Noname manuscript No.
(will be inserted by the editor)

An Epistemic Approach to the Formal Specification of Statistical Machine Learning

Yusuke Kawamoto

Received: date / Accepted: date

Abstract We propose an epistemic approach to formalizing statistical properties of machine learning. Specifically, we introduce a formal model for supervised learning based on a Kripke model where each possible world corresponds to a possible dataset and modal operators are interpreted as transformation and testing on datasets. Then we formalize various notions of the classification performance, robustness, and fairness of statistical classifiers by using our extension of statistical epistemic logic (StatEL). In this formalization, we show relationships among properties of classifiers, and relevance between classification performance and robustness. As far as we know, this is the first work that uses epistemic models and logical formulas to express statistical properties of machine learning, and would be a starting point to develop theories of formal specification of machine learning.

Keywords Modal logic · Possible world semantics · Machine learning · Classification performance · Robustness · Fairness

1 Introduction

With the increasing use of machine learning in real-life applications, the safety and security of learning-based systems have been of great interest. In particular, many recent studies [40],[12] have found vulnerabilities on the robustness of deep neural networks (DNNs) to malicious inputs, which can lead to disasters in security critical systems, such as self-driving cars. To find out these vulnerabilities in advance, there have been researches on the formal verification and testing methods for the robustness of DNNs in recent years [24, 27, 36, 41]. However, relatively little attention has been paid to the formal specification of machine learning [38].

This work was supported by JSPS KAKENHI Grant Number JP17K12667, by the New Energy and Industrial Technology Development Organization (NEDO), and by Inria under the project LOGIS.

Y. Kawamoto
AIST Tsukuba Central 1, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8560 JAPAN

In the research filed of formal specification and verification, logical approaches have been shown useful to characterize desired properties and to develop theories to discuss those properties. For example, temporal logic [37] is a branch of modal logic for expressing time-dependent propositions, and has been widely used to describe requirements of hardware and software systems. For another example, epistemic logic [44] is a modal logic for knowledge and belief that has been employed as formal policy languages for distributed systems (e.g., for the authentication [8] and the anonymity [39] of security protocols). As far as we know, however, no prior work has employed logical formulas to rigorously describe various statistical properties of machine learning, although there are some papers that (often informally) list various desirable properties of machine learning [38].

In this paper, we present a first logical formalization of statistical properties of machine learning. To describe the statistical properties in a simple and abstract way, we extend *statistical epistemic logic* (StatEL) [28], which is recently proposed to describe statistical knowledge and is applied to formalize statistical hypothesis testing and statistical privacy of databases.

A key idea in our modeling of statistical machine learning is that we formalize logical aspects in the syntax level, and statistical distances and dataset operations in the semantics level by using accessibility relations of a Kripke model [31]. In this model, we formalize supervised learning and some of its desirable properties, including performance, robustness, and fairness. More specifically, classification performance and robustness are described as the differences between the correct class label and the classifier's prediction, whereas fairness is expressed as a conditional indistinguishability between different groups.

Our contributions. The main contributions of this work are as follows:

- We propose a logical approach to formalizing statistical properties of machine learning in a simple and abstract way. Specifically, we introduce a principle that logical aspects of statistical properties are described in the syntax level, and statistical distances and datasets are formalized in the semantics level.
- We formalize supervised learning models and datasets by employing a distributional Kripke model [28] where each possible world corresponds to a possible test dataset, and modal operators are interpreted as transformation and testing on datasets. Then we show how probabilistic behaviors of machine learning models and non-deterministic adversarial inputs are formalized in the distributional Kripke model.
- We propose an extension of statistical epistemic logic (StatEL) as a formal language to describe various properties of machine learning models, including the performance, robustness, and fairness of statistical classifiers. Then the satisfaction of logical formulas representing those properties is associated with their testing using a test dataset. As far as we know, this is the first work that uses logical formulas to formalize various statistical properties of machine learning, and that provides an epistemic view on those properties.
- We show some relationships among properties of classifiers, such as different levels of robustness. We also present certain relationships between classification performance and robustness, which suggest robustness-related properties that have not been formalized in the literature as far as we know.

Cautions and limitations. In this paper, we focus on formalizing properties of supervised learning models that may be tested by using a dataset; i.e., we do *not* deal with unsupervised learning, reinforcement learning, the properties of learning algorithms, quality of training data (e.g., sample bias), quality of testing (e.g., coverage criteria), explainability, temporal properties, or system-level specification. It should be noted that most of the properties formalized in this paper have been known in literatures on machine learning, and the novelty of this work lies in the logical formulation of those statistical properties.

We also remark that this work aims to provide a logical approach to the modeling of statistical properties tested with a dataset, and does not present methods for checking, guaranteeing, or improving the performance/robustness/fairness of machine learning models. As for the satisfiability of logical formulas, we leave the development of testing and (statistical) model checking algorithms as future work, since the research area on the testing and verification of machine learning is relatively new and needs further techniques to improve the scalability. Moreover, in some applications such as image recognition, some atomic formulas (e.g., representing whether an input image is panda) cannot be defined mathematically, and require additional techniques based on experiments. Nevertheless, we demonstrate that describing various properties using logical formulas is useful to explore desirable properties and to discuss their relationships in a framework.

Finally, we emphasize that our work is the first attempt to use epistemic models and logical formulas to express statistical properties of machine learning models, and would be a starting point to develop theories of formal specification of machine learning in future research.

Relationship with the preliminary version. The main novelties of this paper with respect to the preliminary version [29] are as follows:

- We add how the satisfaction of a formula at a possible world can be regarded as the testing of a specification using a test dataset (Sect. 3.1).
- We show how modal operators are used to model the transformation and testing on datasets. For example, *data preparation* T (e.g., data cleaning, data augmentation) can also be formalized as a modal operator $\mathit{\Delta}_T$ (Sect. 3.2).
- We re-interpret the non-classical implication $\supset$ for conditional probabilities in StatEL as a modal operator associated with a conditioning relation (Sect. 3.3).
- We introduce a modal operator $\sim_x^{\varepsilon,D}$ for conditional indistinguishability (Sect. 3.4), and provide a more comprehensible formalization of the fairness of supervised learning (Sect. 7) instead of using counterfactual epistemic operators [29].
- We add a formalization of *generalization error* to capture how accurately a classifier is able to classify previously unseen input data (Sect. 5.3).
- We add a formalization of other fairness notions called *separation* (a.k.a. *equalized odds*) and *sufficiency* (Sect. 7.4) so that this paper covers all three categories of fairness notions [5].

Paper organization. The rest of this paper is organized as follows. Sect. 2 presents notations used in this paper and provides background on statistical distances and statistical epistemic logic (StatEL). Sect. 3 introduces a different view on the modal operators in StatEL and extends the logic with additional operators. Sect. 4 introduces a formal model for describing the behaviors of statistical classifiers and non-deterministic adversarial inputs. Sects. 5, 6, and 7 respectively formalize various notions of the performance, robustness, and fairness of classifiers by using our extension of StatEL. Sect. 8 presents related work and Sect. 9 concludes.

2 Preliminaries

In this section we introduce some notations and background on statistical distance notions, and recall the syntax and semantics of *statistical epistemic logic* (StatEL), introduced in [28].

2.1 Notations

Let $\mathbb{R}^{\geq 0}$ be the set of non-negative real numbers, and $[0, 1]$ be the set of non-negative real numbers not greater than 1. We denote by $\mathbb{D}\mathcal{O}$ the set of all probability distributions over a finite set $\mathcal{O}$. Given a finite set $\mathcal{O}$ and a probability distribution $\mu \in \mathbb{D}\mathcal{O}$, the probability of sampling a value v from μ is denoted by $\mu[v]$. For a subset $R \subseteq \mathcal{O}$ we define $\mu[R]$ by $\mu[R] = \sum_{v \in R} \mu[v]$. For a distribution μ over a finite set $\mathcal{O}$, its *support* is defined by $\mathtt{supp}(\mu) = \{v \in \mathcal{O} \colon \mu[v] > 0\}$.

2.2 Statistical Distance

In this section we recall two popular notions of distance between probability distributions: *total variation* and ∞-*Wasserstein distance.*

Informally, total variation between two distributions μ_0 and μ_1 over a set $\mathcal{O}$ represents the largest difference between the probabilities that μ_0 and μ_1 assign to an identical subset R of $\mathcal{O}$.

Definition 1 (Total variation) For a finite set $\mathcal{O}$, the *total variation* D_{tv} of two distributions $\mu_0, \mu_1 \in \mathbb{D}\mathcal{O}$ is defined by:

$$D_{\mathsf{tv}}(\mu_0 \parallel \mu_1) \overset{\text{def}}{=} \sup_{R \subseteq \mathcal{O}} |\mu_0[R] - \mu_1[R]|.$$

We then recall the ∞-Wasserstein metric [43]. Intuitively, the ∞-Wasserstein metric $W_d(\mu_0, \mu_1)$ between two distributions μ_0, μ_1 represents the minimum largest move between points in a transportation from μ_0 to μ_1.

Definition 2 (∞-Wasserstein metric) Let $\mathcal{O}$ be a finite set and $d : \mathcal{O} \times \mathcal{O} \to \mathbb{R}^{\geq 0}$ be a metric over $\mathcal{O}$. The ∞-*Wasserstein metric* W_d w.r.t. d between two distributions $\mu_0, \mu_1 \in \mathbb{D}\mathcal{O}$ is defined by:

$$W_d(\mu_0, \mu_1) = \min_{\mu \in \mathsf{cp}(\mu_0, \mu_1)} \max_{(v_0, v_1) \in \mathtt{supp}(\mu)} d(v_0, v_1)$$

where $\mathsf{cp}(\mu_0, \mu_1)$ is the set of all couplings[1] of μ_0 and μ_1.

