A Model Explanation System

Ryan Turner
Northrop Grumman Corporation
ryan.turner@ngc.com

We propose a general model explanation system (MES) for "explaining" the output of black box classifiers. In this introduction we use the motivating example of a classifier trained to detect fraud in a credit card transaction history. The key aspect is that we provide explanations applicable to a *single prediction*, rather than provide an interpretable set of parameters. The labels in the provided examples are usually negative. Hence, we focus on explaining positive predictions (alerts).

In many classification applications, but especially in fraud detection, there is an expectation of false positives. Alerts are given to a human *analyst* before any further action is taken. Analysts often insist on understanding "why" there was an alert, since an opaque alert makes it difficult for them to proceed. Analogous scenarios occur in computer vision [10], credit risk [8], spam detection [6], etc.

Furthermore, the MES framework is useful for model criticism. In the world of generative models, practitioners often generate synthetic data from a trained model to get an idea of "what the model is doing" [5]. Our MES framework augments such tools. As an added benefit, MES is applicable to completely non-probabilistic black boxes that only provide hard labels. In Section 3 we use MES to visualize the decisions of a face recognition system.

Fraud detection example A simple example explanation is: "Today, there were two in person transactions in the USA, followed by \$1700 in country X." MES would output "$(x_i \geq 2) \wedge (x_j \geq 1700)$" for the appropriate features i and j. The former could be generated via a NLG module [14].

Explanation vs. interpretability We assume the paradigm where prediction accuracy is of paramount importance, but explanation is also important. Therefore, we are not willing to give up any predictive accuracy for explanation. There is a long history of building models that are "interpretable" [1; 20]; such as, (small) decision trees [13] and sparse linear models [19]. MES augments black box predictions with explanations, as the best prediction system may not be "interpretable."

Historically, this dilemma has led to two approaches: 1) the "interpretable" models approach, common in scientific discovery/bioinformatics [15], and 2) the accuracy-focused approach, common in computer vision with methods like deep learning, k-NNs [2], and SVMs [16]. The downside of the interpretable approach is seen in machine learning competitions, where the winning methods are typically nonparametric, or have a very large number of parameters (e.g., deep learning) [4].

MES has elements of both approaches. We do not aim to succinctly summarize how the model "works in general," but only seek explanations of individual cases. Although the distinction is subtle, explanation is a much easier task than explaining the entire model. MES utilizes this weaker requirement to augment black box models with explanations without affecting accuracy.

1 Formal setup

Consider a black box binary classifier f that takes a feature vector $\mathbf{x} \in \mathcal{X} = \mathbb{R}^D$ and provides a binary label: $f \in \mathcal{X} \to \{0, 1\}$. In the examples above, explanations are Boolean statements about the feature vector. In effect, an explanation E is a function from $\mathcal{X}$ to $\{0, 1\}$. The mapping $E^* \in \mathcal{X} \to \mathcal{E}$ finds the best explanation from the set of possible explanations $\mathcal{E} \subset \mathcal{X} \to \{0, 1\}$. In Section 1.1 we describe a set of desiderata on E^*. We define that: 1) $\mathcal{E}$ contains a "null explanation" $E_0(\mathbf{x}) := 1$; 2) the set $\mathcal{F} := \{\mathbf{x} \in \mathcal{X} | f(\mathbf{x}) = 1\}$; 3) an explanation E is true at $\mathbf{x}$ when $E(\mathbf{x}) = 1$.

Dependence on input distribution Discriminative models, even when probabilistic, do not learn the input feature distribution, which we refer to as p [11]. However, they can be augmented by learning an input feature distribution "on the side." An interesting twist of MES is that it *does* depend on the input distribution even when explaining discriminative models.

