DECISION TREES FOR HETEROGENEOUS DOSE-RESPONSE SIGNAL ANALYSIS

Kush R. Varshney, Moninder Singh, and Jun Wang

Business Analytics and Mathematical Sciences Department, IBM Thomas J. Watson Research Center
1101 Kitchawan Rd., Route 134, Yorktown Heights, NY 10598, USA

ABSTRACT

We propose a novel decision tree algorithm for modeling function-valued responses. This algorithm partitions the feature space into homogeneous subpopulations with common dose-response signals using a splitting criterion based on Nadaraya–Watson kernel regression and the Cramér–von Mises statistical test. We formulate an important business problem of sales team composition within the dose-response framework. Experimental results on generated and real-world sales data show the efficacy of the approach.

***Index Terms*— ** Business analytics, Cramér–von Mises test, customer relationship management, decision tree, kernel regression

1. INTRODUCTION

Dose-response signals show the average behavior of patients in response to varying amounts of treatment. Such signals arise in medicine and biology, but also in statistical economics and business analytics [1, 2]. For example, in examining pulmonary function in response to environmental exposure, respirable dust concentration is the dosage and volume of air exhaled in one second of forced expiration the response. In examining welfare-to-work programs, hours of job assistance services is the dosage and labor market earnings the response.

In a business analytics application that motivated our work, we were presented with the problem of determining the best sales team composition to approach a client opportunity. There are typically two types of sellers in sales organizations: client-facing sellers that focus on developing relationships with customers and technical sellers that provide details about products; sales teams can be composed of different ratios of client-facing sellers to technical sellers. By taking the proportion of the sales team that is client-facing as the dosage and the transactional revenue earned as the response, we can estimate the revenue as a function of sales team composition.

The treatment is continuous-valued in these examples and this is also the focus of the paper. Binary- or categorical-valued treatments lead to different problems and solutions. Also, here we focus on one-time responses; a temporal treatment or response component could be considered in future work. In the domain of focus, each patient gives us one sample including their treatment dosage, their response, and other features such as demographics or firmographics. The dose-response signal can be understood by looking at collections of patients with different treatment doses.

Often in medicine and other applications, the objective is to estimate the dose-response signal from data samples while attempting to eliminate any bias from variability in patient features [1, 2]. The idea is to understand the effect of the treatment independently from who the treatment is applied to, and the problem is essentially regression with bias correction. However, patient populations are heterogeneous [3] and different subpopulations may require different treatment doses.

Characterization of these different subpopulations is actionable information, especially in business analytics. (There may be legal and ethical barriers to providing unequal treatments to different subpopulations in public welfare applications.) Returning to our sales team example, it may be better to send a team with more technical sellers to customers in the industrial and government sectors and better to send a team with more client-facing sellers to customers in the transportation, distribution, and computer services sectors.

Patients are characterized by a host of features. Subpopulations can be defined by partitioning along any and any number of these features. One can manually 'slice and dice' the patient feature space to find subpopulations with homogeneous dose-response signals within the heterogeneous overall population, but this is intractable if there are even just a modest number of categorical patient features. In this paper, we propose a decision tree algorithm for partitioning the patient feature space into regions with a common dose-response signal.

Our proposed method is similar to classification and regression trees [4], but instead of having a discrete-valued scalar response (classification tree) or a continuous-valued scalar response (regression tree), here we have a function-valued response. As such, the criterion for splitting is not based on Gini impurity or information gain (classification), or on squared error (regression), but on a novel criterion for decision trees that we propose based on Nadaraya–Watson kernel regression and the Cramér–von Mises statistical test [5, 6, 7, 8]. Once a tree-based partition of the patient feature space has been learned, the dose-response signal for each of the finest-level partitions can be estimated.

Motivated by a real-world need, we are investigating a novel problem that has not been, to the best of our knowledge, considered in the statistical signal processing and statistical learning literature. On first sight, kernel regression trees may seem to be focused on the same problem, but they are not [9]. They solve the standard regression problem without the dose-response aspect. Although the problem we are considering, characterizing heterogeneous dose-response signals, does not appear in the literature, it is possible to compare our method to kernel regression and to regression trees in one respect: dose-response estimation accuracy on a per patient basis.