[1] A *coupling* of two distributions $\mu_0, \mu_1 \in \mathbb{D}\mathcal{O}$ is a joint distribution $\mu \in \mathbb{D}(\mathcal{O} \times \mathcal{O})$ such that μ_0 and μ_1 are μ's marginal distributions, i.e., for each $v_0 \in \mathcal{O}$, $\mu_0[v_0] = \sum_{v_1' \in \mathcal{O}} \mu[v_0, v_1']$ and for each $v_1 \in \mathcal{O}$, $\mu_1[v_1] = \sum_{v_0' \in \mathcal{O}} \mu[v_0', v_1]$. Then for a coupling μ, the support $\mathtt{supp}(\mu)$ is the maximum subset of $\mathcal{O} \times \mathcal{O}$ whose elements are assigned non-zero probabilities in μ.

2.3 Syntax of StatEL

We recall the syntax of statistical epistemic logic (StatEL) [28], which has two levels of formulas: *static* and *epistemic formulas.* Intuitively, a static formula describes a proposition satisfied at a (deterministic) state, while an epistemic formula describes a proposition satisfied at a probability distribution of states. In this paper, the former is used only to define the latter.

Formally, let $\mathtt{Mes}$ be a set of symbols called *measurement variables*, and Γ be a set of atomic formulas of the form $\gamma(x_1, x_2, \ldots, x_n)$ for a predicate symbol γ, $n \geq 0$, and $x_1, x_2, \ldots, x_n \in \mathtt{Mes}$. Let $I \subseteq [0, 1]$ be a finite union of disjoint intervals, and $\mathcal{A}$ be a finite set of indices (e.g., associated with statistical divergences). Then the formulas are defined by:

Static formulas: $\psi ::= \gamma(x_1, x_2, \ldots, x_n) \mid \neg\psi \mid \psi \wedge \psi$
Epistemic formulas: $\varphi ::= \mathbb{P}_I\, \psi \mid \neg\varphi \mid \varphi \wedge \varphi \mid \psi \supset \varphi \mid \mathsf{K}_a\, \varphi$

where $a \in \mathcal{A}$. We denote by $\mathcal{F}$ the set of all epistemic formulas. Note that we have no quantifiers over measurement variables. (See Sect. 2.5 for more details.)

The *probability quantification* $\mathbb{P}_I\, \psi$ represents that a static formula ψ is satisfied with a probability belonging to a set I. For instance, $\mathbb{P}_{(0.95,1]}\, \psi$ represents that ψ holds with a probability greater than 0.95. By $\psi \supset \mathbb{P}_I\, \psi'$ we represent that the conditional probability of ψ' given ψ is included in a set I. The *epistemic knowledge* $\mathsf{K}_a\, \varphi$ expresses that we know φ when our capability of observation is denoted by $a \in \mathcal{A}$.

As syntax sugar, we use *disjunction* $\vee$, *classical implication* $\rightarrow$, and *epistemic possibility* P_a, defined as usual by: $\varphi_0 \vee \varphi_1 ::= \neg(\neg\varphi_0 \wedge \neg\varphi_1)$, $\varphi_0 \rightarrow \varphi_1 ::= \neg\varphi_0 \vee \varphi_1$, and $\mathsf{P}_a\, \varphi ::= \neg\, \mathsf{K}_a\, \neg\varphi$. When I is a singleton $\{i\}$, we abbreviate $\mathbb{P}_I$ as $\mathbb{P}_i$.

2.4 Distributional Kripke Model

Next we recall the notion of a distributional Kripke model [28], where each possible world is associated with a probability distribution over a set of states, and with a stochastic assignment of data to measurement variables.

Definition 3 (Distributional Kripke model) Let $\mathcal{A}$ be a finite set of indices (typically associated with operations and tests on datasets), $\mathcal{S}$ be a finite set of states, and $\mathcal{O}$ be a finite set of data, called a *data domain.* A *distributional Kripke model* is a tuple $\mathfrak{M} = (\mathcal{W}, (\mathcal{R}_a)_{a\in\mathcal{A}}, (V_s)_{s\in\mathcal{S}})$ consisting of:

- a non-empty set $\mathcal{W}$ of multisets of states belonging to $\mathcal{S}$;
- for each $a \in \mathcal{A}$, an accessibility relation $\mathcal{R}_a \subseteq \mathcal{W} \times \mathcal{W}$;
- for each $s \in \mathcal{S}$, a valuation $V_s : \Gamma \rightarrow \mathcal{P}(\mathcal{O}^k)$ that maps each k-ary predicate γ to a set $V_s(\gamma)$ of k-tuples of data.

The set $\mathcal{W}$ is called a *universe*, and its elements are called *possible worlds.* A world is said to be *finite* if it is a finite multiset, i.e., it has a finite number of (possibly duplicated) elements. A world is said to be *infinite* if it is an infinite multiset.

The relation $\mathcal{R}_a$ determines an accessibility between two worlds. For example, $(w, w') \in \mathcal{R}_a$ means that a world w' is accessible from a world w when our capability of distinguishing possible worlds is denoted by $a \in \mathcal{A}$. The valuation V_s

may give a possibly different interpretation of a predicate γ at a different state s. We assume that all measurement variables range over the same data domain $\mathcal{O}$ in every world. The interpretation of measurement variables at a state s is given by a deterministic assignment σ_s defined below.

Definition 4 (Deterministic assignment) For any distributional Kripke model $\mathfrak{M} = (\mathcal{W}, (\mathcal{R}_a)_{a\in\mathcal{A}}, (V_s)_{s\in\mathcal{S}})$, we assume that each world $w \in \mathcal{W}$ is associated with a function $\rho_w : \texttt{Mes} \times \mathcal{S} \to \mathcal{O}$ that maps each measurement variable x to its value $\rho_w(x, s)$ that is observed at a state s belonging to the world w. We also assume that each state s in a world w is associated with the *deterministic assignment* $\sigma_s : \texttt{Mes} \to \mathcal{O}$ defined by $\sigma_s(x) = \rho_w(x, s)$.

Since each world w is a multiset of states, we abuse the notation and denote by $w[s]$ the probability that a state s is randomly chosen from w (i.e., the number of occurrences of s in the multiset w, divided by the total number of elements in w). Here we regard each world w as a probability distribution over the states that corresponds to the multiset.

The probability that a measurement variable $x \in \texttt{Mes}$ has a value $v \in \mathcal{O}$ is given by $\sigma_w(x)[v] = \sum_{s\in w,\sigma_s(x)=v} w[s]$. Note that $\sigma_w : \texttt{Mes} \to \mathbb{D}\mathcal{O}$ maps each measurement variable x to a probability distribution $\sigma_w(x)$ over the data domain $\mathcal{O}$. Hence σ_w represents the joint probability distribution of all variables in $\texttt{Mes}$, and is called the *stochastic assignment* at w. When a state s is uniformly drawn from a multiset w of states, a datum $\sigma_s(x)$ is sampled from the distribution $\sigma_w(x)$.

In later sections, a possible world corresponds to a *dataset* (i.e., a multiset of data tuples) from which data are sampled. For example, suppose that we have only three measurement variables $\texttt{Mes} = \{x, y, z\}$. Then for each state s in a world w, the deterministic assignment $\sigma_s : \texttt{Mes} \to \mathcal{O}$ represents the tuple of data $(\sigma_s(x), \sigma_s(y), \sigma_s(z))$. Hence each state s corresponds to a tuple of data, and the world w corresponds to the dataset $\{(\sigma_s(x), \sigma_s(y), \sigma_s(z)) \mid s \in w\}$.

2.5 Stochastic Semantics of StatEL

Now we recall the *stochastic semantics* [28] for the StatEL formulas over a distributional Kripke model $\mathfrak{M} = (\mathcal{W}, (\mathcal{R}_a)_{a\in\mathcal{A}}, (V_s)_{s\in\mathcal{S}})$ with $\mathcal{W} = \mathbb{D}\mathcal{S}$.

The interpretation of static formulas ψ at a state s is given by:

$$\begin{aligned} s \models \gamma(x_1, x_2, \ldots, x_k) \ &\text{ iff } \ (\sigma_s(x_1), \sigma_s(x_2), \ldots, \sigma_s(x_k)) \in V_s(\gamma) \\ s \models \neg\psi \ &\text{ iff } \ s \not\models \psi \\ s \models \psi \wedge \psi' \ &\text{ iff } \ s \models \psi \ \text{ and } \ s \models \psi'. \end{aligned}$$

The *restriction* $w|_\psi$ of a world w to a static formula ψ is defined by $w|_\psi[s] = \frac{w[s]}{\sum_{s':s'\models\psi} w[s']}$ if $s \models \psi$, and $w|_\psi[s] = 0$ otherwise. Note that $w|_\psi$ is undefined if there is no state s that satisfies ψ and has a non-zero probability in w.

Then the interpretation of epistemic formulas in a world w is defined by:

$$\begin{aligned}
\mathfrak{M}, w \models \mathbb{P}_I\, \psi & \text{ iff } \Pr\Big[s \xleftarrow{\$} w :\ s \models \psi\Big] \in I \\
\mathfrak{M}, w \models \neg\varphi & \text{ iff } \mathfrak{M}, w \not\models \varphi \\
\mathfrak{M}, w \models \varphi \wedge \varphi' & \text{ iff } \mathfrak{M}, w \models \varphi \text{ and } \mathfrak{M}, w \models \varphi' \\
\mathfrak{M}, w \models \psi \supset \varphi & \text{ iff } w|_\psi \text{ is defined and } \mathfrak{M}, w|_\psi \models \varphi \\
\mathfrak{M}, w \models \mathsf{K}_a\, \varphi & \text{ iff for every } w' \text{ s.t. } (w, w') \in \mathcal{R}_a,\ \mathfrak{M}, w' \models \varphi,
\end{aligned}$$

where $s \xleftarrow{\$} w$ represents that a state s is sampled from the distribution w.

Then $\mathfrak{M}, w \models \psi_0 \supset \mathbb{P}_I\, \psi_1$ represents that the conditional probability of satisfying a static formula ψ_1 given another ψ_0 is included in a set I at a world w.