1.1 Desired properties

Before describing our desiderata on E^*, we first define some key terms: **Eligibility**: An explanation is eligible if it increases the probability that the classifier alerts: Explanation $E \in \mathcal{E}$ is eligible if $P(f(\mathbf{x}) = 1|E(\mathbf{x}) = 1) \geq P(f(\mathbf{x}) = 1)$, where $\mathbf{x} \sim p$; and we refer to all eligible explanations as $\mathcal{E}' := \{E \in \mathcal{E}|E \text{ is eligible}\}$. Note that f and E are deterministic functions of the input $\mathbf{x}$; we are *marginalizing* over the inputs $p(\mathbf{x})$. **Generality**: An explanation's generality G is the probability of it being true (among $\mathbf{x}$ in $\mathcal{F}$): $G(E) := P(E(\mathbf{x}) = 1|f(\mathbf{x}) = 1)$. **Accuracy**: An explanation's accuracy A is the probability the classifier alerts given the explanation is true: $A(E) := P(f(\mathbf{x}) = 1|E(\mathbf{x}) = 1)$. **Validity**: An explanation E is *valid* at $\mathbf{x}$ if it is eligible and true at $\mathbf{x}$ ($E(\mathbf{x}) = 1$).

To summarize, eligibility is a property of E that does not depend on $\mathbf{x}$, although it does depend on the marginal $p(\mathbf{x})$. Whether an explanation E is true/valid requires knowledge of $\mathbf{x}$. In a classification context, accuracy A and generality G are analogous to precision and recall, respectively.

We contend that a sensible $E^*(\mathbf{x})$ mapping returns the most preferable explanation that is valid at $\mathbf{x}$:

$$E^*(\mathbf{x}) \in \max_{E \in \mathcal{E}'} E \quad \text{such that} \quad E(\mathbf{x}) = 1\,, \tag{1}$$

where the $\max$ is a preference relation max. We now list the desired properties of the preference relation $(\mathcal{E}, \precsim)$: 1) The set $\mathcal{E}$ is a preference relation (totality and transitivity). 2) If two valid at $\mathbf{x}$ explanations have equal accuracy then the one with more generality is preferred: If $G(E_1) > G(E_2)$ and $A(E_1) = A(E_2)$ then $E_2 \prec E_1$. 3) If two valid at $\mathbf{x}$ explanations have equal generality then the one with more accuracy is preferred: If $A(E_1) > A(E_2)$ and $G(E_1) = G(E_2)$ then $E_2 \prec E_1$. Requiring generality is useful since useless explanations like $x_i \in [a, a + \epsilon]$ often have accuracy 1.

This gives two derived properties: 1) Any valid at $\mathbf{x}$ explanation E not independent of f is preferable to the null explanation E_0: If $E \neq E_0$ is valid at $\mathbf{x}$ and $E \not\perp f$, then $E^*(\mathbf{x}) \neq E_0$. 2) If the decision rule f is in $\mathcal{E}$, then it is preferable to any other explanation: If $f \in \mathcal{E}$ then $E^*(\mathbf{x}) = f$ for all $\mathbf{x} \in \mathcal{F}$.

1.2 Explanation with a scoring function

A clear way to setup the explanation problem is to use a scoring function:

$$E^*(\mathbf{x}) \in \operatorname*{argmax}_{E \in \mathcal{E}'} S(A(E), G(E)) \quad \text{such that} \quad E(\mathbf{x}) = 1\,, \quad \text{where} \quad \frac{\partial S}{\partial A} > 0\,, \quad \frac{\partial S}{\partial G} > 0\,. \tag{2}$$

Any E^* defined via (2) obeys (1) and the desiderata on $(\mathcal{E}, \precsim)$. The converse is also true assuming a continuity condition on $\mathcal{E}$ [12, p. 104]; this justifies the scoring paradigm (2). Note that (2) only cares about the relationship between E and f; we do *not* care if $E(x)$ directly predicts the data.

We describe three clear examples for S that obey (2) and are normalized on $[0, 1]$ for eligible explanations. Being normalized, regardless of the constant background rate $P(f)$, means the score S is suitable as an explanation *quality score*. We get similar results for the three scores: mutual information $S = \text{MI}(E; f)/\text{H}[f]$, correlation $S = \text{Corr}\,[E, f]$, and covariance $S = \text{Cov}\,[E, f]\,/\text{Var}\,[f]$.

