Teaching Note on a Framework for Prescriptive Analytics

Scott Webster[1]

October 23, 2019

Business analytics, in the main, focuses on improving decision making by leveraging data. Improvements are typically achieved through time compression (e.g., faster decisions via automation) and/or higher quality (e.g., decisions that lead to better performance via more effective use of data).

There is a long history and a wide literature on statistical methods for prediction. Similarly, much of the machine-learning literature has focused on supervised learning, i.e., predicting a value (or vector of values) as a function of observables (aka, predictors). For example, we might be interested in predicting the demands or demand probabilities of products. In this example, the observables (or predictors) may be data on product attributes (prices, inventories, design features, etc.), customers, competition, economy, weather, etc.

While the application of machine learning and statistical methods to large datasets for prediction (i.e., *predictive analytics*) is well established, methods to translate large datasets into prescriptions (i.e., *prescriptive analytics*) is emerging. Prescriptive analytics will likely become a very active area of analytics research and industry attention over the coming years.

The purpose of this teaching note is to introduce a new framework for prescriptive analytics. A reader with a reasonable knowledge of predictive analytics methods and with a basic understanding of optimization models/methods should be in a position to design and implement methods for prescriptive analytics. The content in this teaching note draws heavily on Bertsimas and Kallus (2020).

1. Preliminaries

A *decision* is to be made at the beginning of a period. The decision affects a *measure of performance* in the period, and may affect a *response* in the period. The response also affects the measure of performance. In addition, at the time of the decision there may be data on a *predictor* (in addition to the decision), or covariates, that affect the response. For example:

decision:	prices and replenishment order quantities of products the firm sells
predictor (beyond decision):	sales in previous periods, and indicator of whether the period is a weekday or weekend
response (unknown):	demand for each product in the period
performance measure:	profit in the period

[1] Department of Supply Chain Management, W. P. Carey School of Business, Arizona State University

In this example, product prices affect demand whereas replenish order quantities do not affect demand. We use the following notation for the above elements:

s = decision that affects the response, which may be a vector or empty, e.g., **s** = $\left(s_1,\ldots,s_{n_s}\right)$, $n_s \geq 0$

t = decision that doesn't affect the response, which may be a vector or empty, e.g., **t** = $\left(t_1,\ldots,t_{n_t}\right)$, $n_t \geq 0$

z = (**s**, **t**) = decision

Z = set of possible decision values, e.g., the decision may be constrained

x = predictor, which may be a vector or empty, e.g., **x** = $\left(x_1,\ldots,x_{n_x}\right)$, $n_x \geq 0$

y = response, which may be a vector, e.g., **y** = $\left(y_1,\ldots,y_{n_y}\right)$, $n_y \geq 1$

$v(\mathbf{z}, \mathbf{y})$ = performance measure that is a function of the decision and response

We have historical data comprised of N observations that are denoted using a superscript, i.e., the set of observations is

$$S_N = \{(\mathbf{x}^1, \mathbf{y}^1, \mathbf{z}^1), \ldots, (\mathbf{x}^N, \mathbf{y}^N, \mathbf{z}^N)\}.$$

From these data, the historical values of the performance measure can be generated from the function v, i.e., the historical performance measures are

$$\{v(\mathbf{z}^1, \mathbf{y}^1), \ldots, v(\mathbf{z}^N, \mathbf{y}^N)\}.$$

The predictor is needed when considering performance in the future (because that response has not yet been observed). However, there may be some applications for which there is no predictor. In such applications, the response depends only on the decision.

Finally, we illustrate the performance measure function in our example above, but with some elaboration on terms. Suppose there are n products, the inventory of product j is the replenishment quantity t_j, and the replenishment cost per unit of product j is c_j. Recall that s_j is the price of product j in the example. Then the profit in period i (i.e., observation i) is

$$v\left(\mathbf{z}^i,\mathbf{y}^i\right) = v\left(\left(\mathbf{s}^i,\mathbf{t}^i\right),\mathbf{y}^i\right) = \sum_{j=1}^{n}\left(s_j^i \min\left\{t_j^i, y_j^i\right\} - c_j t_j^i\right).$$

In this example, the measure of performance is distinct from the response. In some settings, the response itself is the measure of performance (e.g., $v(\mathbf{z}, y) = y$).

