CognitiveScale
THE TRUSTED AI COMPANY

A Brief Primer on the Certifai App for Data Scientists

OVERVIEW

Certifai is a risk assessment tool that repeatedly probes a predictive model M in terms of its input-output behavior and provides an evaluation of model risk along 3 dimensions:[1] Robustness (R), Explainability (E) and Fairness/Bias (F). It treats the model M as a black box - thus it only needs to know the model's response or output corresponding to a given instance or input. This means that any kind of predictive model can be assessed, including rule-based systems, statistical approaches and neural networks. Certifai can be used to assess regression as well as classification models (binary and multi-class) models. However, its application in a binary classification setting is the easiest to understand and is briefly described below.

CERTIFAI FOR BINARY CLASSIFIERS

Without loss of generality, we refer to the two classes as the positive and negative class respectively, with the positive class being a more desired outcome. As a running example, we consider a loan approval case, in which each instance is a loan application, the positive class is "approved" and the negative class is "denied". Any classifier will partition the input space into regions that are assigned the same class label, separated by decision boundaries.

COUNTERFACTUALS

Certifai is based on the notion of a counterfactual (CF), as used in the recent fairness literature, e.g., Wachter et al., 2017 [1] and Google's What-if Tool [2]. Given an instance, a CF is a data-point for which the classifier returns a different class label.[2] The most relevant CFs are those that are as close to possible to the given instance. Then, the difference between the feature vectors representing the probe and its corresponding CF represents the (minimal) change that needs to be made to the probe in order to flip the outcome. Given an input instance (the "probe") and a black-box classifier, Certifai generates a series of (synthetic) instances based on a genetic algorithm, to efficiently query the model and estimate the location of the CF that is nearest to the probe. We use the L1 distance metric and induce a sparsity prior as well so that a CF tends to not differ from the probe in many attributes. These properties make the CF more explainable and actionable, as explained later.

Note that the CF depends on the given probe as well as the given model M. We guarantee that the model will return the opposite class label for the CF. Moreover, while in general it is not possible to guarantee (given limited compute time) that the CF is the closest possible, the CFs that Certifai finds are in general substantially closer than those from alternative approaches, including the one taken by Google [2], which restricts the CF to be an actual data point in the evaluation dataset. Moreover, the user can specify domain constraints in the UI to ensure that the CF is realistic. For example, in a medical application, the user can specify that demographics (gender, ethnicity) of an instance cannot change, but only behaviors and treatments (exercise, medications) can change in the CF. In a future release, Certifai will have the option of returning multiple CFs, representing qualitatively disparate options, all of which lead to a different outcome.

RISK EVALUATION USING COUNTERFACTUALS

Robustness (R), Explainability (E) and Fairness/Bias (F) are all joint properties of the model being assessed as well as the application the model has been designed for. A proper evaluation thus

[1] For a primer on why these three dimensions are critical for evaluating AI business risk and how such risks become a major barrier to the adoption of machine learning based solutions, please see (link to our overview brochure and video).

[2] Note that this definition of a CF is not to be confused with the notion of a "counterfactual" in philosophy [3], which has been used to suggest causality in certain statistical models.

requires an evaluation dataset, D, that is representative of the application of interest. For each instance in this dataset, Certifai determines a suitable CF. All the CFs thus determined then contribute to the model risk evaluation as follows:

- **Robustness:** The distance of a CF to the corresponding probe point, averaged over D, is a measure of robustness or sensitivity, since it indicates the average amount of (adversarial) perturbation needed to flip the outcome. Higher scores indicate less sensitive or more robust models. Scores can also be normalized to a range from 0 to 100 using a proprietary non-linear but monotonic function of the CF distance after normalization w.r.t intra-class data spreads.
- **Explainability:** A CF can be expressed in terms of the number of attributes that need to change in order to flip the outcome. Fewer attributes changing means that the explanation on what it will minimally take to change the outcome (for example to convert an application from denied to accepted status) is more succinct and hence more explainable. Each CF gets a explainability score that is a monotonically decreasing function of the number of attributes involved in the change vector. The mean explainability score is a measure of the explainability of the entire model, normalized to a range from 0 to 100. In addition to the model-level explainability, if CF individual explanations are needed for specific probes or instances, such instances can be collated and specified in an "explanation dataset" in the Certifai UI.
- **Fairness/Bias**: For fairness studies, one has to first define one or more categorical variables, called the grouping feature(s) (aka protected attribute(s)), which is used to partition the instances into subgroups. For example, the groupings can be based on gender, ethnicity, a combination of gender and age, etc. Then, fairness is evaluated by comparing the outcomes or burden imposed by the model across the different subgroups. For a two-class problem, we consider the burden to be zero for a given instance if it receives the desired outcome (positive class). Otherwise the burden is indicated by the difficulty of recourse required to flip the outcome from negative to positive, as measured by the corresponding CF distance. Finally the average burden across the different subgroups is compared using the gini index, a popular measure of inequality. A score of 100 means that each group has the same average burden, while a score of 0 means that one subgroup has all the burden while other groups have no burden at all.

It is well known that there are many indicators of fairness and different notions of justice, such as distributional justice and procedural justice. Machine learning based fairness assessments (e.g. demographic parity, equalized odds) typically focus on distributive justice and consider only binary outcomes. Our approach is more nuanced, since it considers not only the outcome but also how difficult it is to attain a more preferred outcome.

For instructions on how to use Certifai, see VIDEO 2: Using Certifai – Running Evaluations

EXTENSION TO REGRESSION SETTINGS

For regression, an alternate outcome can be defined in terms of relative or absolute thresholds of the predicted value, depending on the application. For example, if the current model produces a function f(x), then two derived functions, $f(x) + \lambda\sigma$ and $f(x) - \lambda\sigma$ can serve as upper and lower boundaries respectively, where λ is a user defined value, and σ is the standard deviation of the outcome. Alternatively, if one is predicting a credit score, then a fixed value, say 650 for FICO, can be specified, such at any score above this value is deemed as desirable. Thus we split the outcome in two or three regimes, "current", "higher" and "lower", depending on one or two boundaries being specified, converting it into a binary or 3-class problem.

REFERENCES

1. Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2):2018, 2017
2. https://pair-code.github.io/what-if-tool/
3. https://plato.stanford.edu/entries/causation-counterfactual/

About CognitiveScale

CognitiveScale is an enterprise AI software company with solutions that helps customers win with intelligent, transparent and trusted AI/ML powered digital systems. Our Cortex software and industry AI accelerators enable businesses to rapidly build, operate, and evolve intelligent, transparent, and trusted AI systems on any cloud. The company's award-winning software is being used by global leaders in banking, insurance, healthcare and digital commerce to increase user engagement, improve employee expertise and productivity, and protect brand and digital infrastructure from AI Business risks. Headquartered in Austin, Texas, CognitiveScale has offices in New York, London, and Hyderabad, India, and is funded by Norwest Venture Partners, Intel Capital, IBM Watson, Microsoft Ventures, and USAA.

Contact an AI specialist at cognitivescale.com/contact