Comparison with decision trees Although it may appear we are merely "reinventing decision trees," there are key differences between MES and a decision tree. At a high level, MES is explaining why a decision was made on a single input, while decision trees aim to make the entire model easy to understand. More precisely, although we use simple decision functions in $\mathcal{E}$, E^* may be arbitrarily complex. By contrast, simple decision trees cannot match any black box classifier exactly. If the decision tree is not the top performing model, we must sacrifice performance for interpretability. MES augments any model with an explanation, alleviating any need to sacrifice performance.

2 Linear classifier example

Before moving on to general black boxes, we first demonstrate MES when the input density $p(\mathbf{x})$ is Gaussian and the decision boundary is linear, which includes logistic regression, perceptrons, and

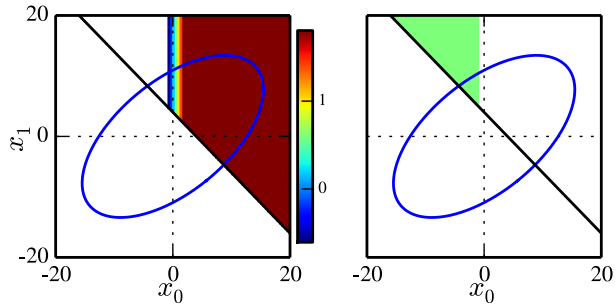

Figure 1: Linear Gaussian example of MES in 2D ($\mathbf{x} \in \mathbb{R}^2$). For illustration, we plot explanations for all inputs $\mathbf{x} \in \mathcal{F}$. The black line is the decision boundary of f. The blue ellipse represents the Gaussian $p(\mathbf{x})$. **Left:** Explanations of the $E(\mathbf{x}) = \mathbb{I}\{x_0 \geq a\}$ form where the heatmap color represents a. **Right:** Same as left except that $E(\mathbf{x}) = \mathbb{I}\{x_1 \geq a\}$. When $x_0 < -1$ the explanation is $E(\mathbf{x}) = \mathbb{I}\{x_1 \geq 0.5\}$ and for $x_0 \geq 2$ it is $E(\mathbf{x}) = \mathbb{I}\{x_0 \geq 2\}$. For $x_0 \in [-1, 2]$ it is $E(\mathbf{x}) = \mathbb{I}\{x_0 \geq a\}$ with $a = x_0$.

linear SVMs. We use axis aligned thresholds (i.e., half spaces) as the explanations. In equations:

$$p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \,, \quad f(\mathbf{x}) = \mathbb{I}\{\mathbf{w}^\top \mathbf{x} + b \geq 0\} \,, \quad \mathcal{E} = \textstyle\bigcup_{i=1}^{D} \left\{ \mathbb{I}\{x_i \lesseqgtr a\} \,, \forall a \in \mathbb{R} \right\} . \tag{3}$$

Next, we solve (2) for each $i \in 1{:}D$ and use the axis with the best optimum. We use the transform

$$\tilde{\mathbf{x}} := [\mathbf{u}_i \quad \mathbf{w}]^\top \mathbf{x} + [-a \quad b]^\top \in \mathbb{R}^2 \,, \tag{4}$$

where $\mathbf{u}_i$ represents a 1-of-k encoding. Notice that now $P(E, f) = P(\text{sign}(\tilde{x}_1), \text{sign}(\tilde{x}_2))$. Since $\tilde{\mathbf{x}}$ is distributed by a bivariate Gaussian, $P(E, f)$ is computed with one bivariate normal CDF and two univariate normal CDF calls. After computing $P(E, f)$ evaluating the score S is straightforward.

Efficient precomputation When using mutual information as the score function S, the score function is convex w.r.t. the threshold a in the linear Gaussian case. Therefore, on each axis i we precompute the optimal threshold a_i via bisection search. Later, when finding an explanation of the form $\mathbb{E}(\mathbf{x}) = \mathbb{I}\{x_i \leq a\}$, we use $a = \max(x_i, a_i)$; we use a similar $\min$ operation for $\mathbb{I}\{x_i \geq a\}$. We then compare the scores for the explanation on each axis and report the axis (and corresponding threshold a) with the highest score. This fast approach scales linearly in dimension D.