In each world w, measurement variables can be interpreted using σ_w. This allows us to assign different values to different occurrences of a variable in a formula; E.g., in $\varphi(x) \to \mathsf{K}_a\, \varphi'(x)$, x occurring in $\varphi(x)$ is interpreted by σ_w in a world w, while x in $\varphi'(x)$ is interpreted by $\sigma_{w'}$ in another w' s.t. $(w, w') \in \mathcal{R}_a$.

Finally, the interpretation of an epistemic formula φ in $\mathfrak{M}$ is given by:

$$\mathfrak{M} \models \varphi \text{ iff for every world } w \text{ in } \mathfrak{M},\ \mathfrak{M}, w \models \varphi.$$

Hereafter we mainly focus on the satisfaction local to a possible world, and $\mathfrak{M}$ may be omitted when it is clear from the context.

3 Modality as Transformation and Testing on Datasets

In this section we introduce a different view on the modal operators in statistical epistemic logic (StatEL), and define additional modal operators that are used to formalize various properties of machine learning in Sects. 5 to 7.

3.1 Checking Satisfaction at a World as Testing with a Dataset

We first show how we regard the satisfaction of a formula φ as testing a system's specification expressed by φ as follows.

As explained in Sect. 2.4, a possible world corresponds to a possible dataset. Thus, given a model $\mathfrak{M}$, a world w, and a formula φ, checking the satisfaction $\mathfrak{M}, w \models \varphi$ can be regarded as testing whether the specification φ of a system (e.g., a machine learning model we formalize in Sect. 4) is satisfied with the dataset w. For example, let φ be a formula representing that a machine learning task (e.g., classification) C fails with probability at most 5%. Then $\mathfrak{M}, w \models \varphi$ represents that when the learning task C is performed using a test dataset w, then it fails for at most 5% of the test data in w.

For simplicity, we discuss the satisfaction of the formulas φ in which neither K_a nor P_a occurs as follows. For each state (namely, data tuple) $s \in w$ and for each static sub-formula ψ of φ, we can efficiently check whether $s \models \psi$.

When the dataset w is finite (i.e., it is a finite multiset of data tuples), we can check the satisfaction $w \models \varphi$ in finite time, more precisely, in linear time in the number of elements in w.

When the dataset w is infinite, however, we cannot check whether $w \models \varphi$ in general. For example, suppose that w be the infinite dataset representing a true distribution from which data are sampled and observed. When we cannot learn w itself, we usually obtain a finite dataset w_{fin} by sampling data from w repeatedly and independently and check a specification φ only with this test dataset w_{fin}.

Hereafter, we mainly deal with distributional Kripke models $\mathfrak{M}$ that have infinite numbers of finite worlds. In the following sections except Sect. 6, we deal only with formulas without K_a nor P_a [2], hence can check their satisfaction at a finite world in finite time.

3.2 Modal Operators for Dataset Transformation

In the rest of Sect. 3, we show that modal operators can be used to model the transformation and testing on datasets.

First, we introduce *modal operators for dataset transformation*. The modal operator $\mathit{\Delta}_T$ defined below is unary (i.e., taking a single formula as argument), and is parameterized with a transformation T between datasets. Intuitively, $w \models \mathit{\Delta}_T\varphi$ represents that a formula φ is satisfied for the dataset w' that is obtained by transforming the current dataset w by T. Formally, the modal operator $\mathit{\Delta}_T$ is interpreted as follows.

Definition 5 (Modality $\mathit{\Delta}_T$ for a dataset transformation T) Given a function $T : \mathcal{W} \rightarrow \mathcal{W}$, we define an accessibility relation as the function $\mathcal{R}_T \overset{\text{def}}{=} \{(w, w') \mid w' = T(w)\}$. Then we define the interpretation of $\mathit{\Delta}_T$ by:

$$\mathfrak{M}, w \models \mathit{\Delta}_T\varphi \text{ iff there exists a } w' \text{ s.t. } (w, w') \in \mathcal{R}_T \text{ and } \mathfrak{M}, w' \models \varphi.$$

For example, machine learning often require *data preparation* to manipulate a given raw dataset into a form that makes a machine learning task feasible and more effective (e.g., *data cleaning*, *data augmentation*). For a data preparation T on a dataset w, $w \models \mathit{\Delta}_T\varphi$ represents that a property φ holds for the prepared dataset $T(w)$.

For another example, the security of machine learning often assumes a certain malicious adversary that can manipulate a given dataset to make a machine learning task fail. Such adversarial operations T on datasets can also be formalized using a different modal operator corresponding to T as we will explain in Sect. 6.

3.3 Modality for Conditioning

We then present another interpretation of the logical connective $\supset$ (defined in Sect. 2.5) used to express conditional probabilities in Sects. 5 and 6. Roughly speaking, we regard the restriction $w|_\psi$ of a world w to a static formula ψ as

[2] The testing of a formula φ is not feasible when an epistemic operator K_a or P_a occurs in φ and the model $\mathfrak{M}$ has a large number of possible worlds. Detailed analysis of time complexity of StatEL is out of the scope of this paper, and should be included in the journal version of our paper [28] that proposed StatEL. As we will discuss in Sect. 6, the robustness of machine learning is formalized using these epistemic operators, hence cannot be tested in practical time unless $\mathfrak{M}$ is comprised of a small number of worlds.

a transformation $\mathcal{R}_\psi$ of w. Then we redefine $\supset$ as a modal operator associated with $\mathcal{R}_\psi$, and call it the *conditioning operator*. Formally, the interpretation of $\supset$ is defined as follows.

Definition 6 (Conditioning operator $\supset$) Assume that the universe $\mathcal{W}$ includes all sub-multisets of each $w \in \mathcal{W}$. Given a static formula ψ, we define an accessibility relation as the *conditioning relation* $\mathcal{R}_\psi \stackrel{\text{def}}{=} \{(w, w|_\psi) \mid w \in \mathcal{W}\}$. Then the interpretation of the conditioning operator $\supset$ is given by:

$$\mathfrak{M}, w \models \psi \supset \varphi \ \text{ iff } \ \text{there exists a } w' \text{ s.t. } (w, w') \in \mathcal{R}_\psi \text{ and } \ \mathfrak{M}, w' \models \varphi.$$

Intuitively, $w \models \psi \supset \varphi$ corresponds to the two operations: (i) transforming the given dataset w to the sub-dataset $w|_\psi$ and (ii) testing whether a property φ holds for the sub-dataset $w|_\psi$.

Note that when no data in the dataset w satisfies the property ψ, then we can describe this as $\mathfrak{M}, w \models \psi \supset \bot$ by using the propositional constant falsum $\bot$.

In Sects. 5 and 6, we show concrete examples using the conditioning operator $\supset$, i.e., the classification performance and robustness of statistical classifiers.

3.4 Modality for Conditional Indistinguishability

Next, we introduce a modal operator that is used to formalize the fairness of machine learning in Sect. 7.

Given two static formulas ψ_0, ψ_1 (e.g., representing male and female), $w|_{\psi_0}(x)$ (resp. $w|_{\psi_1}(x)$) represents the probability distribution of values of a measurement variable x generated from the sub-dataset $w|_{\psi_0}$, e.g., the sub-dataset about male (resp. $w|_{\psi_1}$, e.g., about female). To formalize a certain similarity between x's values generated from the two sub-datasets (e.g., between the benefits for male and for female), we introduce a modal operator $\sim_x^{\varepsilon,D}$ for conditional indistinguishability as follows. We write $\psi_0 \sim_x^{\varepsilon,D} \psi_1$ to represent that the two distributions $w|_{\psi_0}(x)$ and $w|_{\psi_1}(x)$ are indistinguishable up to a threshold ε in terms of a divergence or distance D. Formally, this modality is defined as follows[3].

Definition 7 (Conditional indistinguishability operator $\sim_x^{\varepsilon,D}$) Assume that the universe $\mathcal{W}$ includes all sub-multisets of each $w \in \mathcal{W}$. Given an $x \in \mathtt{Mes}$, an $\varepsilon \in \mathbb{R}^{\geq 0}$, and a divergence or distance $D : \mathbb{D}\mathcal{O} \times \mathbb{D}\mathcal{O} \to \mathbb{R}^{\geq 0}$, we define an accessibility relation by:

$$\mathcal{R}_x^{\varepsilon,D} \stackrel{\text{def}}{=} \{(w_0, w_1) \in \mathcal{W} \times \mathcal{W} \mid D(\sigma_{w_0}(x) \parallel \sigma_{w_1}(x)) \leq \varepsilon\}. \tag{1}$$

Then for static formulas ψ_0 and ψ_1, we define the interpretation of $\psi_0 \sim_x^{\varepsilon,D} \psi_1$ by:

$$\begin{aligned}\mathfrak{M}, w \models \psi_0 \sim_x^{\varepsilon,D} \psi_1 \ \text{ iff } \ & \text{there exist } w_0, w_1 \text{ s.t. } (w, w_0) \in \mathcal{R}_{\psi_0},\ (w, w_1) \in \mathcal{R}_{\psi_1}, \\ & \text{and } (w_0, w_1) \in \mathcal{R}_x^{\varepsilon,D},\end{aligned}$$

where $\mathcal{R}_{\psi_0}$ and $\mathcal{R}_{\psi_1}$ are two conditioning relations in Definition 6.

[3] The semantics for the (binary) composite operator in the arrow logic [7] resembles that for $\sim_x^{\varepsilon,D}$ in Definition 7, although it has a totally different meaning and motivation.

Note that two worlds are related by $\mathcal{R}_x^{\varepsilon,D}$ if they have close probability distributions of the values of x. Intuitively, $w \models \psi_0 \sim_x^{\varepsilon,D} \psi_1$ corresponds to the two operations: (i) transforming the given dataset w to the two sub-datasets $w|_{\psi_0}$ and $w|_{\psi_1}$, and (ii) testing whether the probability distribution of x generated by the dataset $w|_{\psi_0}$ is indistinguishable from that by the dataset $w|_{\psi_1}$.