2. FEATURES, TREATMENT DOSES, AND RESPONSES

The dose-response characterization problem that we are considering arises in observational studies in natural environments in which patients are given treatments without any experiment in mind. For each patient $i = 1, \ldots, n$, we observe features $\mathbf{x}_i \in \mathcal{X}$, treatment dosage $t_i \in [0, 1]$, and response $y_i \in \mathbb{R}$. The patient feature space is generally multivariate; we refer to a particular feature dimension j as $\mathcal{X}_j$. The responses are assumed to be a smooth function of the treatment dosage with additive noise. A dose-response signal $y(t)$ is thus a

978-1-4673-0183-1/12/$31.00 ©2012 IEEE

function mapping $[0, 1] \to \mathbb{R}$. In a homogeneous population of patients in which all patients have the same underlying distribution, we would obtain an estimate $\hat{y}(t)$ of the dose-response signal from the pairs (t_i, y_i), $i = 1, \ldots, n$ via regression.

In the business analytics example, the patient is a client whereas the features include dimensions such as the sector to which the client belongs, the market segment to which it belongs, the type of selling relationship between the vendor and the client, and the geographic region in which the client is located. The task is to partition $\mathcal{X}$ into sets $\mathcal{R}_k$, $k = 1, \ldots, m$. The sets $\mathcal{R}_k$ are leaves of a binary decision tree constructed using splits along single feature dimensions $\mathcal{X}_j$. Associated with each leaf $\mathcal{R}_k$, we can also estimate a specific dose-response signal $\hat{y}_k(t)$ using the patients that fall in that leaf, i.e. $\{(t_i, y_i) | \mathbf{x}_i \in \mathcal{R}_k\}$.

3. KERNEL REGRESSION AND DENSITY ESTIMATION

The estimator we use in estimating dose-response signals from samples and also in calculating the split criterion to construct the decision tree is the Nadaraya–Watson estimator, a form of kernel regression [5, 6]. The functional form of the estimate from the samples $\{(t_1, y_1), \ldots, (t_n, y_n)\}$ is:

$$\hat{y}(t) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{t_i - t}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{t_i - t}{h}\right)}, \tag{1}$$

where $K(\cdot)$ is a kernel function, such as a Gaussian kernel or Epanechnikov kernel, and h is a bandwidth.

Bandwidth selection is a critical aspect of kernel regression and kernel density estimation to which an entire literature is devoted. In the remainder of the paper, we assume that an appropriate bandwidth has been selected. In the numerical examples we present in Sec. 5 and Sec. 6, we use the Gaussian kernel and the following rule of thumb to determine the bandwidth [10]:

$$h = \frac{\mathrm{med}(|t - \tilde{t}|)\,\mathrm{med}(|y - \tilde{y}|)}{0.6745^2} \sqrt[5]{\frac{16}{9n^2}}, \tag{2}$$

where $\tilde{t}$ is the median of $\{t_1, \ldots, t_n\}$, $\tilde{y}$ is the median of $\{y_1, \ldots, y_n\}$, $\mathrm{med}(|t - \tilde{t}|)$ is the median of $\{|t_1 - \tilde{t}|, \ldots, |t_n - \tilde{t}|\}$, and $\mathrm{med}(|y - \tilde{y}|)$ is the median of $\{|y_1 - \tilde{y}|, \ldots, |y_n - \tilde{y}|\}$.