Illustrative example In order to visualize every possible input we present a linear example of MES in a 2D feature space (x_0 and x_1). We use mutual information as the score S in Fig. 1. The reader can verify visually that MES finds explanations that have high accuracy, generality, or both.

3 Black box models

In this section we use simple Monte Carlo (MC) methods to extend Section 2 to black box models. We merely require the classifier f be queryable at an arbitrary input $\mathbf{x}$ and that we can obtain samples from the input density $p(\mathbf{x})$. We retain the axis aligned explanations from (3) for $\mathcal{E}$.

Although we can optimize a MC estimate of any scoring function, we use a scoring function equivalent to the *covariance* in this section. This allows us to provide worst-case guarantees about the closeness of the MC explanation to the true optimal explanation. In this section we use:

$$S(E) = P(E|f = 1) - P(E|f = 0) = G(E) \left(1 - (P(f)^{-1} - 1)^{-1}(A(E)^{-1} - 1)\right) . \tag{5}$$

The reader can verify that (5) satisfies the requirements in (2) for eligible explanations ($P(f) < A(E)$). First, consider explanations of the form $E(\mathbf{x}) = \mathbb{I}\{x_i \leq a\}$. Then, we compute $S(E)$ using

$$P(x_i \leq a|f = 1) - P(x_i \leq a|f = 0) = \text{CDF}_{x_i}(a|f = 1) - \text{CDF}_{x_i}(a|f = 0) =: F(a) - H(a) \,.$$

We refer to $S(E)$ as the Kolmogorov score since $\max_a |S(E)|$ is known as the *Kolmogorov distance* between $p(x_i|f = 1)$ and $p(x_i|f = 0)$. Utilizing the law of total probability, one can show that $S(E) \propto \text{Cov}\,[E, f]$ in the case of binary variables (E and f) and constant $P(f)$. Additionally, using Bayes' rule and the law of total probability, one can show that E is eligible iff $S(E) \geq 0$.

We approximate (5) using $S_n(a) := F_n(a) - H_n(a)$, where F_n and H_n are empirical CDFs of F and H from n MC samples. Sampling from $p(x_i|f = 1)$ and $p(x_i|f = 0)$ is made efficient by sampling from $p(\mathbf{x})$ and rejection sampling with f. Next, we provide an upper confidence bound (UCB) on the suboptimality e of the approximate explanation threshold $\hat{a}$ vs the exact threshold a^*:

$$e := S(a^*) - S(\hat{a}) \geq 0, \quad a^* \in \text{argmax}_a S(a), \quad \hat{a} \in \text{argmax}_a S_n(a) \,. \tag{6}$$

Figure 2: **Far Left:** A modified face where the classifier correctly predicts Chavez. **Left:** The mean removed input face leading to a prediction of Bush (gray=0, white>0, black<0). **Right:** Eigenface 5, which MES selects for the explanation. **Far Right:** The Hadamard product of the input face and eigenface 5. MES explanation: This eigenface product has net white balance $\geq$1%. Think of the white areas as being the parts of the image that contribute to the SVM predicting Bush, and the dark areas as though the Bush prediction is made in spite of them. Eigenface 5 sees dark eyebrows, shading above the eyes/under the nose, and a dark open mouth.

Utilizing $S_n(a^*) - S_n(\hat{a}) \leq 0$ and then adding/canceling $(S_n(a^*) + S_n(\hat{a}))$ we bound (6):

$$e \leq (S(a^*) - S_n(a^*)) + (S_n(\hat{a}) - S(\hat{a})) \leq 2 \max_a |S_n(a) - S(a)| \,. \tag{7}$$

Then utilizing $\max_a |S_n(a) - S(a)| \leq \max_a |F_n(a) - F(a)| + \max_a |H_n(a) - H(a)|$, the union bound, and the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality [3; 9] we get that:

$$P(e \geq \epsilon) \leq 4 \exp(-2n(\epsilon/4)^2) = 4 \exp(-n\epsilon^2/8) =: \delta_0 \,. \tag{8}$$

This bounds the suboptimality when searching on a single axis. When searching D axes in two directions ($\leq$ and $\geq$ explanations) we use a Bonferroni correction for $\delta_0 = \delta/(2D)$. This allows us to reuse MC samples for searches on each axis. Therefore, to obtain a score suboptimality ϵ with confidence δ we need $n = \lceil 8 \log(8D/\delta)/\epsilon^2 \rceil$ MC samples.