When $\varepsilon = 0$, the operator $\sim_x^{\varepsilon,D}$ represents the identity of two distributions.

Proposition 1 *For a world w, static formulas ψ_0, ψ_1, and a measurement variable x, $w \models \psi_0 \sim_x^{0,D} \psi_1$ iff the distribution $w|_{\psi_0}(x)$ is identical to $w|_{\psi_1}(x)$.*

This proposition is immediate from the following lemma.

Lemma 1 *For a world w, static formulas ψ_0, ψ_1, and a measurement variable x, $w \models \psi_0 \sim_x^{\varepsilon,D} \psi_1$ iff $D(\sigma_{w|_{\psi_0}}(x) \parallel \sigma_{w|_{\psi_1}}(x)) \leq \varepsilon$.*

Proof Let $w_0 = w|_{\psi_0}$ and $w_1 = w|_{\psi_1}$. Then by Definition 6, we have $(w, w_0) \in \mathcal{R}_{\psi_0}$ and $(w, w_1) \in \mathcal{R}_{\psi_1}$. Hence this lemma follows from Definition 7. □

In Sect. 7, we present concrete examples using the conditional indistinguishability operator $\sim_x^{\varepsilon,D}$, i.e., we formalize various notions of fairness in machine learning by using this operator and the above proposition and lemma.

3.5 Summary on the Modal Language

In summary, modal operators are used to represent transformation and testing on datasets. The unary modal operator $\mathit{\Delta}_T$ can be regarded as a transformation T on datasets, while the binary modal operators $\supset$ and $\sim_x^{\varepsilon,D}$ can be regarded as transforming-then-testing on datasets.

Now the syntax of the formulas is given by:

Static formulas: $\psi ::= \gamma(x_1, x_2, \ldots, x_n) \mid \neg\psi \mid \psi \wedge \psi$
Dataset formulas: $\varphi ::= \mathbb{P}_I\, \psi \mid \neg\varphi \mid \varphi \wedge \varphi \mid \mathit{\Delta}_T\varphi \mid \psi \supset \varphi \mid \psi_0 \sim_x^{\varepsilon,D} \psi_1 \mid \mathsf{K}_a\, \varphi,$

where the epistemic formulas with the additional modality are called *dataset formulas*, since they are interpreted in a world that corresponds to a dataset.

When multiple transformations/testing are sequentially applied to datasets, we can use dataset formulas in which different modal operators are nested. For example, $w \models \mathit{\Delta}_T(\psi \supset \varphi)$ represents that after applying a data preparation T to a dataset w, a property φ holds for the sub-dataset $T(w)|_\psi$ that satisfies ψ.

4 Epistemic Model for Supervised Learning

In this section we introduce a formal model for supervised learning. Specifically, we employ a distributional Kripke model (Definition 3), and formalize a probabilistic behavior of a classifier C and a non-deterministic input x from an adversary in the model. In this formalization, we focus only on the testing of supervised learning models, and do *not* formalize the training of supervised learning models or learning algorithms themselves.

4.1 Classification Problems

Multiclass classification is the problem of classifying a given input into one of multiple classes. Let $\mathtt{L}$ be a finite set of *class labels*[4], and $\mathcal{D}$ be a finite set of *input data* (called *feature vectors*) that we want to classify. Then a *classifier* is a function $C : \mathcal{D} \to \mathtt{L}$ that receives an input datum v and predicts which class (among $\mathtt{L}$) the input v belongs to. In this work, we deal with a situation where some classifier C has already been obtained and its properties should be evaluated, and do *not* model or reason about how classifiers are trained from a training dataset.

We assume a *scoring function* $f : \mathcal{D} \times \mathtt{L} \to \mathbb{R}$ that gives a score $f(v, \ell)$ of predicting the class of an input datum (feature vector) v as a label ℓ. Then for each input $v \in \mathcal{D}$, we denote by $H(v) = \ell$ to represent that a label ℓ maximizes $f(v, \ell)$. For example, when the input v is an image of an animal and ℓ is the animal's name, then $H(v) = \ell$ may represent that an oracle (or a "human") classifies the image v as ℓ.

4.2 Modeling the Behaviors of Classifiers

Classifiers are formalized on a distributional Kripke model $\mathfrak{M} = (\mathcal{W}, (\mathcal{R}_a)_{a\in\mathcal{A}}, (V_s)_{s\in\mathcal{S}})$ with $\mathcal{W} = \mathbb{D}\mathcal{S}$. Then $\mathcal{W}$ is an infinite set of possible worlds that corresponds to *all possible datasets we can imagine*. We denote by $w_{\mathsf{test}} \in \mathcal{W}$ a real world that corresponds to a test dataset. Recall that each world $w \in \mathcal{W}$ is a multiset of states over $\mathcal{S}$ and is associated with a stochastic assignment $\sigma_w : \mathtt{Mes} \to \mathbb{D}\mathcal{O}$ that is consistent with the deterministic assignments σ_s for all $s \in w$, as explained in Sect. 2.4.

We present an overview of our formalization in Fig. 1. We denote by $x \in \mathtt{Mes}$ an input datum given to the classifier C (and to the oracle H), by $y \in \mathtt{Mes}$ a correct label given by the oracle H, and by $\hat{y} \in \mathtt{Mes}$ a label predicted by C. We assume that the input variable x (resp. the output variables $y, \hat{y}$) ranges over the set $\mathcal{D}$ of input data (resp. the set $\mathtt{L}$ of labels); i.e., the deterministic assignment σ_s at each state $s \in \mathcal{S}$ has the range $\mathcal{O} = \mathcal{D} \cup \mathtt{L}$ and satisfies $\sigma_s(x) \in \mathcal{D}$ and $\sigma_s(y), \sigma_s(\hat{y}) \in \mathtt{L}$.

A key idea in our modeling is that we describe logical aspects of statistical properties in the syntax level by using logical formulas, and model statistical distances and dataset operations in the semantics level by using accessibility relations in the distributional Kripke model. In this way, we can formalize various statistical properties of classifiers in a simple and abstract way.

To formalize the classifier C, we introduce a static formula $\psi(x, \hat{y})$ to represent that C classifies a given input x as a class $\hat{y}$. We also introduce a static formula $h(x, y)$ to represent that y is the actual class of an input x. As an abbreviation, we write $\psi_\ell(x)$ (resp. $h_\ell(x)$) to denote $\psi(x, \ell)$ (resp. $h(x, \ell)$). Formally, these static formulas are interpreted at each state $s \in \mathcal{S}$ as follows:

$$\begin{aligned} s \models \psi(x, \hat{y}) \quad &\text{iff} \quad C(\sigma_s(x)) = \sigma_s(\hat{y}). \\ s \models h(x, y) \quad &\text{iff} \quad H(\sigma_s(x)) = \sigma_s(y). \end{aligned}$$

[4] The regression can be regarded as the classification problem when the label ranges over the real numbers, hence it can be formalized using a distributional Kripke model analogously. For simplicity, however, we deal only with the classification problems in this paper.

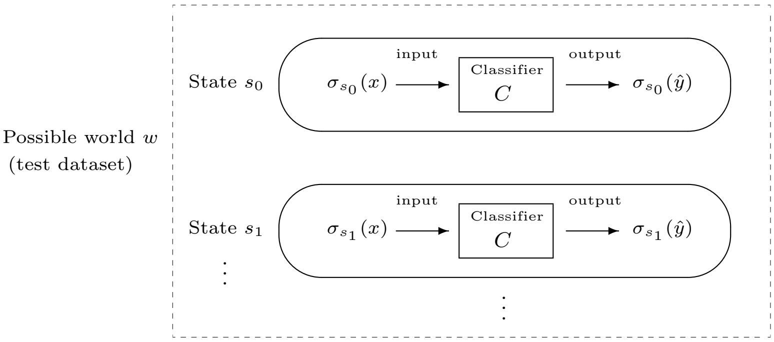

Fig. 1: A world w is chosen non-deterministically and corresponds to a test dataset. With probability $w[s_i]$, the world w is in a deterministic state s_i where the classifier C receives the input value $\sigma_{s_i}(x)$ and returns the output value $\sigma_{s_i}(\hat{y})$. Each state s_i can be regarded as a tuple $(\sigma_{s_i}(x), \sigma_{s_i}(y), \sigma_{s_i}(\hat{y})) \in \mathcal{D} \times \mathtt{L} \times \mathtt{L}$ consisting of an input datum, an actual label, and a predicted label.

4.3 Modeling the Non-deterministic Inputs from Adversaries

We first observe that a distributional Kripke model $\mathfrak{M}$ can formalize an input x that is probabilistically chosen from a given dataset. As explained in Sect. 2.4, each world w corresponds to a *test dataset*. When a state s is drawn from a multiset w of states, an input value $\sigma_s(x)$ is sampled from the distribution $\sigma_w(x)$, and assigned to the measurement variable x. The set of all possible probability distributions of inputs is represented by $\Lambda \overset{\text{def}}{=} \{\sigma_w(x) \mid w \in \mathcal{W}\}$, which is possibly an infinite set.

For example, let us consider testing the classifier C with the actual test dataset $\sigma_{w_{\mathsf{test}}}(x)$. When C classifies an input x as a label ℓ with probability 0.2, i.e.,

$$\Pr\Big[\ v \overset{\$}{\leftarrow} \sigma_{w_{\mathsf{test}}}(x)\ :\ C(v) = \ell\ \Big] = 0.2,$$

then this can be expressed by:

$$\mathfrak{M}, w_{\mathsf{test}} \models \mathbb{P}_{0.2}\, \psi_\ell(x).$$

Next we observe that our model can formalize a non-deterministic input x from an adversary as follows. Although each state s in a possible world w is assigned the probability $w[s]$, each world w itself is not assigned a probability. Thus, each input distribution $\sigma_w(x) \in \Lambda$ itself is also not assigned a probability, hence our model assumes no probability distribution over Λ. In other words, we assume that a world w and thus an input distribution $\sigma_w(x)$ are non-deterministically chosen. This is useful to model an adversary that provides malicious inputs to the classifier C to make its prediction fail, because we usually do not have a prior knowledge of the probability distribution of malicious inputs from adversaries, and need to reason about the worst cases caused by the attack. In Sect. 6, this formalization of non-deterministic inputs is used to express the robustness of classifiers.

