Diagnostic Model Explanations: A Medical Narrative

Umang Bhatt, Brian Davis, & José M.F. Moura
Carnegie Mellon University
{umang, braindavis, moura}@cmu.edu

Abstract

Explainability techniques are important to understanding machine learning models used in decision critical settings. We explore how pattern recognition techniques ought to couple requisite transparency with predictive power. By leveraging medical data with the task of predicting the onset of sepsis, we expose the most important features for the model prediction. We uncover how important training points and consensus feature attributions vary over the learning process of these models. We then pose a counterfactual question to explore trained predictors in the medical domain.

Overview

As machine learning becomes pervasive, transparency and intelligibility of underlying machine learning models precedes adoption of these technologies (Doshi-Velez and Kim 2017; Lipton 2016). Recent machine learning interpretability techniques fall under either gradient-based methods that compute the gradient of the output with respect to the input, treating gradient flow as a saliency map (Shrikumar, Greenside, and Kundaje 2017; Sundararajan, Taly, and Yan 2017), or perturbation-based methods that approximate a complex model using a locally additive model, thus explaining the difference between test output-input pair and some reference output-input pair (Lundberg and Lee 2017; Ribeiro, Singh, and Guestrin 2016). While gradient-based techniques like (Shrikumar, Greenside, and Kundaje 2017) consider infinitesimal changes to the decision surface and then take the first-order term in the Taylor expansion as the additive explanatory model, perturbation-based additive models consider the difference between an input and reference vector. There have also been approaches that assign attributions to features with more complex counterfactuals than that discussed above for gradients (Kusner et al. 2017; Datta, Sen, and Zick 2016).

There is also a burgeoning line of research in using deep learning for medical diagnostics (Choi et al. 2016; Caruana et al. 2015; Rajkomar et al. 2018). Starting with tabular data (where each feature is semantically meaningful) from the medical domain (Purushotham et al. 2017), we fuse the aforementioned interpretability techniques into these nascent medical diagnostic models.

We note that medical professionals and doctors undergo rigorous training to learn how to determine the outcome for a given patient. This training mandates doctors learn how to proactively search for particular risk predictors upon seeing a patient; for example, a cardiologist learns to look at past patients to determine if a patient has a given valve disease. Doctors are not only trained to identify which attributes of a patients (vital signs, personal information, family history, etc.) deem the patient at risk for a particular disease or outcome, but also develop a intuition from past patients based on years of experience; for example, if a doctor treats a rare disease was seen over a decade ago, that patient can be extremely vital to a doctors diagnosis engine when attributes alone are uninformative about how to proceed. The doctor treats the patient from ten years ago as an **anchor** for future patients with similar symptoms. Over time, the doctor learns a larger set of diagnosis he or she feels comfortable with diagnosing: this growth of a doctor describes a powerful narrative that uncovers how a doctor reasons overtime.

Dataset

The MIMIC-III (Medical Information Mart for Intensive Care III) is a large electronic health record dataset compromised of health related data of over 40,000 patients who were admitted to the the critical care units of Beth Israel Deaconess Medical Center between the years 2001 and 2012 (Johnson et al. 2016). MIMIC-III consists of demographics, vital sign measurements, lab test results, medications, procedures, caregiver notes, imaging reports, and mortality of the ICU patients. Using MIMIC-III dataset, we extracted seventeen real-valued features deemed critical in the sepsis diagnosis task as per (Purushotham et al. 2018). These are the processed features we extracted for every sepsis diagnosis (a binary variable indicating the presence of sepsis): Glasgow Coma Scale, Systolic Blood Pressure, Heart Rate, Body Temperature, Pao2 / Fio2 ratio, Urine Output, Serum Urea Nitrogen Level, White Blood Cells Count, Serum Bicarbonate Level, Sodium Level, Potassium Level, Bilirubin Level, Age, Acquired Immunodeficiency Syndrome, Hematologic Malignancy, Metastatic Cancer, Admission Type.

Approach

Once we train a predictor, f, on the aforementioned dataset, we use the attribution aggregation approach, AVA, proposed

in (Bhatt, Ravikumar, and Moura 2019) to concurrently find influential patients and the features that were important to sepsis diagnoses in the test set. First let us introduce some notation. Let $x \in R^d$ be a datapoint's feature vector where the $x_i \in R$ is a specific feature of this datapoint. Let $\mathcal{D} = \{x^{(j)}\}_{j=1}^N$ represent the training datapoints, where $\mathcal{D} \in R^{d \times N}$ is the entire training set in matrix form with $\mathcal{D}_{i,j} = x_i^{(j)}$. Let f be the learned predictor we wish to explain. Using the tractable approximation derived in (Koh and Liang 2017), we define the influence weight, ρ_j of training point, $x^{(j)}$, on a test point, x_{test} as:

$$\rho_j = \mathcal{I}_{\text{up,loss}}(x^{(j)}, x_{\text{test}}) = \frac{d}{d\epsilon}\mathcal{L}(f_{\epsilon,x^{(j)}}, x_{\text{test}})\big|_{\epsilon=0}$$