During implementation, we precompute the optimal thresholds a as in the linear Gaussian case, but instead of doing a bisection search we merely store the *cumulative maximum* of S_n. During test, we find the optimal explanation on an axis by evaluating the precomputed cumulative maximum. Then, as in the linear case, we compare the best scores found on each axis.

Face recognition example We now demonstrate the black box MES implementation on the sci-kit learn demo "Faces recognition example using eigenfaces and SVMs" [7; 17]. The faces are reduced to dimension $D = 150$ from $50 \times 37 = 1{,}850$ using PCA [18]. Then 966 training examples are plugged into a (RBF kernel) multi-class SVM for classifying the faces as one of seven political figures. When explaining a classification of face k (Bush) we convert the SVM to a binary black box, informally as $f(\mathbf{x}) = \mathbb{I}\{\text{SVM}(\mathbf{x}) = k\}$. We use $\epsilon = 0.025$ and $\delta = 0.05$ implying $n = 129{,}099$. Induced from the assumptions of PCA, we use a standard multivariate Gaussian for the input density $p(\mathbf{x})$.

In Fig. 2 MES explains why the SVM classifies Hugo Chavez as George W Bush. In the far right image we see the classifier is utilizing the black arc under his teeth and the dark area around his right eye. In most training photos of Bush he has an open (dark) mouth and a lot of shading above his eyes. When we subtract the (normalized) eigenface from the original image, the classifier correctly predicts the face as Chavez. The corrected face (far left) has a lighter right eye and whiter smile.

4 Conclusions

We have presented a general framework for explaining black box models. In doing so, we have made clear the subtle distinction between interpretable models and augmentation through explanation. The framework alleviates the tension between performance and interpretability in suitable use cases.

First, we proposed and implemented an efficient algorithm using bisection search for linear classifiers. This algorithm works well in high dimensions as its cost scales linearly in input dimension D. We then demonstrated a MC algorithm (with accuracy guarantees) that provides explanations on (nonlinear) black box classifiers. The methodology was then used to explain, and then correct, why a classifier from a standard face recognition demo misclassified a seemingly standard test input.

References

[1] Apté, C. and Weiss, S. (1997). Data mining with decision trees and decision rules. *Future Generation Computer Systems*, 13(2):197–210.

[2] Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern Classification*. John Wiley & Sons.

[3] Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669.

[4] Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(1):3133–3181.

[5] Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760.

[6] Guzella, T. S. and Caminhas, W. M. (2009). A review of machine learning approaches to Spam filtering. *Expert Systems with Applications*, 36(7):10206–10222.

[7] Huang, G. B. and Learned-Miller, E. (2014). Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst.

[8] Martens, D., Baesens, B., Van Gestel, T., and Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, 183(3):1466–1476.

[9] Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283.

[10] Mordvintsev, A., Olah, C., and Tyka, M. (2015). Inceptionism: Going deeper into neural networks. *Google Research Blog*.

[11] Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems 14*, pages 841–848.

[12] Ok, E. A. (2007). *Real Analysis with Economic Applications*. Princeton University Press.

[13] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

[14] Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press.

[15] Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517.

[16] Schölkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press.

[17] scikit learn (2015). Faces recognition example using eigenfaces and SVMs. Online demo.

[18] Sirovich, L. and Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, 4(3):519–524.

[19] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288.

[20] Vellido, A., Martin-Guerroro, J., and Lisboa, P. (2012). Making machine learning models interpretable. In *Proceedings of the 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 163–172.