Sensors: Predicting the future with simple convenience signals

Sean J. Taylor
Facebook
sjt@fb.com

Alexander Peysakhovich
Facebook
alexpeys@fb.com

Abstract

Forecasting which products, restaurants, or websites will become popular is difficult. We introduce a general method for finding individuals whose behaviors act as leading indicators for the future popularity of a product, URL, place, or application. We call these individuals "sensors" and their behaviors "signals." Importantly, signals generated by sensors need not be elicited with any particular forecasting problem in mind in order to be helpful.

We formulate the inclusion of sensors signals as a regularized regression problem which admits a simple and scalable estimation technique that finds a candidate set of sensors and incorporates their signals into any forecasting model. As an example, we apply our technique to identify sensors for the popularity of Facebook pages for musicians. Our method does not rely on causal effect of word-of-mouth: sensors can be useful either because they are individuals who are ahead of the taste curve (i.e. early adopters) or because they generate trends themselves (i.e. are influencers). Our next steps include a detailed comparison across more data sets, analysis of the demographics and local network structure of sensors, and applying qualitative methods to understand why and how they contribute to forecast accuracy.

Keywords social learning, big data, machine learning, forecasting

1. Introduction

The world is full of forecasting problems. These can include many important questions we face daily: who is the next Taylor Swift? What will crime rates be next year? Will a YouTube video go viral? Will my favorite restaurant start being crowded on Friday nights?

In this paper we develop a technique that can be used to find individuals whose actions are reliable predictors of future aggregate behavior. Our approach finds individuals such that when functions of their early adoption events (eg. watching a particular video) are added to an existing model, the out-of-sample forecasts of a collection of target time series becomes more accurate. We call these individuals "sensors" and their early adoption events "signals."

Many forecasting methods rely on the availability of leading indicators or borrowing information across related time series to learn from common trends. The sensors technique we introduce here provides potentially abundant leading indicators "for free" when the individual level behaviors which make up some aggregate are available – as is the case in most settings where individual-level event data are available.

2. Method

2.1 Simple Model

Suppose that we are modeling a collection of time series for items $i \in [1, I]$ at times t, with random values $Y_i(t)$. We claim this sequence of future values can be predicted by individuals $j \in [1, J]$ who can emit signals $k \in [1, K]$ about i at time T_{ijk}.[1] For example, we may use "liking" the page, "unliking" the page, and mentioning the page in a comment as our three signals. We can use our current set of available signals as features which help predict our collection of time series beyond some arbitrary baseline forecasting model $f_i(t)$:

$$Y_i(t) = f_i(t) + \sum_{j=1}^{J} \sum_{k=1}^{K} \beta_{jk} \cdot g_k(t, T_{ijk}).$$

2.2 Integrating signals from sensors

Functions g_k map sensor signal events into time-varying features for the model. For instance, we have explored unit-step function $(1(t - T_{ijk} > 0))$, the unit-ramp function $(t \cdot 1(t - T_{ijk} > 0))$, the rectangular function $(1(t-T_{ijk} > 0) \cdot 1(t-T_{ijk} < h))$, and logistic curves. The choice of function g depends on one's objective. Each has a slightly different interpretation but they may all be used if they are sufficiently uncorrelated.

In practice we use weakly positive functions for g_k, so the weight parameters β_{jk} can be interpreted as the contribution of the individual j's signal k to the future value of the time series. Large absolute values of β_{jk} are associated with the most important signals from the most important sensors.

Example: Suppose $Y_i(t)$ is the number of likes page i has a time t. Letting the signal $k = 1$ be "Liking a page" and g_1 be a unit step function amounts to allowing sensors to change the long run expected fans of a page, so if $\beta_{j1} = 5$ for some j, then our model can be interpreted: "if j likes a page i at t then at t you should now adjust your forecast upward by 5 likes."

2.3 Regularization

To prevent overfitting, we regularize the model by minimizing the squared in sample errors plus a penalty function:

$$C(\beta) = \lambda_1 (\sum_j \sum_k |\beta_{jk}|) + \lambda_2 \sum_j \sum_k (\beta_{jk}^2).$$

Individuals whose signal parameters, β_{jk}, remain different from zero (either positively or negatively) are useful sensors. We interpret the two penalty components as follows. First, the L_1 penalty λ_1 corresponds to our prior belief that not all individuals should be sensors and will lead to many candidate sensors having zero coefficients. If their penalized score is zero, we may safely ignore their signals, allowing us to pay attention to a smaller set of individuals. The L_2 penalty corresponds to our prior belief that, due to sampling variation, sensors may have gotten lucky or unlucky and that their true parameters are closer to zero than our estimate.