Kernel density estimation, specifically Parzen window density estimation [11], is also used in calculating the split criterion proposed in Sec. 4, and is based on the same principle as the Nadaraya–Watson estimate. Given samples $\{t_1, \ldots, t_n\}$, the Parzen window density estimate is:

$$\hat{p}(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t_i - t}{h}\right). \tag{3}$$

4. DECISION TREE CHARACTERIZATION

In this section we propose a decision tree method to partition $\mathcal{X}$ into leaves $\mathcal{R}_k$ with homogeneous dose-response signals. Like classification and regression trees [4], the high-level algorithm proceeds by starting at a root node containing all samples $(\mathbf{x}_i, t_i, y_i)$. These samples are divided into two sets based on a split of one of the dimensions $\mathcal{X}_j$. If $\mathcal{X}_j$ is continuous or ordinal, then a split takes the form of a single threshold; $x_{i,j}$ less than or equal to the threshold go to one child and $x_{i,j}$ greater than the threshold go to the other child. If $\mathcal{X}_j$ is nominal with C different values, then there are $(2^{C-1} - 1)$ possible splits. For example, if $C = 3$ with $\mathcal{X}_j = \{\alpha, \beta, \gamma\}$, then the three possible splits are: α and β in one child and γ in the other, α and γ in one child and β in the other, and β and γ in one child and α in the other.

Among all possible splits in all of the feature dimensions, the split that is chosen is the one that maximizes a split criterion. If all possible splits have a split criterion value less than a parameter τ, then the node is not split and it is designated a leaf node. After the root node is split, both of its children are split, and so on, thus constructing a decision tree with leaf nodes $\mathcal{R}_k$, $k = 1, \ldots, m$.

The split criterion that we propose for a dose-response characterizing decision tree is based on nonparametric comparison of regression curves as follows [8]. Let us denote the two children of a particular split as A and B with samples $\{(t_1^A, y_1^A), \ldots, (t_{n_A}^A, y_{n_A}^A)\}$ and $\{(t_1^B, y_1^B), \ldots, (t_{n_B}^B, y_{n_B}^B)\}$ respectively, where $n = n_A + n_B$. Let us also denote the estimated density of the treatment dose samples in child A as $\hat{p}_A(t)$ and of the dose samples in B as $\hat{p}_B(t)$.

We base the split criterion on the residual between the A samples and the kernel regression of all samples at the node, and on the residual between the B samples and the kernel regression of all samples at the node. These residuals, normalized by kernel density estimates of the A and B samples are:

$$f_i^A = \frac{n(y_i^A - \hat{y}(t_i^A))}{n_A \hat{p}(t_i^A)} \tag{4}$$

$$f_i^B = \frac{n(y_i^B - \hat{y}(t_i^B))}{n_B \hat{p}(t_i^B)}. \tag{5}$$

If the A samples and B samples are from the same dose-response signal, then their pooled kernel regression should provide a good estimate for each set of samples and thus both sets of residuals should be small and similar on average. If they are from different dose-response signals, then the residuals should be large and dissimilar on average.

To measure the average similarity between the residuals, taking a cue from the two-sample Cramér–von Mises criterion [7], we construct the cumulative sum functions of the residual samples and calculate their squared L^2 distance. The cumulative sum functions are [8]:

$$F_A(t) = \frac{1}{n} \sum_{i=1}^{n_A} f_i^A \,\mathrm{step}(t - t_i^A) \tag{6}$$

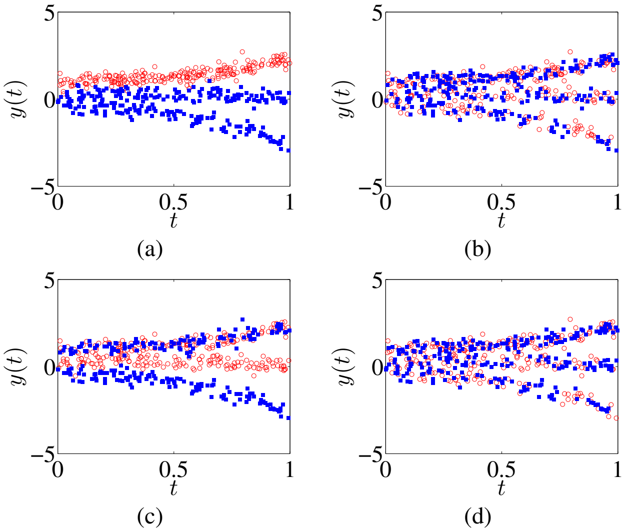
$$F_B(t) = \frac{1}{n} \sum_{i=1}^{n_B} f_i^B \,\mathrm{step}(t - t_i^B), \tag{7}$$