Next, we find the feature attribution of x_{test} via an explanation function g. Suppose we let g be a Shapley Value attribution from classical game theory and from (Lundberg and Lee 2017), then we find that attribution of the i^{th} feature of point x is given by the Shapley value, which is the sum of the contributions to f for the i^{th} feature in all possible subsets S of the features F given by, where $R = \left(\frac{|S|!(|F|-|S|-1)!}{|F|!}\right)$:

$$g_i(x) = \sum_{S \subseteq F \setminus \{i\}} R(f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S))$$

If we let g be be gradient-based attribution from (Sundararajan, Taly, and Yan 2017), we find that attribution of the i^{th} feature of point x is given by the gradient of $f(x)$ along the i^{th} dimension of x with respect to a baseline $\bar{x}$.

$$g_i(x) = (x_i - \bar{x_i}) \int_{\alpha=0}^{1} \frac{\partial f(\bar{x} + \alpha(x - \bar{x}))}{\partial x_i} d\alpha$$

Using the aggregation methodology from AVA, we aggregate the attributions for past patients (training set) to explain model predictions for new patients (test set). For a given test set example, we provide an aggregate feature attribution to explain why the given prediction was made. Aggregate feature attribution can be found via the weighted variants of AVA from the original paper or via rank aggregation techniques like Borda Count and Markov Chain aggregation. We then provide the most important patient from the previously seen patients. We find this important patient, x_{imp}, as follows.

$$x_{imp} = \arg\max_{x^{(j)} \in \mathcal{D}} \rho_j$$

We are interested in how over time different training points become more or less influential to the retrained predictor. To be concrete, to make a prediction on Day 10, the doctor can use a patient she saw on Day 5. After "retraining" her internal predictor that day, she can now use the patients from Day 10 to explain and predict patients on subsequent days. The influential anchors in the training data change as a function of time; therefore, model explanations capture how different patients serve as anchors based on the exhaustiveness of the predictor's training set. We are also interested in understanding how the predictor deals with *unknown unknowns* that lie in an uncharted portion of the feature space. The predictor might not be confident about its predictions in a given region, but as more training data is added, the predictor may be able to learn about a particular region unbeknownst to it a few time steps ago. Note: we do NOT make any assumptions about the model class of f. We assume black box access to the predictor we wish to explain.

Experimentation

We first create different candidate feed-forward models to be explained and then train them on the aforementioned sepsis data set. We varied the depth of the models from 1 to 3 hidden layers, with ReLU or Sigmoid activations, trained with the ADAM optimizer and cross-entropy loss. To explain the model, we use SHAP, IG, and various proposed techniques ($\mathcal{W}_{\text{SHAP}}$ is the AVA weighted SHAP technique, $\mathcal{W}_{\text{IG}}$ is the AVA weighted IG technique, $\mathcal{M}_{\text{IG}}$ is Markovian aggregation with IG attribution, $\mathcal{M}_{\text{SHAP}}$ is Markovian aggregation with SHAP attribution, $\mathcal{B}_{\text{IG}}$ is Borda aggregation with IG attribution, $\mathcal{B}_{\text{SHAP}}$ is Borda aggregation with SHAP attribution). We report recall of a decision tree's *gold set* averaged over all the test instances of the sepsis dataset in Table 1, as done in (Ribeiro, Singh, and Guestrin 2016). For these experiments, we fix k to be five, use the mean values of the training input as the region of perturbation for SHAP, and use the aforementioned greedy technique to determine m to be five. Note random attribution will recall m/d, where d is the total number of features; for these experiments, random attribution will have a recall of 13%.

m-Sensitivity

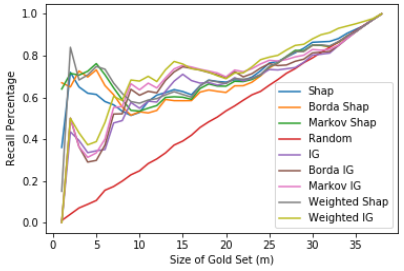
We also run experiments where we analyze how gold set recall changes as a function of m, the size of a gold set. If $m = d$, then all attribution techniques, including random, will have 100% recall. In Figure 1, we see that all attributions (other than random) recall a high percentage of the important features. As such for all following experiments we set m to 5.