Finally, it should be noted that we cannot enumerate all possible adversarial inputs, hence cannot enumerate all possible datasets to construct the universe $\mathcal{W}$. Since $\mathcal{W}$ can be an infinite set and is unspecified, we cannot check whether a formula expressing a security property against an adversary is satisfied in all possible worlds of $\mathcal{W}$. Nevertheless, as shown in later sections, describing various properties using our extension of StatEL is useful to explore desirable properties and to discuss relationships among them.

5 Formalizing the Classification Performance

In this section we show a formalization of classification performance using our extension of StatEL. We formalize popular measures of classification performance, including precision, recall, and accuracy, and measures for evaluating overfitting, such as the generalization error. See Fig. 2 for basic ideas on these formalizations.

5.1 Classifier's Prediction and its Correctness

In classification problems, the terms *positive/negative* represent the result of the classifier's prediction, and the terms *true/false* represent whether the classifier predicts correctly or not. Then the following terminologies are commonly used:

- *true positive* (*tp*): both the prediction and actual class are positive;
- *true negative* (*tn*): both the prediction and actual class are negative;
- *false positive* (*fp*): the prediction is positive but the actual class is negative;
- *false negative* (*fn*): the prediction is negative but the actual class is positive.

These terminologies can be formalized using static formulas as shown in Table 1. For example, when an input x shows true positive at a state s, this can be expressed as $s \models \psi_\ell(x) \wedge h_\ell(x)$. Note that the value of the measurement variable x is uniquely determined by the assignment σ_s at the state s. True negative, false positive (type I error), and false negative (type II error) are respectively expressed as $s \models \neg\psi_\ell(x) \wedge \neg h_\ell(x)$, $s \models \psi_\ell(x) \wedge \neg h_\ell(x)$, and $s \models \neg\psi_\ell(x) \wedge h_\ell(x)$.

5.2 Precision, Recall, Accuracy, and Other Performance Measures

Next we formalize three popular measures for binary classification performance: *precision*, *recall*, and *accuracy*. In Table 1 we summarize the formalization of various notions of classification performance using our dataset formulas.

In theory, these notions should be formalized with the infinite dataset w_{true} representing the true distribution. However, we usually cannot obtain w_{true} or test the performance measures using w_{true}. Hence, we often sample a finite test dataset w_{test} from the true distribution and regard it as an approximation of w_{true}[5].

Given a test dataset w_{test}, *precision* (*positive predictive value*) is defined as the conditional probability that the prediction is correct given that the prediction is positive; i.e., *precision* $= \frac{tp}{tp+fp}$. Since the probability distribution of the input x

[5] Since the test dataset w_{test} is finite, there can be *missing data* that are not included in w_{test} but are sampled from the true distribution w_{true} with a very small probability.

Table 1: Logical description of the table of confusion

	Actual class			
	positive $h_\ell(x)$	negative $\neg h_\ell(x)$	$\mathsf{Prevalence}_{\ell,I}(x)$ $\stackrel{\text{def}}{=} \mathbb{P}_I(h_\ell(x))$	$\mathsf{Accuracy}_{\ell,I}(x)$ $\stackrel{\text{def}}{=} \mathbb{P}_I(\psi_\ell(x) \leftrightarrow h_\ell(x))$
Positive prediction $\psi_\ell(x)$	$\mathit{tp}(x) \stackrel{\text{def}}{=}$ $\psi_\ell(x) \wedge h_\ell(x)$	$\mathit{fp}(x) \stackrel{\text{def}}{=}$ $\psi_\ell(x) \wedge \neg h_\ell(x)$	$\mathsf{Precision}_{\ell,I}(x) \stackrel{\text{def}}{=}$ $\psi_\ell(x) \supset \mathbb{P}_I\, h_\ell(x)$	$\mathsf{FDR}_{\ell,I}(x) \stackrel{\text{def}}{=}$ $\psi_\ell(x) \supset \mathbb{P}_I\, \neg h_\ell(x)$
Negative prediction $\neg\psi_\ell(x)$	$\mathit{fn}(x) \stackrel{\text{def}}{=}$ $\neg\psi_\ell(x) \wedge h_\ell(x)$	$\mathit{tn}(x) \stackrel{\text{def}}{=}$ $\neg\psi_\ell(x) \wedge \neg h_\ell(x)$	$\mathsf{FOR}_{\ell,I}(x) \stackrel{\text{def}}{=}$ $\neg\psi_\ell(x) \supset \mathbb{P}_I\, h_\ell(x)$	$\mathsf{NPV}_{\ell,I}(x) \stackrel{\text{def}}{=}$ $\neg\psi_\ell(x) \supset \mathbb{P}_I\, \neg h_\ell(x)$
	$\mathsf{Recall}_{\ell,I}(x) \stackrel{\text{def}}{=}$ $h_\ell(x) \supset \mathbb{P}_I\, \psi_\ell(x)$	$\mathsf{FallOut}_{\ell,I}(x) \stackrel{\text{def}}{=}$ $\neg h_\ell(x) \supset \mathbb{P}_I\, \psi_\ell(x)$		
	$\mathsf{MissRate}_{\ell,I}(x) \stackrel{\text{def}}{=}$ $h_\ell(x) \supset \mathbb{P}_I\, \neg\psi_\ell(x)$	$\mathsf{Specificity}_{\ell,I}(x) \stackrel{\text{def}}{=}$ $\neg h_\ell(x) \supset \mathbb{P}_I\, \neg\psi_\ell(x)$		

in the world w_{test} is expressed by $\sigma_{w_{\mathsf{test}}}(x)$ as explained in Sect. 4.3, the precision being within an interval I is given by:

$$\Pr\Big[\, v \xleftarrow{\$} \sigma_{w_{\mathsf{test}}}(x) \,:\, H(v) = \ell \;\Big|\; C(v) = \ell \,\Big] \in I,$$

which can be written as:

$$\Pr\Big[\, s \xleftarrow{\$} w_{\mathsf{test}} \,:\, s \models h_\ell(x) \;\Big|\; s \models \psi_\ell(x) \,\Big] \in I.$$

By using StatEL, this can be formalized as:

$$\mathfrak{M}, w_{\mathsf{test}} \models \mathsf{Precision}_{\ell,I}(x) \quad \text{where} \quad \mathsf{Precision}_{\ell,I}(x) \stackrel{\text{def}}{=} \psi_\ell(x) \supset \mathbb{P}_I\, h_\ell(x), \qquad (2)$$

where $\supset$ is the conditioning operator defined in Sect. 3.3. Note that the value of precision depends on the test dataset w_{test}, and can be computed in finite time since w_{test} is finite.

Symmetrically, *recall* (*true positive rate*) is defined as the conditional probability that the prediction is correct given that the actual class is positive; i.e., *recall* $= \frac{tp}{tp+fn}$. Then the recall being within I is formalized as:

$$\mathsf{Recall}_{\ell,I}(x) \stackrel{\text{def}}{=} h_\ell(x) \supset \mathbb{P}_I\, \psi_\ell(x). \qquad (3)$$

Finally, *accuracy* is defined as the probability that the classifier predicts correctly; i.e., *accuracy* $= \frac{tp+tn}{tp+tn+fp+fn}$. Then the accuracy being within I is formalized as:

$$\mathsf{Accuracy}_{\ell,I}(x) \stackrel{\text{def}}{=} \mathbb{P}_I\big(\psi_\ell(x) \leftrightarrow h_\ell(x)\big), \qquad (4)$$

which can also be defined as $\mathbb{P}_I\big(\mathit{tp}(x) \vee \mathit{tn}(x)\big)$. When we measure the accuracy after a data preparation operation T (e.g., data cleaning) to the test dataset w_{test}, this can be represented by $w_{\mathsf{test}} \models \Delta_T \mathsf{Accuracy}_{\ell,I}(x)$.

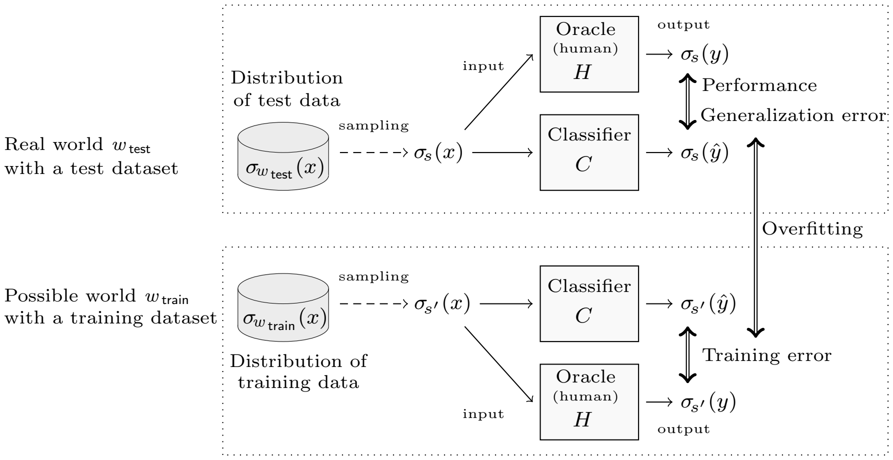

Fig. 2: The classification performance compares the conditional probability of the oracle H's output with that by the classifier C's, while the evaluation of overfitting compares the expected loss by the test dataset with that by the training dataset.

5.3 Generalization Error

We next formalize the *generalization error* of a classifier, i.e., a measure of how accurately a classifier is able to predict the class of *previously unseen* input data. Since a classifier has been trained on a finite sample training dataset w_{train}, it may be *overfitted* to w_{train} and have worse classification performance on new input data that have not been included in the training dataset w_{train}.