2. Prescriptive Analytics Framework

We wish to identify a mapping from the predictor to a prescription using the historical data S_N. The objective is to identify a decision (i.e., prescription) that maximizes the measure of performance. This mapping is called a *predictive prescription* and is denoted $\hat{\mathbf{z}}_N\left(\mathbf{x}\right)$.

In order to help clarify the concept of a predictive prescription, we begin with the simple and (usually) unrealistic setting where we know the exact probability distribution of the response for any given **z** and **x**. For a given predictor **x**, an optimal mapping is the solution to the following problem:

$$\max_{\mathbf{z}\in\mathbf{Z}} \bar{v}\left(\mathbf{z}\,|\,\mathbf{x}\right).$$

where $\bar{v}\left(\mathbf{z}\,|\,\mathbf{x}\right)=E\left[v\left(\mathbf{z},\mathbf{y}\right)|\,\mathbf{x}\right]$. $E[\cdot]$ is the expectation operator that is present because **y** may be uncertain (but affected by the values of **z** and **x**). In other words, for any given **x** and **z** values, **y** may be a random variable, and $\bar{v}\left(\mathbf{z}\,|\,\mathbf{x}\right)$ is the expected (or average) value of the performance measure given predictor **x** and decision **z**.

The optimal mapping, or predictive prescription, can be expressed as

$$\mathbf{z}^{*}(\mathbf{x}) \in \arg\max_{\mathbf{z}\in\mathbf{Z}} \bar{v}\left(\mathbf{z}\,|\,\mathbf{x}\right). \tag{1}$$

The above says that, for any given predictor **x**, $\mathbf{z}^*(\mathbf{x})$ is given by any “argument” that maximizes the objective function $\bar{v}\left(\mathbf{z}\,|\,\mathbf{x}\right)$ subject to the constraint $\mathbf{z} \in \mathbf{Z}$. Note that while $\mathbf{z}^*(\mathbf{x})$ is a predictive prescription (i.e., it maps a predictor into a prescription), the problem defined in (1) would not be described as prescriptive analytics. The reason is that there are no data, i.e., we are not translating a dataset into a prescription.

2.1. Framework

In most real-life settings, the response is uncertain and we don’t know the probability distribution of the response; we need to identify an effective predictive prescription from the data. The problem of finding a prescription from the data (i.e., *prescriptive analytics problem*) can be generically expressed as

$$\max_{\mathbf{z}\in\mathbf{Z}} \hat{v}_N\left(\mathbf{z}\,|\,\mathbf{x}\right) \tag{2}$$

where $\hat{v}_N\left(\mathbf{z}\,|\,\mathbf{x}\right)=\sum_{i=1}^{N} w_N\left(\mathbf{s},\mathbf{x},i\right)v\left(\mathbf{z},\mathbf{y}^i\right)$ is the objective function. The predictive prescription is

$$\hat{\mathbf{z}}_N\left(\mathbf{x}\right) \in \arg\max_{\mathbf{z}\in\mathbf{Z}} \hat{v}\left(\mathbf{z}\,|\,\mathbf{x}\right).$$

The term $w_N(\mathbf{s}, \mathbf{x}, i)$ represents the “weight” (or probability) of performance given decision **s** (recall that **s** is the subset of decision **z** that affects the response), predictor **x**, and observation *i*. The function $w_N(\mathbf{s}, \mathbf{x}, i)$ is estimated from historical data. Importantly, the function is estimated with the goal of *minimizing prescription error*, i.e., maximizing the quality of the prescription (Section 3 describes a way to measure prescription quality).