Expectation over Explanations

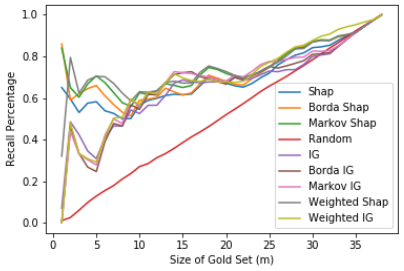
We cannot declare the absolute feature attribution for any arbitrary test point. We therefore aim to see how our methods perform in expectation by iterating 1000 times to find a probability distribution over the rank of the explanation algorithms. We used gold set recall to rank every method on every iteration, keep the ordinal position of each method, and iterate. For every iteration, we sample 100 points at random with replacement: we find explanations for those points using every single method in question. After 1000 iterations, we can then say with 55% probability weighted aggregation with SHAP attribution yields the best explanation (in the first position) and with 44% probability Markovian aggregation with SHAP attribution yields the best explanation. Interestingly, Markovian aggregation with SHAP attribution appears in the top two positions 99.5% of the time, while weighted aggregation with SHAP attribution appears only 94.3% of the time. A graph of the distribution for each method can be found in Figure 2.

MODEL	ACCURACY	SHAP	IG	$\mathcal{W}_{\text{SHAP}}$	$\mathcal{W}_{\text{IG}}$	$\mathcal{B}_{\text{SHAP}}$	$\mathcal{B}_{\text{IG}}$	$\mathcal{M}_{\text{SHAP}}$	$\mathcal{M}_{\text{IG}}$
1-SIGMOID	85.3	60	29	**68**	37	65	26	67	31
1-RELU	82.8	62	33	**69**	47	65	37	**69**	38
2-SIGMOID	86.7	61	34	75	41	73	75	**76**	40
2-RELU	87.2	55	35	**64**	35	60	30	62	33
3-SIGMOID	83	64	31	68	41	67	29	**71**	31
3-RELU	87	55	38	**65**	48	57	44	64	43

Table 1: Gold set recall on important features from an interpretable classifier to explain models trained on the sepsis dataset



(a)



(b)

Figure 1: m-Sensitivity for the sepsis dataset for different models trained with the ADAM optimizer (a) recall for 2 Hidden Layer Sigmoid model (b) recall for 2 Hidden Layer ReLU model

For every iteration of every method, we can also keep track of the position of each feature. This gives us a probability distribution of rankings for each feature, which gives us better insight into how important features are in expectation. In Figure 3, we find that sofa and sepsis_cdc appear in the top position of importance among all explanations most often: this is expected because both are highly correlated with the onset of sepsis. Simultaneously, as a sanity check, we find the race (e.g. hispanic) does not matter (appears at a lower rank) in expectation for all explanations; therefore, the top model learns not to correlate race and sepsis.

Counterfactual Intuition

It is instructive to consider the counterfactual entailed in temporal explanations. Feature attribution techniques like (Sundararajan, Taly, and Yan 2017) calculate attribution by finding the partial derivative of the output with respect to every input feature (Ancona et al. 2018). One perspective of

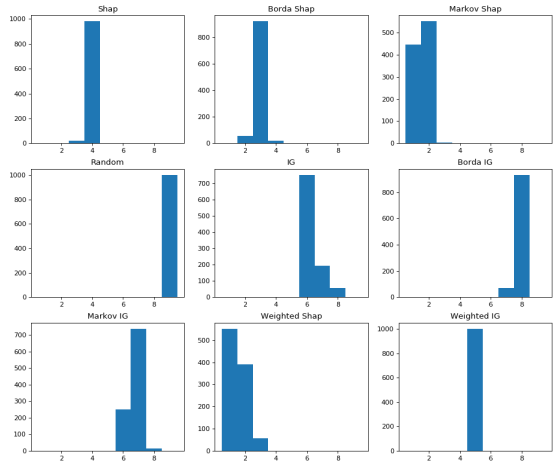


Figure 2: Representative Distribution of Methods Ranks

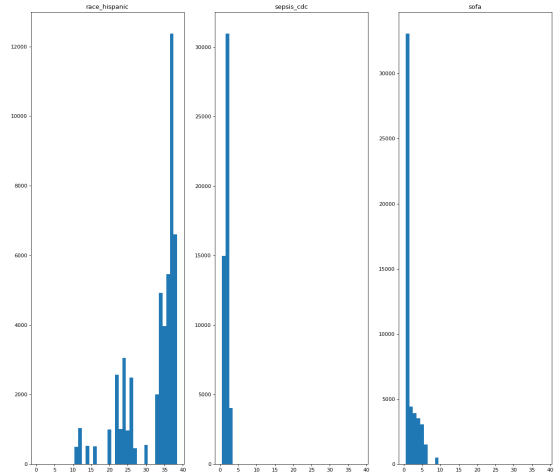


Figure 3: Feature rank distribution for race_hispanic, sepsis_cdc, and sofa (left to right)

