Medumo - Identifying Patient Engagement Patterns

Zhonghao Guo, Zihao Yuan, Tianzhi Wu, Jialiang Shi
{gzh1994,yuan1z,tianzhi,markshi}@bu.edu

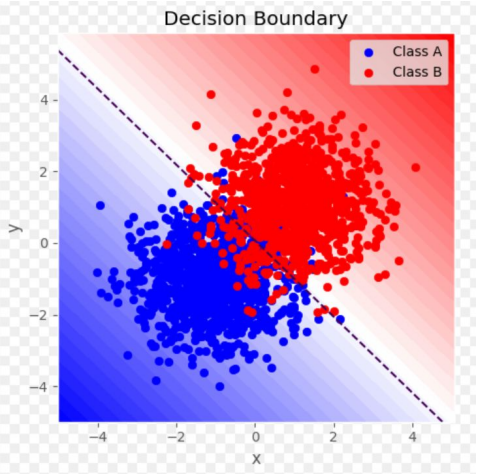

Figure 1. Binary Classification Model

1. Project Task

We would like to build a ML model to predict whether a customer will show up to an appointment or not. Additionally, which factors affect patients that don't show up and which combination of factors are most effective in predicting that a patient will not show up. This is critical to determining whether a patient's medical care will be successful or not.

Other Strategic Questions to be Answered:

1. Are demographic variables indicative of patients more/less likely to cancel appointments at the last minute?
2. What other factors are responsible for driving patient engagement? e.g. demographics (age, gender, time-related factors, engagement channels - mobile, web, email, etc., engagement events -notifications, web-clicks, etc.

2. Related Work

Machine learning has become a major approach for data analysis, pattern recognition, classification, and natural language processing (NLP). Various research efforts have been put into machine learning algorithm development and apply machine learning techniques to practical problems. Schuldt et al. apply support vector machines (SVM) to recognize human actions in videos through local space time features capturing [1]. Liaw et al. describe using random forest to do classification and regression [2]. [3] use neural network to classify subcellular structures in fluorescence microscope image. Krizhevsky et al. train Deep Convolutional Neural Network (DCNN) using ImageNet [5] dataset and achieve top-1 and top-5 error rates of 39.7% and 18.9%. In order to improve the performance of CNN, different hardware platform including FPGA, GPU, and AISC are being used to speed up CNN training and inference process. The authors in [4] build a stochastic computing based DCNN as a machine learning hardware accelerator to speed up Alexnet inference process. [6] implement Fast R-CNN in GPU to do Real-Time Object Detection. For a specific classification problem like identifying Patient engagement patterns, we will try different machine learning classification methods such as SVM, Regression, Neural Network etc., and pick the algorithm that result in best F-1 score. Following that, we will optimize the approach we select to further improve our model's performance.

3. Approach

The given dataset consists of m training samples. We denote them by $(x^{(i)}, y^{(i)})$, which $x^{(i)}$ is an input with n-dimensional features, and $y^{(i)}$ is the label of the datapoint, $y^{(i)} \in (0, 1)$. What we are asked to implement is a classification model, so that we plan to adopt supervised learning algorithms. Here are 4 typical algorithms:

(1) Logistic Regression

-Hypothesis:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

"y=1" if $h_\theta(x) \geq 0.5$, "y=0" if $h_\theta(x) \leq 0.5$

θ: learnable parameters

-Cost Function:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} Cost(h_\theta(x^{(i)}, y^{(i)})$$

$$= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)} + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$$

-Goal: Minimize cost $minJ(\theta)$

(2) SVM

SVM aim to find the hyperplane with the maximum margin to classify positive and negative datasets. Take the hard-margin SVM for example:

-Hyperplane model:

$$f(x) = \omega^T x + b$$

-Objective Function:

$$L(\omega, b, \alpha) = \frac{1}{2}||\omega||^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\omega^T x_i + b))$$

$\alpha = (\alpha_1; \alpha_2; ...; \alpha_m)$ is Lagrange multiplier

-Goal: Minimize L

(3) Decision Tree & Random Forest

A decision tree is a tree-like structure, where each non-leaf node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf node holds a class label. Algorithms for constructing decision trees: choosing a variable at each step that best splits the set of items. For example, the information gain is used by the ID3, C4.5 and C5.0 tree-generation algorithms.

-Entropy:

$$H(T) = I_E(p_1, p-2, ..., p_J) = -\sum_{i=1}^{J} p_i \log p_i$$

-Information Gain:

$$-\sum_{i=1}^{J} p_i \log p_i - \sum_{a} p(a) \sum_{i=1}^{J} -Pr(i|a) \log Pr(i|a)$$

(4) Neural Network

-Input: $x = [x_1, x_2, ..., x_m]^T$

-Output: $h_\theta(x) = g(\theta^{(i-1)} a)$, where g(z) is activation function and i is the # of layers.

-Cost function:

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log(h_\theta(x^{(i)}))_k + (1 - y_k^{(i)}) \log(1 - (h_\theta^{(i)})_k)]$$

$$+\frac{\lambda}{2m} \sum_{l=1}^{L-1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_l+1} (\theta_{ji}^{(l)})^2$$

4. Dataset and Metric

Dataset description

The engagement data is structured in three different tables Enrollment, Engagement Events and Outcomes. Enrollment Data has a list of patients that were enrolled for their colonoscopies(procedure) along with the date they were enrolled and their procedure date. It also has their date of birth and gender.

Engagement events is the engagement data that includes all the interaction events for every patient. It includes a lot of events and the meanings of each event type will be explained to each team in person.

Outcomes data contains the binary variable that we are trying to predict. For the scope of this project, cancellations within 3 days of colonoscopies and no show are combined and flagged as 1. The procedures that took place are flagged as 0. This puts the limitation that latest we can predict is at the end of 4th day before procedure date. Ideally we would want to predict as soon as possible but accuracy of the model increases as we try to predict closer to the procedure.

Dataset Size

The size of training set is 1995, and test set size is 221, which is relatively small.

Data preprocessing

Since there are 3 separate tables, the first thing we need to deal with is to represent the data into one uniformed format.

We want to use several data representation techniques to preprocess our data, which may involve handling missing data, combining multiple dataset, data standardizing/ normalization/ re-scaling, etc.

Evaluation of success

By trying out different models to classify whether a patients would show up or not, we hope to show that the best model can achieve accuracy, precision, and recall are all higher than the 70%.

5. Timeline and Roles

Task	Deadline	Lead
Implement Logistic Regression	11/05/2018	Tianzhi Wu
Implement SVM	11/18/2018	Zhonghao Guo
Implement Random Forest	11/30/2018	Jialiang Shi
Implement Neural Network	12/15/2018	Zihao Yuan
Report and Presentation	12/20/2018	all

References

1) Schuldt, Christian, Ivan Laptev, and Barbara Caputo. "Recognizing human actions: a local SVM approach." *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. Vol. 3. IEEE, 2004.
2) C. J. C. Burges. A tutorial on support vector machine for pattern recognition. Data Min. Knowl. Disc. 2 (1998) 121.
3) P. Xu, A. K. Chan. Support vector machine for multi-class signal classification with unbalanced samples. Proceedings of the International Joint Conference on Neural Networks 2003. Portland, pp.1116-1119, 2003.
4) Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." *R news* 2.3 (2002): 18-22.
5) Boland, Michael V., and Robert F. Murphy. "A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells." *Bioinformatics* 17.12 (2001): 1213-1223.