where $\mathrm{step}(\cdot)$ is a unit step function. The squared L^2 distance is the split criterion:

$$\text{split criterion} = \int_0^1 \left(F_A(t) - F_B(t)\right)^2 dt. \tag{8}$$

Thus at a node while constructing the dose-response decision tree, we calculate (8) for all possible splits. If there is at least one split criterion greater than τ, then the split with the maximum split criterion is chosen.[1]

[1] As we proceed deeper in constructing the tree, the number of samples per node decreases, resulting in greater regression variance; we are currently investigating the effect of this phenomenon.



(a) (b) (c) (d)

Fig. 1. One instantiation of data set 1 marked according to all possible splits at the root node with split criterion values (a) $\mathcal{X}_1$: 3.2411, (b) $\mathcal{X}_2$: 0.0137, (c) $\mathcal{X}_3$: 0.3968, (d) $\mathcal{X}_4$: 0.0134.

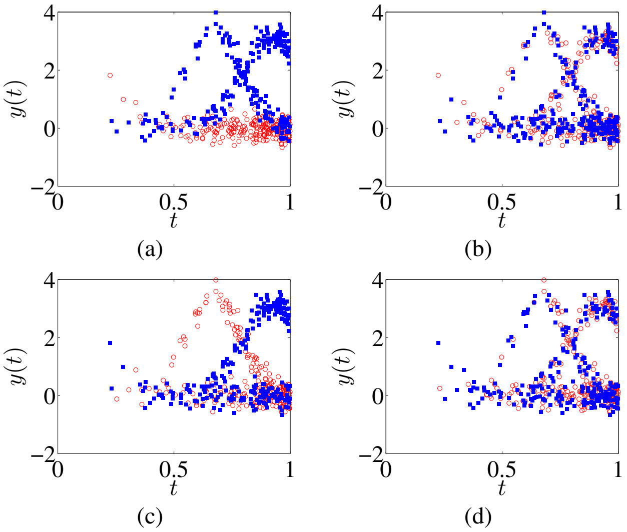
5. GENERATED DATA RESULTS

In this section, we present results showing the performance of the dose-response decision tree proposed in Sec. 4 and provide y prediction comparisons to baseline kernel regression and regression tree methods. We consider sample features, doses, and responses generated as follows. We construct two generated data sets. In both data sets, the patient feature space $\mathcal{X}$ has four dimensions, each of which is an equiprobable binary random variable. The treatment dose is sampled from a uniform distribution over $[0, 1]$ in the first data set and from the beta distribution with parameters $(4, 1)$ in the second.

In both data sets, there is a true tree structure to the dose-response signals with three leaves: $\mathcal{R}_1$ has $\mathcal{X}_1$ true, $\mathcal{R}_2$ has $\mathcal{X}_1$ false and $\mathcal{X}_3$ true, and $\mathcal{R}_3$ has $\mathcal{X}_1$ false and $\mathcal{X}_3$ false. The two dimensions $\mathcal{X}_2$ and $\mathcal{X}_4$ do not affect the dose-response signals. In the first data set, the dose-responses are fifth order polynomials with random coefficients: uniform in the range $[0, 1]$ for $\mathcal{R}_1$, range $[-0.5, 0.5]$ for $\mathcal{R}_2$, and range $[-1, 0]$ for $\mathcal{R}_3$. In the second data set, the dose-responses are Gaussian bumps with random standard deviations uniform in the range $[0.1, 0.15]$ and random means uniform in the range $[0, 0.25]$ for $\mathcal{R}_1$, $[0.5, 0.75]$ for $\mathcal{R}_2$, and $[0.75, 1]$ for $\mathcal{R}_3$. The sample response values have additive Gaussian noise with standard deviation σ. The two data sets are shown in Fig. 1 and Fig. 2.