[1] Without loss of generality, we assume that $T_{ijk} = \infty$ if individual j has not yet adopted i.

Future time series	Signals
Page likes	likes, unlikes, mentions
Popular restaurants	check-ins, mentions
Trending Topics	use of n-grams in status updates
URL shares	shares, clicks, likes

Table 1. Potential practical applications of the sensors technique at Facebook.

3. Discussion

3.1 Sensors are a convenience prediction market

One possible interpretation of the signals emitted by sensors is that they are bets on the future values of a collection of future time series. The payoff function for signal k is then determined by function g_k. For instance, assume an individual likes a page when it has 500 likes and it has 5000 likes a week later. If we use the unit-step function for g_k, then that individual would have earned about 4500 likes (corresponding to a "buy and hold" trading strategy). If others liked the page around the same time, our regression approach would have essentially distributed credit for the increase in likes among them.

Thus, instead of recruiting skilled forecasters to actively participate in a prediction market and provide monetary incentives for them to reveal their information, we find a set of individuals who behave *as if* they are skilled bettors through their normal behaviors. The result is potentially vastly larger participation by using convenience signals instead of directly eliciting some future value of interest. Under this metaphor, sensors are bettors with consistently good track records.

3.2 Estimation is fast and simple

The sensors model can be easily estimated by leveraging efficient stochastic gradient descent methods [1] to fit all model parameters. Estimation is scalable to extremely large numbers of items and individuals for two reasons. First, the ability to use stochastic gradient descent means we we do not require all historical data to fit into memory. Third, if many signals have been emitted prior to some observation (i, t), then feature hashing [2] can be used to ensure the sparse signal vector is sufficiently small.

3.3 Data for sensors are abundant

The sensors technique can be applied in a large variety of empirical settings (c.f. Table 1). We can view any set of time-varying aggregate statistics for any population of items as potentially predictable using many individual-level behaviors toward those items. The assumption crucial to employing the sensors technique is that we observe a large set of individuals engage in repeated behaviors toward some set of related items. Sensors relies on sufficient density of individuals emitting signals to effectively learn which of the individuals are reliable leading indicators and which are not. We leave characterizing this density to future work, but applying the prediction market metaphor – we want to find sensors who have earned sufficient returns on their signals over a long enough period.

4. Preliminary empirical evaluation

We now turn to demonstrating a simple version of the sensor forecasting method. We look for sensors that can help forecast the future popularity of Facebook Pages. To do this, we use an event log of all *like* events[2] for a sample ($N = 49,858$) of of Facebook Pages for musicial artists which were created between 2011 and 2014 and have between 100 and 100,000 fans as of November 2014.

[2] A *like* event means that an individual declares herself a fan of the page and has subscribed to its status updates.

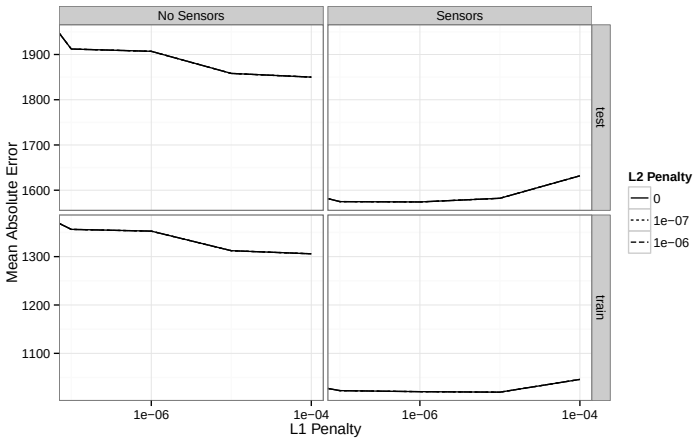

Figure 1. Out of sample error is minimized using sensors and positive L_1 penalty. The method is not sensitive L_2 penalty for this forecasting problem.

All data we use is completely aggregated and anonymized. The sensors technique does not require us to retain any personal information about individuals, or even their past signals. After we have estimated sensor-level parameters, all other information can be safely ignored during forecasting.

