Feature Selection and Validation for Human Classifiers

Jay B. Martin, Eric Colson, and **Brad Klingenberg**
Stitch Fix, Inc.
{jmartin,ecolson,brad}@stitchfix.com

Abstract

Algorithmic approaches to prediction and recommendation can often be improved by combining the results with the curation of human experts. Hybrid machine-human recommendation systems can combine the best of both large-scale machine learning and expert-human judgement. In this paper, we outline an approach for measuring, training, and understanding the human contribution to the combined system. This approach provides a practical strategy for optimizing the role and experience of the human experts. We share a motivating example from Stitch Fix, an online personal styling service that commits to its recommendations through the physical delivery of merchandise to clients.

Introduction

At Stitch Fix, human computation enables our recommendation system to deal with the complexity and nuances of personalization that is not yet possible with even state-of-the-art algorithms (Colson 2013). Our system relies on a two-stage process: The first stage performs a series of machine learning (ML) computations to produce rank-ordered recommendations over our available inventory based on client attributes and purchase history. The second stage passes those rankings to in-house experts (i.e., stylists) who use their domain knowledge to select a subset of machine-recommended items to be shipped to our clients.

A key challenge in this process is assessing which client attributes (e.g., height, age, location, style), when processed by expert-humans, provide additional information beyond what has already been extracted by machines. That is, the machine algorithms use some of the same information in their processing. The goal of including additional processing by expert humans is to be able to extract incremental value though better contextualization. This process requires a validation system to assess the impact of different sets of client attributes on stylist judgment. In this article, we outline one such system that borrows insights from the cognitive science and ML literature.

Humans as Classifiers

Our recommendation system narrows down the available inventory and scores items based on how likely the client is to buy them. Our styling platform presents the scored results along with the client's profile, which includes attributes such as age, height, weight, location, style preferences, purchase history, etc. The stylists use this information to curate the recommendations and put together a cohesive shipment. We are now working to understand how stylists use this information to decide which items ultimately go into a shipment. For example, how much does seeing a picture of a client help in deciding which items should be sent to her? Does it help or hinder performance to know where the client lives, or if she normally wears bohemian style clothing?

Generally, some information might bias stylist behavior in unpredictable ways. For example, only seeing a client's age could bias stylists to send items they feel are age appropriate. However, new information made available via a client photo may reveal that the 50 year old client they are styling has a much younger aesthetic. On the other hand, a client photo could result in the stylist basing decisions on potentially irrelvent aspects like the appearence of the client's friend in the photo, or by drawing other invalid conclusions from the client's physical appearence.

We gain an understanding of these and other biases by studying how stylists are influenced by various pieces of information. We assume that stylists use a mental styling algorithm to classify each client's affinity for items. Casting stylist behavior as a classification problem, we can treat stylists' decisions as the output of a classifier. To make this connection clear, suppose each stylist from the set of stylists S uses a mental styling algorithm h_s. h_s is presumably a complicated, non-linear function of the attributes $\mathbf{x}$ presumably learned over a life time of experiences and observations. To keep things simple, we assume it is a linear function of $\mathbf{x}$. Moreover, let us also make the unlikely but convenient assumption that stylists are exchangeable, i.e., $h_s = h_t$ $s, t \in S$. The objective of h is to assess the likelihood that client j will buy item i given j's attributes $\mathbf{x_j}$. Buy/not buy is a binary random variable $Y_i \in \{0, 1\}$, and thus h can be expressed simply as $P(Y_i = 1 | X = x_j)$ which can be modeled with standard tools like logistic regression.

Recall that we have a machine-learned styling algorithm f that precomputes scores based on client attributes, and those scores are passed to the stylist algorithm h which then integrates those scores along with a similar set of client attributes to select which items go into a shipment. That is, the stylist considers a similar set of client attributes as the machines. Our proposed validation system aims to quantify the incremental value afforded by the use of expert-humans.

Framing the human decisions as the output of classification algorithms suggests that the application of standard ML validation techniques might be possible. In the next section, we will consider the styling problem from this perspective.

Validation

Treating stylists as classifiers opens up many possibilities for employing established ML methods to validate the impact of features and stylist abilities, as well as assessing the potential contribution of humans to our recommender system. For example, our approach uses an experimental apparatus and simple cross-validation to validate the performance of a classifier (stylist + attributes).

Recall the basic idea of cross-validation is to split a dataset into a test set $\mathcal{T}$ and a training/validation set $\mathcal{V}$. The validation set is used to train a model, whereas the test data establishes how well the model generalizes to new examples instead of simply recalling examples from memory.