We create 1000 different instantiations of the dose-response signals and take $n = 500$ samples. We also vary the magnitude of the additive response noise. We learn dose-response decision trees from the samples for each of the instantiations, each different noise deviation σ, and three different threshold values τ. The percentage of instantiations in which the true tree structure with three leaves is recovered is reported in Table 1 and Table 2 for the two data sets. The true tree is recovered a high percentage of the time. As would be expected, smaller threshold values have better performance when the noise variance is smaller.

As a comparison on the task of predicting y, we construct test sets of samples from the same distributions as the original data sets, also containing 500 samples each. We train on the original data sets and test on the test sets for all 1000 instantiations using the proposed dose-response decision tree with a separate kernel regression at each leaf, and also three other baseline methods: kernel regression on all of the (t_i, y_i) samples, regression tree on all of the (t_i, y_i) samples, and regression tree with both the t_i and $\mathbf{x}_i$ as the predictive variables and y_i as the response. The average root mean squared error (RMSE) results of the four methods are shown in Table 3. A threshold of $\tau = 0.1$ is used for both data sets. The RMSE of the dose-response tree method is best and is almost equal to the standard deviation of the noise, which is about as well as one can expect.



(a) (b) (c) (d)

Fig. 2. One instantiation of data set 2 marked according to all possible splits at the root node with split criterion values (a) $\mathcal{X}_1$: 0.3118, (b) $\mathcal{X}_2$: 0.0314, (c) $\mathcal{X}_3$: 0.0824, (d) $\mathcal{X}_4$: 0.0053.

	$\tau = 0.05$	$\tau = 0.10$	$\tau = 0.15$
$\sigma = 0.20$	92.3%	90.4%	88.0%
$\sigma = 0.25$	89.1%	90.3%	88.0%
$\sigma = 0.30$	82.4%	87.9%	87.5%
$\sigma = 0.35$	70.0%	85.3%	86.1%

Table 1. Percentage of data set 1 instantiations in which true tree is recovered for different amounts of noise and different thresholds.

6. SALESFORCE ANALYTICS DATA RESULTS

We also present results on the real-world salesforce analytics problem discussed in Sec. 1. Data from the customer relationship man-

	$\tau = 0.05$	$\tau = 0.10$	$\tau = 0.15$
$\sigma = 0.20$	92.6%	88.0%	83.6%
$\sigma = 0.25$	90.3%	86.7%	82.6%
$\sigma = 0.30$	87.1%	85.1%	80.8%
$\sigma = 0.35$	82.8%	83.4%	79.0%

Table 2. Percentage of data set 2 instantiations in which true tree is recovered for different amounts of noise and different thresholds.

Method	Data Set 1	Data Set 2
Kernel Reg.	1.1151	1.1481
Reg. Tree 1	1.3800	1.4045
Reg. Tree 2	0.3105	0.3320
DR Tree	0.2622	0.3023

Table 3. Average RMSE in predicting the y_i for $\sigma = 0.25$.

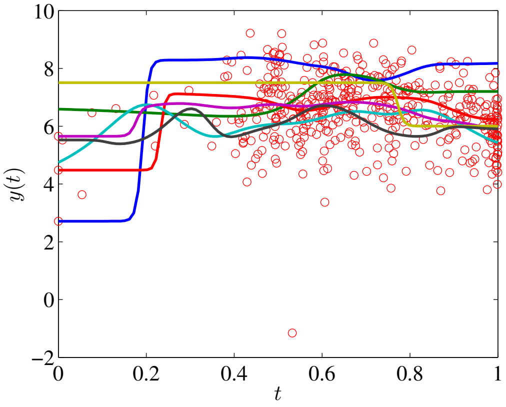


Fig. 3. Dose-response samples of sales team composition as well as estimated dose-response signal for each leaf of decision tree.

agement system of IBM is shown in Fig. 3, where as discussed earlier, t is the fraction of a sales team that is client-facing (as opposed to technical). The y is a particular transformation of transactional revenue earned by that team. We consider three feature dimensions: $\mathcal{X}_1$ takes six values and is the economic sector; $\mathcal{X}_2$ takes three values and is the coverage type of how clients are approached, such as the sales team having office space at the client location or coming to the client infrequently; $\mathcal{X}_3$ takes three values and is a market segmentation into core clients, clients with whom sales could grow, and clients whose business is taken opportunistically.