To formalize the generalization error, we introduce a formula $\lambda_L(y, \hat{y})$ to represent that given a correct label y and a predicted label $\hat{y}$, the expected value of losses (i.e., real numbers representing the penalty for incorrect classification) is equal to or less than a non-negative real number L. Formally, the semantics of $\lambda_L(y, \hat{y})$ is given by:

$$w \models \lambda_L(y, \hat{y}) \quad \text{iff} \quad \mathop{\mathbb{E}}_{(v,\hat{v})\sim\sigma_w(y,\hat{y})} \mathit{loss}(v, \hat{v}) \leq L,$$

where *loss* is a loss function selected according to the data domain $\mathcal{O}$, and a pair (v, v') of a correct label and a predicted label follows the joint distribution $\sigma_w(y, \hat{y})$.

Now the generalization error being L or smaller at a true distribution w_{true} is formalized as $w_{\mathsf{true}} \models \mathsf{GE}_L(x, y, \hat{y})$ where:

$$\mathsf{GE}_L(x, y, \hat{y}) \stackrel{\text{def}}{=} \big(h(x, y) \wedge \psi(x, \hat{y})\big) \supset \lambda_L(y, \hat{y}).$$

Since we usually cannot obtain the true distribution w_{true} and cannot check the satisfaction $w_{\mathsf{true}} \models \mathsf{GE}_L(x, y, \hat{y})$, we often compute an empirical error (as an approximation of the generalization error) by using a finite test dataset w_{test} that is believed to be an approximation of w_{true}. This testing can be expressed as $w_{\mathsf{test}} \models \mathsf{GE}_L(x, y, \hat{y})$.

On the other hand, given a training dataset w_{train}, the *training error* being L_{train} or smaller is represented by $w_{\mathsf{train}} \models \mathsf{GE}_{L_{\mathsf{train}}}(x, y, \hat{y})$. Then the overfitting of the classifier can be evaluated by comparing the empirical error L with the training error L_{train}. For example, when the empirical error is smaller than $L_{\mathsf{train}}+\varepsilon$ for some error bound $\varepsilon > 0$, then this can be represented by $w_{\mathsf{test}} \models \mathsf{GE}_{L_{\mathsf{train}}+\varepsilon}(x, y, \hat{y})$.

6 Formalizing the Robustness of Classifiers

Many recent studies have found attacks on machine learning where a malicious adversary manipulates the input to cause a malfunction in a machine learning task [12]. Such input data, called *adversarial examples* [40], are designed to make a classifier fail to predict the actual class ℓ of the input, but are recognized to belong to ℓ from human eyes. In computer vision, for example, Goodfellow et al. [21] create an adversarial example by adding undetectable noise to a panda's photo so that humans can still recognize the perturbed image as a panda, but a classifier misclassifies it as a gibbon. To prevent or mitigate such attacks, the classifier should be *robust* against perturbed input, i.e., it should return similar predicted labels given similar input data.

In this section we formalize robustness notions for classifiers by using epistemic operators in StatEL (See Fig. 3 for an overview of the formalization). Furthermore, we present certain relationships between classification performance and robustness, and suggest robustness-related properties that have not been formalized in the literature as far as we know. We present an overview of these formalizations and relationships in Fig. 4.

6.1 Total Correctness of Classifiers

We first note that the *total correctness* of classifiers could be formalize as a classification performance (e.g., precision, recall, or accuracy) in the presence of all possible inputs from adversaries. For example, the total correctness could be formalized as $\mathfrak{M} \models \mathsf{Recall}_{\ell,I}(x)$, which represents that $\mathsf{Recall}_{\ell,I}(x)$ is satisfies in all possible worlds of $\mathfrak{M}$.

In practice, however, it is not possible or tractable to test whether the classification performance is achieved for all possible test datasets (corresponding to an infinite number of possible worlds in $\mathfrak{M}$). Hence we need a weaker form of correctness notions, which may be verified or tested in some way. In the following sections, we deal with robustness notions that are weaker than total correctness.

6.2 Accessibility Relation for Robustness

To formalize robustness notions, we introduce an accessibility relation $\mathcal{R}_x^{\varepsilon,W_d}$ that relates two worlds having closer inputs as follows.

Definition 8 (Accessibility relation for robustness) We define an accessibility relation $\mathcal{R}_x^{\varepsilon,W_d} \subseteq \mathcal{W} \times \mathcal{W}$ by:

$$\mathcal{R}_x^{\varepsilon,W_d} \stackrel{\text{def}}{=} \big\{(w, w') \in \mathcal{W} \times \mathcal{W} \;\mid\; W_d(\sigma_w(x),\, \sigma_{w'}(x)) \leq \varepsilon\big\},$$

where W_d is ∞-Wasserstein distance w.r.t. a metric d in Definition 2.

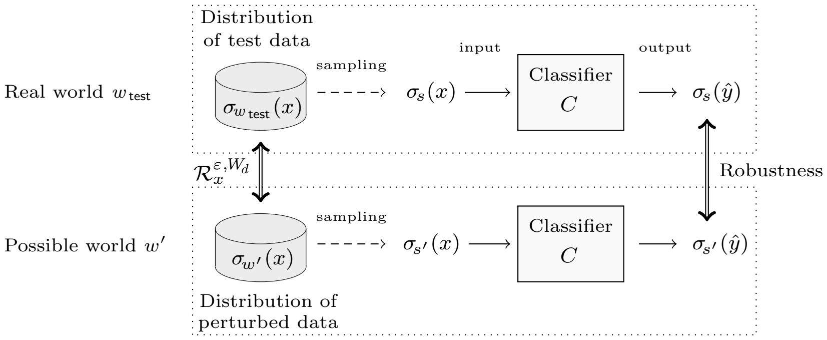

Fig. 3: The robustness compares the conditional probability in the test dataset w_{test} with that in another possible world w' that is close to w_{test} in terms of $\mathcal{R}_x^{\varepsilon,W_d}$. Note that an adversary's choice of the input distribution $\sigma_{w'}(x)$ is formalized as a non-deterministic choice of the possible world w'.

Then $(w, w') \in \mathcal{R}_x^{\varepsilon,W_d}$ represents that the two distributions $\sigma_w(x)$ and $\sigma_{w'}(x)$ of inputs to the classifier C are close in terms of the distance W_d [6]. Intuitively, for example, W_d means the distance between two image datasets $\sigma_w(x)$ and $\sigma_{w'}(x)$ when the distance between individual images are measured by a metric d.

Then an epistemic formula $\mathsf{K}^{\varepsilon, W_d}\,\varphi$ represents that we are confidence that φ is true even when the input data are perturbed by noise of the level ε or smaller.

6.3 Probabilistic Robustness against Targeted Attacks

When a robustness attack aims at misclassifying an input as a specific target label $\hat{\ell}_{\mathsf{target}}$, then it is called a *targeted attack*. For instance, in the above-mentioned attack by [21], a gibbon is the target into which a panda's photo is misclassified.

In this section, we discuss how we formalize robustness using the epistemic operator $\mathsf{K}^{\varepsilon, W_d}$. We denote by $v \in \mathcal{D}$ an original input image in the test dataset w_{test}, and by $\widetilde{v} \in \mathcal{D}$ an image obtained by perturbing the original image v by noise.

A first definition of robustness against targeted attacks might be:

$$\text{For any } v, \widetilde{v} \in \mathcal{D},\ \text{ if } H(v) = \mathsf{panda} \text{ and } d(v, \widetilde{v}) \leq \varepsilon,\ \text{ then } C(v') \neq \mathsf{gibbon},$$

which represents that when an image $\widetilde{v}$ is obtained by perturbing a panda's photo v by noise, then it will not be classified as the target label gibbon at all. This can be formalized using StatEL by:

$$\mathfrak{M}, w_{\mathsf{test}} \models h_{\mathsf{panda}}(x) \supset \mathsf{K}^{\varepsilon, W_d}\, \mathbb{P}_0\, \psi_{\mathsf{gibbon}}(x).$$

However, this notion does not accept a negligible probability of misclassification, and does not cover the case where the human cannot recognize the perturbed image

[6] $W_d(\sigma_w(x), \sigma_{w'}(x)) \leq \varepsilon$ expresses that each value of the input x from the dataset w is close to the corresponding value of x from w' in terms of the metric d between individual data. For example, each input image x in the dataset w looks similar to the corresponding image in w' from the human' eyes.

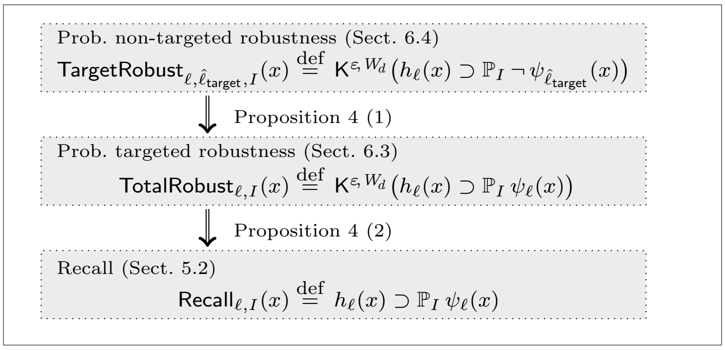
Prob. non-targeted robustness (Sect. 6.4)
$\mathsf{TargetRobust}_{\ell,\hat{\ell}_{\mathsf{target}},I}(x) \overset{\text{def}}{=} \mathsf{K}^{\varepsilon,W_d}\big(h_\ell(x) \supset \mathbb{P}_I \neg \psi_{\hat{\ell}_{\mathsf{target}}}(x)\big)$
Proposition 4 (1)
Prob. targeted robustness (Sect. 6.3)
$\mathsf{TotalRobust}_{\ell,I}(x) \overset{\text{def}}{=} \mathsf{K}^{\varepsilon,W_d}\big(h_\ell(x) \supset \mathbb{P}_I\, \psi_\ell(x)\big)$
Proposition 4 (2)
Recall (Sect. 5.2)
$\mathsf{Recall}_{\ell,I}(x) \overset{\text{def}}{=} h_\ell(x) \supset \mathbb{P}_I\, \psi_\ell(x)$

Fig. 4: Robustness notions and their relationships.