There are two hurdles for applying this approach to humans: (1) We lack explicit knowledge of their mental model h and the data it was trained on, $\mathcal{V}$, and (2) Finding a test set $\mathcal{T}$ of gold standard reference labels could potentially be difficult. (1) is out of the scope of this article, but see Sanborn and Griffiths (2007) for possible approaches. For (2), we use our clients' historical data similar to standard cross-validation techniques. Note, there is some chance that the training set leaks into the test set, $\mathcal{T} \subset \mathcal{V}$ because a stylist may have been exposed to this historical data in the past, but for simplicity we will ignore this issue.

Assessing the Impact of Human Classifiers

Using historical data for the test set $\mathcal{T}$, we would like to assess the performance of our human classifier, h. Here, $\mathcal{T}$ is a set of tuples containing client attributes $\mathcal{C}$, item attributes $\mathcal{I}$, and outcome (e.g., buy/no buy) Y, $\mathcal{T} = \{(\mathcal{C}_j, \mathcal{I}_i, Y_{ij}) \mid j \in C, i \in I_j\}$, where C is the set of clients, I_j is the set of items sent to client j. In machine learning, we would normally apply some model to the $(\mathcal{C}, \mathcal{I})$ pairs in $\mathcal{T}$ to predict the outcome of $\hat{Y}_{ij}$ and then estimate the corresponding prediction error $\hat{L}(f(\mathcal{C}, \mathcal{I}))$. When studying human performance, however, we need a way to query h, so that we can estimate $\hat{L}(h(\mathcal{C}, \mathcal{I}))$. The remainder of the section outlines our proposed method for estimating $\hat{L}$ directly from human applications of h to $\mathcal{T}$.

Our method simulates the styling platform normally used by stylists to perform their tasks. Because it is a simulation, we can easily manipulate the information that is presented, allowing us to perform carefully controlled experiments on any existing client information such as age, location, style preferences, etc. Additionally, we can experiment with adding new information such as social network data or client photos.

We gauge performance by presenting stylists with two (or more) sets of client profiles: one baseline profile without the attribute under investigation ($P_{\setminus a} = \mathcal{C} \setminus a$, $a \in \mathcal{C}$), and another profile with the attribute ($P_a = \mathcal{C}$). Profiles are presented alongside an item i that client j may or may not have purchased. The stylists' task is to use the provided information to decide whether or not client j bought item i. After the stylists complete the experiment, it is straightforward to compare changes in performance with the attribute $\hat{L}(h(P_a, \mathcal{I}))$ and without it $\hat{L}(h(P_{\setminus a}, \mathcal{I}))$. By implementing a simple A/B test, for instance, we can randomly assign half of the stylists to receive P_a and the other half receive $P_{\setminus a}$. We can then compare stylists' performance across conditions to identify the effect of the attribute in question.

This method can also be used to compare stylist and machine performance by having the machine-algorithm and stylist make predictions on the same test set and then compare their performance using a metric like AUC. A significant boost in AUC would suggest that either the stylist or machine algorithm is performing better, whereas no result suggests little to no value added to the machine predictions.

Conclusion and Future Work

We have presented a basic framework for evaluating the predictive performance of human stylists in a system that combines statistical prediction and human curation. The ability to carefully measure the incremental performance gained by presenting different features to the stylist algorithm h is important to balancing the inherent strengths and weaknesses that come with human judgement. For example, humans generally excel at processing unstructured data, being empathetic with clients and taking a holistic view when selecting inventory. On the other hand, there are many challenges to removing bias from human judgment. Humans easily succumb to confirmation bias, availability bias, narrative fallacies, selective memories, and many other phenomenon. In addition, humans have a difficult time assessing multivariate tradeoffs and properly weighting the saliency of information (Stillings, Chase, and Feinstein 1995). Further, human learning is hindered when feedback is ambiguous or received only long-after a tasks has been completed.

By measuring the impact of changing the inputs to the stylist algorithm h we can systematically and iteratively measure the resulting improvement in human judgement. In future work, we will present further techniques to mitigate these limitation of human judgement in order to leverage human-computation in a way that is accretive to machine computation.

References

Colson, E. 2013. Using human and machine processing in recommendation systems. In *First AAAI Conference on Human Computation and Crowdsourcing*.

Sanborn, A., and Griffiths, T. L. 2007. Markov chain monte carlo with people. In *Advances in neural information processing systems*, 1265–1272.

Stillings, N. A.; Chase, C. H.; and Feinstein, M. H. 1995. *Cognitive science: An introduction*. MIT press.