2.2. Estimating w_N

There are many statistical and machine-learning methods that can be used to compute the function $w_N(\mathbf{s}, \mathbf{x}, i)$. Linear and logistic regression fall within the statistical category. In the following sections we briefly illustrate how $w_N(\mathbf{s}, \mathbf{x}, i)$ can be computed using methods that fall within the machine-learning category. We do not go into the details of how the methods work (e.g., see James et al. 2015 for explanations).

2.2.1. *k*-nearest-neighbor

$$w_N^{k\text{NN}}\left(\mathbf{s},\mathbf{x},i\right) = \frac{I\left(\left(\mathbf{s}^i,\mathbf{x}^i\right) \text{ is the } k\text{NN of } \left(\mathbf{s},\mathbf{x}\right)\right)}{k}$$

where $I(\cdot)$ is an indicator function that returns a value of 1 if the condition is satisfied, and 0 otherwise. We see that the function includes parameter k, the value of which may be tuned using machine-learning methods.

The above expression reflects the following idea: if the point $(\mathbf{s}^i, \mathbf{x}^i)$ is one of the k closest points to $(\mathbf{s}, \mathbf{x})$, then $\mathbf{y}^i$ is likely to be reasonably close to the response at $(\mathbf{s}, \mathbf{x})$. For example, suppose there are $N = 10^{10}$ observations, $k = 3$, and for a particular point $(\mathbf{s}', \mathbf{x}')$, the 3 “close” observations are number 15, 722, and 1.23×10^9. The objective function value at decision $\mathbf{z}' = (\mathbf{s}', \mathbf{t}')$ and predictor $\mathbf{x}'$ is

$$\hat{v}_N^{k\text{NN}}\left(\mathbf{z}' \mid \mathbf{x}'\right) = \frac{1}{3}\left[v\left(\mathbf{z}',\mathbf{y}^{15}\right) + v\left(\mathbf{z}',\mathbf{y}^{722}\right) + v\left(\mathbf{z}',\mathbf{y}^{1.23\times 10^9}\right)\right].$$

2.2.2. Kernel methods

$$w_N^{\mathrm{K}}\left(\mathbf{s},\mathbf{x},i\right) = \frac{K\left(\left(\left(\mathbf{s}^i,\mathbf{x}^i\right) - \left(\mathbf{s},\mathbf{x}\right)\right)/h_N\right)}{\sum_{j=1}^{N} K\left(\left(\left(\mathbf{s}^j,\mathbf{x}^j\right) - \left(\mathbf{s},\mathbf{x}\right)\right)/h_N\right)}$$

where K is a kernel function and h_N is the bandwidth parameter, the value of which may depend on the number of observations, or more generally, dataset S_N. The value of h_N may be tuned using machine-learning methods. Kernel functions can be used to measure the “size” or magnitude of a vector, and thus can be used identify observations that are “close” to a point $(\mathbf{s}, \mathbf{x})$. Here are a few examples of kernel functions that map a vector $\mathbf{u} = (u_1, \ldots, u_n)$ into a scalar:

$K(\mathbf{u}) = I\left(\|\mathbf{u}\| \le 1\right)$ (uniform, aka, naïve)

$K(\mathbf{u}) = \left(1 - \|\mathbf{u}\|^2\right) I\left(\|\mathbf{u}\| \le 1\right)$ (Epanechnikov)

$K(\mathbf{u}) = \left(1 - \|\mathbf{u}\|^3\right)^3 I\left(\|\mathbf{u}\| \le 1\right)$ (tricubic)