$\widetilde{v}$ as panda (e.g., when the perturbed image $\widetilde{v}$ is obtained by linear displacement, rescaling, and rotation [2], then $H(\widetilde{v}) \neq \mathsf{panda}$ may hold).

To overcome these issues, we introduce the following definition with some conditional probability δ of misclassification as follows.

Definition 9 (Probabilistic targeted robustness) Let $\delta \in [0,1]$. Given a dataset w_{test}, a classifier C satisfies *probabilistic targeted robustness* w.r.t. an actual label ℓ and a target label $\hat{\ell}_{\mathsf{target}}$ if for any input $v \in \mathsf{supp}(\sigma_{w_{\mathsf{test}}}(x))$ from the dataset w_{test}, and for any perturbed input $\widetilde{v} \in \mathcal{D}$ s.t. $d(v, v') \leq \varepsilon$, we have:

$$\Pr[\, C(\widetilde{v}) = \hat{\ell}_{\mathsf{target}} \mid H(\widetilde{v}) = \ell \,] \leq \delta. \tag{5}$$

For instance, when the actual class ℓ is panda and the target label $\hat{\ell}_{\mathsf{target}}$ is gibbon, then the classifier C misclassifies a panda's photo as gibbon with only a small probability δ.

Now we express this robustness notion with $I = [1-\delta, 1]$ by using StatEL.

Proposition 2 (Probabilistic targeted robustness) *Let $I \subseteq [0,1]$. The probabilistic targeted robustness w.r.t. an actual label ℓ and a target label $\hat{\ell}_{\mathsf{target}}$ under a given test dataset w_{test} is expressed by $w_{\mathsf{test}} \models \mathsf{TargetRobust}_{\ell,\hat{\ell}_{\mathsf{target}},I}(x)$ where:*

$$\mathsf{TargetRobust}_{\ell,\hat{\ell}_{\mathsf{target}},I}(x) \overset{def}{=} \mathsf{K}^{\varepsilon,W_d}\big(h_\ell(x) \supset \mathbb{P}_I \neg \psi_{\hat{\ell}_{\mathsf{target}}}(x)\big).$$

Proof Let w' be a possible world such that $(w_{\mathsf{test}}, w') \in \mathcal{R}_x^{\varepsilon,W_d}$. Then w' corresponds to the dataset obtained by perturbing each data in w. Let $\widetilde{v} \in \mathsf{supp}(\sigma_{w'}(x))$. Then $\widetilde{v}$ represents a perturbed input. Let $w'' = w'|_{h_\ell(x)}$. Then (5) is logically equivalent to $w'' \models \mathbb{P}_{[0,\delta]}\, \psi_{\hat{\ell}_{\mathsf{target}}}(x)$. By Definition 6, $w' \models h_\ell(x) \supset \mathbb{P}_{[0,\delta]}\, \psi_{\hat{\ell}_{\mathsf{target}}}(x)$. By $I = [1-\delta, 1]$, $w' \models h_\ell(x) \supset \mathbb{P}_I \neg \psi_{\hat{\ell}_{\mathsf{target}}}(x)$. Therefore this proposition follows from the semantics for $\mathsf{K}^{\varepsilon,W_d}$. □

Since the L^p-norms are often regarded as reasonable approximations of human perceptual distances [10], they are used as distance constraints on the perturbation in many researches on targeted attacks (e.g. [40,21,10]). Our model can represent the robustness against these attacks by using the L^p-norm as a metric d for $\mathcal{R}_x^{\varepsilon,W_d}$.

6.4 Probabilistic Robustness against Non-Targeted Attacks

In this section we formalize *non-targeted attacks* [34,33] in which adversaries try to misclassify inputs as some arbitrary incorrect labels (i.e., not as a specific label like a gibbon). Compared to targeted attacks, this kind of attacks are easier to mount, but harder to defend.

We first define the notion of robustness against non-targeted attacks as follows.

Definition 10 (Probabilistic non-targeted robustness) Let $\delta \in [0,1]$. Given a dataset w_{test}, a classifier C satisfies *probabilistic non-targeted robustness* w.r.t. an actual label ℓ if for any input $v \in \mathsf{supp}(\sigma_{w_{\mathsf{test}}}(x))$ from the dataset w_{test}, and for any perturbed input $\widetilde{v} \in \mathcal{D}$ s.t. $d(v, v') \leq \varepsilon$, we have:

$$\Pr[\, C(\widetilde{v}) = \ell \mid H(\widetilde{v}) = \ell \,] > 1 - \delta.$$

Now we express this robustness notion with $I = [1-\delta, 1]$ by using StatEL.

Proposition 3 (Probabilistic non-targeted robustness) *Let* $I \subseteq [0,1]$. *The probabilistic non-targeted robustness under a given test dataset* w_{test} *is expressed by* $w_{\mathsf{test}} \models \mathsf{TotalRobust}_{\ell,I}(x)$ *where:*

$$\mathsf{TotalRobust}_{\ell,I}(x) \overset{def}{=} \mathsf{K}^{\varepsilon, W_d}\big(h_\ell(x) \supset \mathbb{P}_I\, \psi_\ell(x)\big) = \mathsf{K}^{\varepsilon, W_d}\, \mathsf{Recall}_{\ell,I}(x).$$

Proof The proof is analogous to that for Proposition 2. □

6.5 Relationships among Robustness Notions

In this section we present relationships among notions of robustness and performance, and discuss properties related to robustness.

We first present the following proposition immediate from the definitions.

Proposition 4 (Relationships among notions) *Let* $I \subseteq [0,1]$ *and* $\ell, \hat{\ell}_{\mathsf{target}} \in \mathsf{L}$. *Then we have:*

1. $w_{\mathsf{test}} \models \mathsf{TotalRobust}_{\ell,I}(x)$ *implies* $w_{\mathsf{test}} \models \mathsf{TargetRobust}_{\ell,\hat{\ell}_{\mathsf{target}},I}(x)$.
2. $w_{\mathsf{test}} \models \mathsf{TotalRobust}_{\ell,I}(x)$ *implies* $\mathfrak{M}, w_{\mathsf{test}} \models \mathsf{Recall}_{\ell,I}(x)$.

The first claim means that probabilistic non-targeted robustness is not weaker than probabilistic targeted robustness for the same I. The second claim means that probabilistic non-targeted robustness implies recall without perturbation noise. Note that this is immediate from the reflexivity of $\mathcal{R}_x^{\varepsilon,W_d}$.

Next we remark that our extension of StatEL can be used to describe a certain situation where adversarial attacks are mitigated. When we apply some mechanism T that preprocesses a given input to mitigate attacks on robustness, then the probabilistic targeted robustness is expressed as $w_{\mathsf{test}} \models \Delta_T \mathsf{TotalRobust}_{\ell,I}(x)$ where Δ_T is the modality for the dataset transformation T.

Finally, we recall that by Proposition 3, robustness can be regarded as recall in the presence of perturbed noise. This implies that for each property φ in the table of confusion (Table 1), we could consider $\mathsf{K}^{\varepsilon, W_d}\,\varphi$ as a property to evaluate the classification performance in the presence of adversarial inputs although this has not been formalized in the literature of robustness of machine learning as far as we recognize. For example, $\mathsf{K}^{\varepsilon, W_d}\,\mathsf{Precision}_{\ell,i}(x)$ represents that in the presence of perturbed noise, the prediction is correct with a probability i given that it is positive. For another example, $\mathsf{K}^{\varepsilon, W_d}\,\mathsf{Accuracy}_{\ell,i}(x)$ represents that in the presence of perturbed noise, the prediction is correct (whether it is positive or negative) with a probability i.

7 Formalizing the Fairness of Classifiers

Many studies have proposed and investigated various notions of fairness in machine learning [5]. Informally, these fairness notions mean that the results of machine learning tasks are irrelevant of some sensitive attributes, e.g, gender, age, race, disease, political/religious view. In a recently few years, there have been studies on the testing methods for fairness of machine learning [19,1,42].

In this section, we formalize popular notions of fairness of supervised learning by using our extension of StatEL. Here we focus on the fairness that should be maintained in the *impact* (i.e., the results of machine learning tasks) rather than the *treatment* (i.e., the process of machine learning tasks). This is because previous research show that many seemingly neutral features have statistical relationships with sensitive attributes, and hence just ignoring or removing sensitive attributes in the process of data preparation and training[7] is often ineffective or harmful to achieve the fairness and performance of learning tasks.

7.1 Basic Ideas and Notations

Various notions of fairness in supervised learning are classified into three categories: *independence*, *separation*, and *sufficiency* [5]. All of these have the form of (conditional) independence or its relaxation, and thus can be formalized using the modal operator $\sim_x^{\varepsilon,D}$ for conditional indistinguishability (defined in Sect. 3.4) in our extension of StatEL[8]

In the formalization of fairness notions, we use a distributional Kripke model $\mathfrak{M} = (\mathcal{W}, (\mathcal{R}_a)_{a\in\mathcal{A}}, (V_s)_{s\in\mathcal{S}})$. Recall that x, y, and $\hat{y}$ are measurement variables respectively denoting the input datum, the actual class label (given by the oracle H), and the predicted label (output by the classifier C). Given a real world w_{test} (corresponding to a given test dataset), $\sigma_{w_{\mathsf{test}}}(x)$ is the probability distribution of C's test input over $\mathcal{D}$, $\sigma_{w_{\mathsf{test}}}(y)$ is the distribution of the actual label over $\mathtt{L}$, and $\sigma_{w_{\mathsf{test}}}(\hat{y})$ is the distribution of C's output over $\mathtt{L}$.

