an inductive approach for sifting out predictive bivariate relationships from noisy, high-dimensional data sets.

References

1. Pearson K. Notes on regression and inheritance in the case of two parents. P R Soc London. 1895. 58:240-242.
2. Hedelin R, Wiklund U, Bjerle P, Henriksson-Larsen K. Cardiac autonomic imbalance in an overtrained athlete. Med Sci Sports Exerc. 2000. 32:1531-1533.
3. Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. Ann Stat. 2007. 35(6):2769-2794.
4. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948. 27:379-423.
5. Baldi P, Brunak S. Bioinformatics: the machine learning approach 2nd ed. Cambridge, Massachusetts: MIT Press; 2001. 357 p.
6. Cover TM, Thomas JA. Elements of information theory 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2006. 11 p.
7. Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat. 1951. 21(1):79-86.
8. Deisboeck TS, Kresh JY. Complex systems science in biomedicine. New York, NY: Springer Inc.; 2006. 77 p.
9. Emmert-Streib F, Dehmer M. Information theory and statistical learning. New York, NY: Springer Inc.; 2009. 365 p.
10. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC. Detecting novel associations in large data sets. Science. 2011. 334(6062):1518-1524.