$K(\mathbf{u}) = \dfrac{1}{2\sqrt{\pi}} e^{-\|\mathbf{u}\|_2^2/2}$ (Gaussian)

where $\|\mathbf{u}\| = \left(\sum_{i=1}^{n} u_i^2\right)^{1/2}$. As an example, consider the uniform kernel function applied to the prescriptive analytics problem:

$$K\left(\left(\left(\mathbf{s}^i,\mathbf{x}^i\right)-\left(\mathbf{s},\mathbf{x}\right)\right)/\,h_N\right)=I\left(\left\|\left(\left(\mathbf{s}^i,\mathbf{x}^i\right)-\left(\mathbf{s},\mathbf{x}\right)\right)/\,h_N\right\|\le 1\right).$$

For the uniform kernel, all neighbors of ($\mathbf{s}$, $\mathbf{x}$) in the set {($\mathbf{s}^1$, $\mathbf{x}^1$), …, ($\mathbf{s}^N$, $\mathbf{x}^N$)} that are within bandwidth h_N are weighted equally in the computation of the objective function

$$\hat{v}_N^{\mathrm{K}}\left(\mathbf{z}\,|\,\mathbf{x}\right)=\sum_{i=1}^{N}w_N^{\mathrm{K}}\left(\mathbf{s},\mathbf{x},i\right)v\left(\mathbf{z},\mathbf{y}^i\right).$$

2.2.3. Tree-based methods

$$w_N^{\mathrm{T}}\left(\mathbf{s},\mathbf{x},i\right)=\frac{I\left(R\left(\mathbf{s}^i,\mathbf{x}^i\right)=R\left(\mathbf{s},\mathbf{x}\right)\right)}{\sum_{j=1}^{N}I\left(R\left(\mathbf{s}^j,\mathbf{x}^j\right)=R\left(\mathbf{s},\mathbf{x}\right)\right)}$$

where R is a function that maps a vector into a terminal node (or leaf) of the tree. We see that all points in {($\mathbf{s}^1$, $\mathbf{x}^1$), …, ($\mathbf{s}^N$, $\mathbf{x}^N$)} that result in the same terminal node as ($\mathbf{s}$, $\mathbf{x}$) are weighted equally.

2.2.4. Ensembles, e.g., random forest

$$w_N^{\mathrm{RF}}\left(\mathbf{s},\mathbf{x},i\right)=\frac{1}{T}\sum_{k=1}^{T}\frac{I\left(R^k\left(\mathbf{s}^i,\mathbf{x}^i\right)=R^k\left(\mathbf{s},\mathbf{x}\right)\right)}{\sum_{j=1}^{N}I\left(R^k\left(\mathbf{s}^j,\mathbf{x}^j\right)=R^k\left(\mathbf{s},\mathbf{x}\right)\right)}$$

where T is the number of trees and R^k is the function for tree $k = 1, \ldots, T$.

3. Coefficient of Prescriptiveness

The R^2 of a predictive model is a measure of the model's accuracy or explanatory power. It measures the fraction of the variance of the uncertain response that is reduced (or explained) by a prediction based on observables.

For example, suppose model m_N that maps ($\mathbf{s}$, $\mathbf{x}$) into response y is estimated using dataset S_N = {($\mathbf{x}^1$, y^1, $\mathbf{z}^1$) , …, ($\mathbf{x}^N$, y^N, $\mathbf{z}^N$)}. We may wish to measure the predictive quality of model m_N using a validation dataset $\overline{S}_{N_v}=\left\{\left(\overline{\mathbf{x}}^1,\overline{y}^1,\overline{\mathbf{z}}^1\right),\ldots,\left(\overline{\mathbf{x}}^{N_v},\overline{y}^{N_v},\overline{\mathbf{z}}^{N_v}\right)\right\}$ of N_v observations that is separate from S_N. The coefficient of determination of model m_N on dataset $\overline{S}_{N_v}$ is

$$R^2=1-\frac{\sum_{i=1}^{N_v}\left(\overline{y}^i-m_N\left(\overline{\mathbf{s}}^i,\overline{\mathbf{x}}^i\right)\right)^2}{\sum_{i=1}^{N_v}\left(\overline{y}^i-\sum_{j=1}^{N_v}\overline{y}^j\,/\,N_v\right)^2},$$

e.g., R^2 is a measure of out-of-sample predictiveness of model m_N.