The dose-response decision tree that is learned from this data is shown in Fig. 4. The split at the root node divides opportunities by coverage type into those with the sales team sitting at the client location and others. The next split on the right branch is according to market segmentation with growth clients on the left. As can be seen in Fig. 3, there exist multiple homogeneous subpopulations that exhibit distinctive behavior in revenue attainment as a function of sales team composition. In particular, completely technical sales teams ($t = 0$) produce the lowest revenue in most but not all instances. These results have been confirmed by subject matter experts within IBM during 'deep-dive' investigations. For each of these subpopulations, we can identify the sales team composition that maximizes revenue by identifying peaks in the $\hat{y}_k(t)$.

7. CONCLUSION

We presented a novel heterogeneous dose-response signal analysis framework in this paper. Due to the existence of subpopulations for whom different treatment strategies should be applied, we develop a novel decision tree algorithm to partition the population based on dose-response characteristics. Finally, partitioning analysis and dose-response regression are conducted concurrently. Experimental results on synthetic and sales data show the effectiveness of the proposed method. Directions for future work include: addressing the issue of sample selection bias [1], designing methods to optimize the split parameter τ, and applying the method to other data domains.

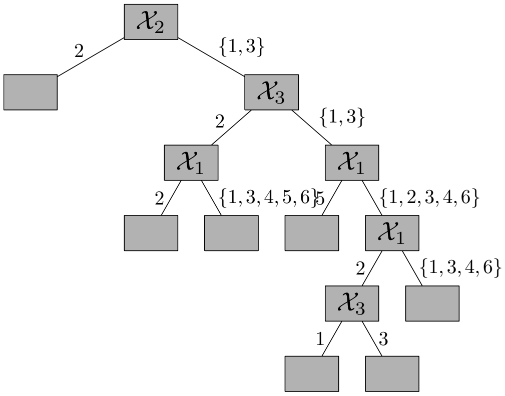


Fig. 4. Decision tree learned from sales team composition data.

8. REFERENCES

[1] K. Hirano and G. W. Imbens, "The propensity score with continuous treatments," in *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*. Chichester, UK: Wiley, 2005, pp. 73–84.

[2] C. A. Flores, "Estimation of dose-response functions and optimal doses with a continuous treatment," Dept. Econ., Univ. Miami, Coral Gables, FL, 2007.

[3] R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik, "Nonparametric tests for treatment effect heterogeneity," *Rev. Econ. Stat.*, vol. 90, no. 3, pp. 389–405, Aug. 2008.

[4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Boca Raton, FL: Chapman & Hall, 1984.

[5] E. A. Nadaraya, "On estimating regression," *Theory Probab. Appl.*, vol. 9, no. 1, pp. 141–142, 1964.

[6] G. S. Watson, "Smooth regression analysis," *Sankhyā Ser. A*, vol. 26, no. 4, pp. 359–372, Dec. 1964.

[7] T. W. Anderson, "On the distribution of the two-sample Cramér-von Mises criterion," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 1148–1159, Sep. 1962.

[8] N. Neumeyer and H. Dette, "Nonparametric comparison of regression curves: An empirical process approach," *Ann. Stat.*, vol. 31, no. 3, pp. 880–920, Jun. 2003.

[9] L. Torgo, "Functional models for regression tree leaves," in *Proc. Int. Conf. Machine Learn.*, Nashville, TN, Jul. 1997, pp. 385–393.

[10] A. W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford, UK: Oxford University Press, 1997.

[11] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Stat.*, vol. 33, no. 3, pp. 1065–1075, Sep. 1962.