this is as a counterfactual of how perturbing the j-th input infinitesimally would perturb the learnt predictor f. Indeed, such counterfactual intuition allows humans to intuit about the impact of a cause by having the baseline be the absence of the cause: from here, humans can tell the importance of a cause by seeing how the output changes in the causes' absence. Influence functions from (Koh and Liang 2017) consider the counterfactual of how upweighting a training data point x infinitesimally will affect the loss at a test point x_{test}.

The counterfactual posed by temporal explanations is as follows: what training point (past patients) perturbations which when used to train the predictor (doctor) would influence the test prediction (current patient) the most. Suppose we perturb a training point (past patient) as $z_\delta = z + \delta$, and we denote the predictor obtained by upweighting substituting the training data point x by x_δ and moreover upweighting this by some constant ϵ as $\widehat{f}_{\epsilon,x_\delta,-x}$. Then the counterfactual mentioned above at a test point x_{test}, would compute:

$$\nabla_\delta \frac{d}{d\epsilon} \mathcal{L}(f_{\epsilon,x_\delta,-x}, x_{\text{test}})\big|_{\epsilon=0}\big|_{\delta=0}.$$

We "freeze" the predictor function, since we only have black box access to the model, and ask: given influential training points, what would be the change to the frozen and trained predictor as we perturb those training points. This counterfactual allows us to create explanations that capture global patterns in the local neighborhood of the test point: this allows users to better audit the global trends of a model whilst still having the fidelity of local explanations. Essentially, we can explore what would happen if a doctor had seen a different patient at time step $t-1$ who would have expanded the doctor (that is, the predictor's understanding of the feature space), would the doctor have made a different prediction at time step t? Such an understanding would not only debug the predictors learned from real data but also ensure diagnostic models align with doctor intuition.

References

Ancona, M.; Ceolini, E.; Oztireli, C.; and Gross, M. 2018. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*.

Bhatt, U.; Ravikumar, P.; and Moura, J. M. F. 2019. Towards aggregating weighted feature attributions. *In the AAAI 2019 Workshop on Network Interpretability* abs/1901.10040.

Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; and Elhadad, N. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, 1721–1730.

Choi, E.; Bahadori, M. T.; Schuetz, A.; Stewart, W. F.; and Sun, J. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In Doshi-Velez, F.; Fackler, J.; Kale, D.; Wallace, B.; and Wiens, J., eds., *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56 of *Proceedings of Machine Learning Research*, 301–318. Children's Hospital LA, Los Angeles, CA, USA: PMLR.

Datta, A.; Sen, S.; and Zick, Y. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, 598–617.

Doshi-Velez, F., and Kim, B. 2017. Towards a rigorous science of interpretable machine learning.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; Mark, R. G.; and et al. 2016. Mimic-iii, a freely accessible critical care database.

Koh, P. W., and Liang, P. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 1885–1894. International Convention Centre, Sydney, Australia: PMLR.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc. 4066–4076.

Lipton, Z. C. 2016. The mythos of model interpretability.

Lundberg, S. M., and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*. 4765–4774.

Purushotham, S.; Meng, C.; Che, Z.; and Liu, Y. 2017. Benchmark of deep learning models on large healthcare mimic datasets.

Purushotham, S.; Meng, C.; Che, Z.; and Liu, Y. 2018. Benchmarking deep learning models on large healthcare datasets. *Journal of Biomedical Informatics* 83:112–134.

Rajkomar, A.; Oren, E.; Chen, K.; Dai, A. M.; Hajaj, N.; Liu, P. J.; Liu, X.; Sun, M.; Sundberg, P.; Yee, H.; Zhang, K.; Duggan, G. E.; Flores, G.; Hardt, M.; Irvine, J.; Le, Q. V.; Litsch, K.; Marcus, J.; Mossin, A.; Tansuwan, J.; Wang, D.; Wexler, J.; Wilson, J.; Ludwig, D.; Volchenboum, S. L.; Chou, K.; Pearson, M.; Madabushi, S.; Shah, N. H.; Butte, A. J.; Howell, M.; Cui, C.; Corrado, G.; and Dean, J. 2018. Scalable and accurate deep learning for electronic health records. *CoRR* abs/1801.07860.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. *CoRR* abs/1704.02685.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 3319–3328. International Convention Centre, Sydney, Australia: PMLR.