From the above expression, it is clear that R^2 can also be interpreted as the fraction of the way that the predictive model goes from the extreme of a naïve prediction model that is based on the average response,

$$\text{prediction} = \sum_{j=1}^{N_v} \bar{y}^j \,/\, N_v \,, \tag{3}$$

to the extreme of a perfect prediction model (i.e., model that maps $\left(\bar{\mathbf{s}}^i, \bar{\mathbf{x}}^i\right)$ into $\bar{y}^i$ for any *i*). In other words, it measures the relative predictive content of the data through model m_N.

Bertsimas and Kallus (2020) propose a coefficient of prescriptiveness, denoted *P*, that parallels this latter interpretation of the coefficient of determination. Suppose that $\hat{\mathbf{z}}_N(\mathbf{x})$ is a predictive prescription that is estimated using dataset S_N. Recall that $\bar{S}_{N_v}$ is a validation dataset that is separate from S_N. The value of *P* for $\hat{\mathbf{z}}_N$ on dataset $\bar{S}_{N_v}$ is computed using three terms:

$$\hat{v}_{N_v}(\hat{\mathbf{z}}_N) = \frac{1}{N_v}\sum_{i=1}^{N_v} v\left(\hat{\mathbf{z}}_N\left(\bar{\mathbf{x}}^i\right), \bar{\mathbf{y}}^i\right)$$

$$\hat{v}_{N_v}^* = \frac{1}{N_v}\sum_{i=1}^{N_v} \max_{\mathbf{z}\in\mathbf{Z}} v\left(\mathbf{z}, \bar{\mathbf{y}}^i\right)$$

$$\hat{v}_{N_v}^{\text{SAA}} = \frac{1}{N_v}\sum_{i=1}^{N_v} v\left(\hat{\mathbf{z}}_N^{\text{SAA}}, \bar{\mathbf{y}}^i\right)$$

where

$$\hat{\mathbf{z}}_N^{\text{SAA}} \in \arg\max_{\mathbf{z}\in\mathbf{Z}} \frac{1}{N}\sum_{i=1}^{N} v\left(\mathbf{z}, \mathbf{y}^i\right).$$

We see that $\hat{v}_{N_v}(\hat{\mathbf{z}}_N)$ is the average out-of-sample performance of predictive prescription $\hat{\mathbf{z}}_N$. The value of $\hat{v}_{N_v}^*$ represents the extreme of perfect out-of-sample performance. For each response $\bar{\mathbf{y}}^i$, the value of $\max_{\mathbf{z}\in\mathbf{Z}} v\left(\mathbf{z}, \bar{\mathbf{y}}^i\right)$ is the highest possible performance, e.g., best performance with a perfect prediction model. The value of $\hat{v}_{N_v}^{\text{SAA}}$ represents the other extreme. It is based on decision $\hat{\mathbf{z}}_N^{\text{SAA}}$ that maximizes average performance over all observed responses in the training dataset, and ignores the predictor. Decision $\hat{\mathbf{z}}_N^{\text{SAA}}$ is called the *sample average approximation* (SAA) prescription; it is analogous to a prediction based on the average of observed responses (as in (3)).[2]

The coefficient of prescriptiveness is

$$P = 1 - \frac{\hat{v}_{N_v}^* - \hat{v}_{N_v}(\hat{\mathbf{z}}_N)}{\hat{v}_{N_v}^* - \hat{v}_{N_v}^{\text{SAA}}}.$$

[2] If **y** is independent of **x** (i.e., **x** has no predictive content), then the SAA prescription is a reasonable choice because it appropriately ignores irrelevant data.

Similar to the spirit of the coefficient of determination, *P* measures the fraction of the way that the prescriptive model goes from the extreme of a naïve prescription model that is based on the average performance to the extreme of the best prescription under perfect prediction. In other words, it measures the relative prescriptive content of the data through model $\hat{\mathbf{z}}_N$ on validation dataset $\bar{S}_{N_v}$.