In this paper we discuss a very basic forecasting model which assumes that popularity growth for each page can be captured by a linear/quadratic trend:[3]

$$f_i(t) = \alpha_{i0} + \alpha_{i1}t + \alpha_{i2}t^2$$

This baseline model is for illustration only. The technique we describe here allows sensors to be incorporated into arbitrary forecasting techniques through a boosting procedure.

To look for sensors, we take the 44,267 individuals who have Liked more than 30 Facebook Music Pages in the sample period.[4]

We divide the data set into a training period (pre 2013) and a 30-day test period (January, 2013) in order to evaluate out of sample performance. Figure 1 shows the test set RMSE of the baseline model with and without sensors for various hyperparameters – the technique substantially improves forecast accuracy. Figure 2 shows how the number of sensors chosen varies by L_1 penalty.

5. Future work

In the full paper, we extend this work in three main directions:

5.1 More detailed evaluation

First, we are interested in characterizing the effectiveness of sensors across more forecasting problems. We apply the sensors technique (in combination with more or less sophisticated baseline forecasts, f_i) to the use-cases in Table 1 and evaluate out of sample errors for different time horizons. Our hypothesis is that sensors will be broadly useful across a range of applications.

[3] It is worth pointing out that our procedure finds sensors that improve on a specific "base" forecasting technique, for different models the optimal sensors may be different.

[4] Another way to think of this is that good sensors need always emit signals at some sufficient rate. On the other hand, sensors that signal too frequently will also be less useful due to a "boy who cried wolf" effect.

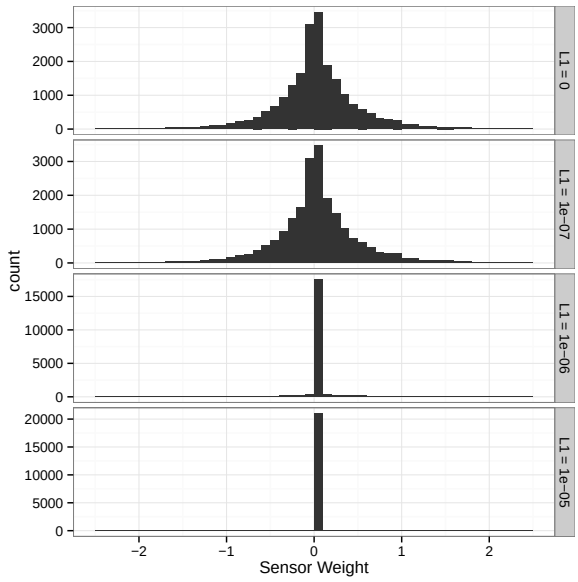

Figure 2. As we add L_1 penalty, the number of sensors decreases. The best performance on the test set was achieved with a small, but positive penalty.

5.2 Refinements to sensors

At the moment we are interested in two main refinements. First, sensors rests on selecting good functions g_k. In our exploration, we have found that rectangular functions work well (and help bound the number of signals per unit time), but they require selection of a bandwidth prior to parameter fitting. What is the optimal bandwidth for effectively using sensors and does this affect which sensors are selected? It may be the case that some signals are better for forecasting the near future while others are better for longer horizons.

Second, we are interested how levels of aggregation for outcomes of interest will affect sensor performance. For instance, we could forecast the total number of page likes for music pages using all Facebook users, or only for a specific demographic (say 18-25 year old males) using only candidate sensors from that group. Our hypothesis is that different sensors will work better for different levels of aggregation. E.g. some people's behaviors provide better "global" signals while others are better leading indicators of their particular demographic group.

5.3 How do sensors work? Are they causal?

We are very interested in understanding whether sensors work because the individuals identified as sensors have good (or bad) taste or because they are influential. We conduct a analysis of their demographics, online behaviors, and local social network structure. In addition, we are planning to survey and interview a small number of individuals identified as sensors in order to gain a better understanding of why they may be able to forecast the future. Without an experiment, we will be unable to say for sure how sensors work but we believe this exploratory work will provide some suggestive evidence.

Acknowledgments

We thank Eytan Bakshy, Adrien Friggeri, Ta Chiraphadhanakul, John Myles White and other members of the Facebook Core Data Science team for valuable discussions.

References

[1] A. Agarwal, O. Chapelle, M. Dudík, and J. Langford. A reliable effective terascale linear learning system. *CoRR*, abs/1110.4198, 2011. URL `http://arxiv.org/abs/1110.4198`.

[2] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1113–1120, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. . URL `http://doi.acm.org/10.1145/1553374.1553516`.