The value of *P* is a unitless measure of the prescriptive content, and thus like R^2, has the advantage of providing a universal measure of "model fit" across applications. However, the ranking of *P* values for alternative models on validation dataset $\bar{S}_{N_v}$ is determined by the ranking of $\hat{v}_{N_v}$. For example, suppose two predictive prescriptions $\hat{\mathbf{z}}_N^1$ and $\hat{\mathbf{z}}_N^2$ are derived from a training dataset S_N. Then

$$\hat{v}_{N_v}\left(\hat{\mathbf{z}}_N^1\right) = \frac{1}{N_v}\sum_{i=1}^{N_v} v\left(\hat{\mathbf{z}}_N^1\left(\bar{\mathbf{x}}^i\right), \bar{\mathbf{y}}^i\right) > \frac{1}{N_v}\sum_{i=1}^{N_v} v\left(\hat{\mathbf{z}}_N^2\left(\bar{\mathbf{x}}^i\right), \bar{\mathbf{y}}^i\right) = \hat{v}_{N_v}\left(\hat{\mathbf{z}}_N^2\right) \text{ implies } P_{\hat{\mathbf{z}}_N^1} > P_{\hat{\mathbf{z}}_N^2}$$

(because the values of $\hat{v}_{N_v}^{*}$ and $\hat{v}_{N_v}^{\mathrm{SAA}}$ are independent of the predictive prescription). We raise this point to clarify a risk that can arise when training a prescriptive model that does not arise when training a predictive model. In particular, if the decision $\mathbf{z}$ only affects the performance measure and does not affect the uncertain response (i.e., settings where $\mathbf{z} = \mathbf{t}$), then $v\left(\hat{\mathbf{z}}_N^1\left(\bar{\mathbf{x}}^i\right), \bar{\mathbf{y}}^i\right)$ is an accurate measure of performance of decision $\hat{\mathbf{z}}_N^1\left(\bar{\mathbf{x}}^i\right)$ applied to observation *i* in the validation dataset. Just as a measure of out-of-sample predictive error can be used to train a predictive model, the value of *P* can be used to train a prescriptive model (e.g., train function w_N that is used in (2)). However, the value of *P* as an out-of-sample measure of performance becomes less accurate in settings where the decision has a large effect on the uncertain response.[3] In such settings, promising predictive prescriptions identified via machine-learning methods may be further evaluated through pilot tests in the field (e.g., A/B testing).

4. Summary

Predictive analytics is the application of statistical methods and machine learning to data with the goal of minimizing out-of-sample prediction error. The output of predictive analytics is a predictive model, which may be used as an input to an optimization algorithm. This is a classical approach to prescriptive analytics, i.e., estimate a model with the objective of minimizing prediction error, then use the model in an optimization algorithm.

This teaching note outlines an alternative integrated approach that explicitly accounts for estimation error in the optimization step. It presents a framework for directly translating data into a prescription. It

[3] On the negative side, settings in which $\mathbf{z}$ affects $\mathbf{y}$ can introduce error when training a prescriptive model. But there is also a positive result. There is theory that shows that a predictive prescription obtained from the framework defined in (2) approaches the true optimal predictive prescription as the sample size *N* increases. This result holds when kernel methods are used to estimate function w_N and under several relatively mild assumptions (see Bertsimas and Kallus 2020).

explains the essence of how to apply statistical methods and machine learning to data with the goal of minimizing out-of-sample prescription error. This is an important, new, and growing research area. Every student of business analytics and data science and should be aware of this emerging area.

5. References

Bertsimas D, Kallus N (2020) From predictive to prescriptive analytics. *Management Science* to appear.

James G, Witten D, Hastie T, Tibshirani R (2015) *An Introduction to Statistical Learning with Applications in R* (Springer, New York).