Fairness notions are usually defined in terms of some *sensitive attribute* (e.g, gender, age, race, disease, political/religious view), which is defined as a tuple of subsets of the input data domain $\mathcal{D}$. For example, a sensitive attribute based on ages can be defined as a pair of groups G_0 (input data with ages 21 to 60) and G_1

[7] Such *unawareness* requires that sensitive attributes are not explicitly used in the learning process. However, StatEL may not be suited to formalizing this requirement.

[8] Compared to the preliminary version [29] of this paper, we corrected errors and changed the formalization into a more comprehensible form by introducing the operator $\sim_x^{\varepsilon,D}$ and by removing the *counter factual epistemic operators* and a formula ξ_d representing that the input is drawn from a dataset d.

Table 2: Popular notions of fairness of machine learning

Sect.	Formalization of fairness notions
7.2	Independence (a.k.a. group fairness) $\mathsf{GrpFair}_{\varepsilon}(x, \hat{y}) \overset{\text{def}}{=} \big(\eta_{G_0}(x) \wedge \psi(x, \hat{y})\big) \sim_{\hat{y}}^{\varepsilon, D_{\mathsf{tv}}} \big(\eta_{G_1}(x) \wedge \psi(x, \hat{y})\big)$
7.3	Separation (a.k.a. equalized odds) $\mathsf{EqOdds}_{\varepsilon}(x, \hat{y}) \overset{\text{def}}{=} \bigwedge_{\ell \in \mathsf{L}} \Big(\big(\eta_{G_0}(x) \wedge \psi(x, \hat{y}) \wedge h_\ell(x)\big) \sim_{\hat{y}}^{\varepsilon, D_{\mathsf{tv}}} \big(\eta_{G_1}(x) \wedge \psi(x, \hat{y}) \wedge h_\ell(x)\big)\Big)$
7.3	Equal opportunity (a relaxation of separation) $\mathsf{EqOpp}(x, \hat{y}) \overset{\text{def}}{=} \big(\eta_{G_0}(x) \wedge \psi(x, \hat{y}) \wedge h_\ell(x)\big) \sim_{\hat{y}}^{0, D_{\mathsf{tv}}} \big(\neg\eta_{G_0}(x) \wedge \psi(x, \hat{y}) \wedge h_\ell(x)\big)$
7.4	Sufficiency (a.k.a. conditional use accuracy equality) $\mathsf{Sufficency}_{\varepsilon}(x, y) \overset{\text{def}}{=} \bigwedge_{\hat{\ell} \in \mathsf{L}} \Big(\big(\eta_{G_0}(x) \wedge \psi_{\hat{\ell}}(x) \wedge h(x, y)\big) \sim_{y}^{\varepsilon, D_{\mathsf{tv}}} \big(\eta_{G_1}(x) \wedge \psi_{\hat{\ell}}(x) \wedge h(x, y)\big)\Big)$

(ages 61 to 100). For each group $G \subseteq \mathcal{D}$ of inputs, we introduce a static formula $\eta_G(x)$ representing that an input x belongs to G. Formally, this is interpreted by:

$$\text{For each state } s \in \mathcal{S}, \;\; s \models \eta_G(x) \;\text{ iff }\; \sigma_s(x) \in G.$$

Roughly speaking, a machine learning task is said to be fair if the outcome of the task for a group G_0's input is similar to that for another group G_1's input[9]. In the following sections, we formalize the three categories of fairness of classifiers and their relaxation. A summary of this formalization is presented in Table 2.

7.2 Independence (a.k.a. Group Fairness, Statistical Parity) and its Relaxation

In this section we explain and formalize the notion of *independence* [9], which is also known as *group fairness* [15] [10], and its relaxed notion. Intuitively, independence means that the predicted label $\hat{y}$ does not have statistical relationships with the membership in a sensitive group. For example, independence does not allow a bank's lending rate to be correlated with a sensitive attribute such as gender.

We first present the definition of a relaxed notion of independence, called *group fairness up to bias* ε [15] as follows. Intuitively, this is the property that the output distributions of the classifier are roughly identical when input data belong to different groups.

Formally, this fairness notion is defined as follows.

Definition 11 (Independence & group fairness up to bias ε) Let $G_0, G_1 \subseteq \mathcal{D}$ be sets of input data constituting a sensitive attribute. For each $b = 0, 1$, let $\mu_{G_b} \in \mathbb{D}\mathsf{L}$ be the probability distribution of the predicted label $\hat{\ell}$ output by a classifier C when an input v is sampled from a test dataset w_{test} and belongs to G_b; i.e., for each $\hat{\ell} \in \mathsf{L}$,

$$\mu_{G_b}[\hat{\ell}] \overset{\text{def}}{=} \Pr[\, C(v) = \hat{\ell} \mid v \xleftarrow{\$} \sigma_{w_{\mathsf{test}}}(x) \text{ and } v \in G_b \,]. \tag{6}$$

[9] Some fairness notions (e.g., equal opportunity) assume $G_1 = \mathcal{D} \setminus G_0$.

[10] In previous literature, independence has been referred to also as different terminologies, such as *statistical parity*, *demographic parity*, and *disparate impact*.

Then a classifier C satisfies the *group fairness between groups G_0 and G_1 up to bias ε* if we have $D_{\mathsf{tv}}(\mu_{G_0} \| \mu_{G_1}) \leq \varepsilon$, where D_{tv} is the total variation between distributions (defined in Sect. 2.2). A classifier C satisfies *independence* w.r.t. groups G_0 and G_1 if it satisfies the group fairness between G_0 and G_1 up to bias 0.

Now we express this fairness notion using our extension of StatEL as follows.

Proposition 5 (Independence & group fairness up to bias ε) *The group fairness between groups G_0 and G_1 up to bias ε under a given test dataset w_{test} is expressed as $w_{\mathsf{test}} \models \mathsf{GrpFair}_\varepsilon(x, \hat{y})$ where:*

$$\mathsf{GrpFair}_\varepsilon(x, \hat{y}) \stackrel{def}{=} \big(\eta_{G_0}(x) \land \psi(x, \hat{y})\big) \sim_{\hat{y}}^{\varepsilon, D_{\mathsf{tv}}} \big(\eta_{G_1}(x) \land \psi(x, \hat{y})\big).$$

Independence (without bias ε) is expressed by $w_{\mathsf{test}} \models \mathsf{GrpFair}_0(x, \hat{y})$.

Proof Let $w_b = w_{\mathsf{test}}|_{\eta_{G_b}(x) \land \psi(x, \hat{y})}$. It follows from (6) that for each $\hat{\ell} \in \mathsf{L}$,

$$\mu_{G_b}[\hat{\ell}] = \Pr[\, \sigma_s(\hat{y}) = \hat{\ell} \mid s \xleftarrow{\$} w_b \,],$$

hence $\mu_{G_b} = \sigma_{w_b}(\hat{y})$. Thus, by Definition 11, the group fairness between groups G_0 and G_1 up to bias ε is given by $D_{\mathsf{tv}}(\sigma_{w_0}(\hat{y}) \| \sigma_{w_1}(\hat{y})) \leq \varepsilon$. Therefore, this proposition follows from Lemma 1. □

7.3 Separation (a.k.a. Equalized Odds) and its Relaxation (Equal Opportunity)

In this section we explain and formalize the notion of *separation* [5] [11], which is well-known as *equalized odds* [23], and its relaxed notion called *equal opportunity* [23]. The motivation behind these notions is to capture typical scenarios in which sensitive characteristics may have statistical relationships with the actual class label. For instance, even when some sensitive attribute is correlated with an actual default rate on loans, banks might want to have a different lending rate for people who have a higher default rate. However, independence (group fairness) does allow this, since it requires that the lending rate should be statistically independent of the sensitive attribute.

To overcome this problem, the notion of separation allows statistical relationships between a sensitive attribute and the predicted label $\hat{y}$ output by the classifier C to the extent that this is justified by the actual class label y. More precisely, separation means that the predicted label $\hat{y}$ is conditionally independent of the membership in a sensitive group, given an actual class label y.

Formally, separation is defined as a property that recall (true positive rate) and specificity (true negative rate, explained in Table 1) are the same for all the groups, and equal opportunity is defined a special case of separation only for an advantageous class label.

[11] In previous literature, separation has been referred to also as *disparate mistreatment* [45] and *conditional procedure accuracy equality* [6].

Definition 12 (Separation & equal opportunity) Given a group $G_b \subseteq \mathcal{D}$ and an actual class label ℓ, let $\mu_{G_b,\ell} \in \mathbb{D}\mathtt{L}$ be the probability distribution of the predicted label $\hat{\ell}$ output by a classifier C when an input $v \in G_b$ is sampled from a test dataset w_{test} and is associated with an actual label ℓ; i.e., for each $\hat{\ell} \in \mathtt{L}$,

$$\mu_{G_b,\ell}[\hat{\ell}\,] \overset{\text{def}}{=} \Pr[\,C(v) = \hat{\ell} \mid v \overset{\$}{\leftarrow} \sigma_{w_{\mathsf{test}}}(x), \quad v \in G_b, \quad \text{and} \quad H(v) = \ell\,]. \tag{7}$$

A classifier C satisfies *separation* between two groups G_0 and G_1 if $\mu_{G_0,\ell} = \mu_{G_1,\ell}$ holds for all $\ell \in \mathtt{L}$. A classifier C satisfies *equal opportunity* of an advantageous label ℓ w.r.t. a group G_0 if $\mu_{G_0,\ell} = \mu_{G_1,\ell}$ where $G_1 = \mathcal{D} \setminus G_0$.

Now we express these two notions using our extension of StatEL as follows.

Proposition 6 (Separation) *The separation between two groups G_0 and G_1 under a given test dataset w_{test} is expressed as $w_{\mathsf{test}} \models \mathsf{EqOdds}_0(x, \hat{y})$ where:*

$$\mathsf{EqOdds}_\varepsilon(x, \hat{y}) \overset{def}{=} \bigwedge_{\ell \in \mathtt{L}} \Big(\big(\eta_{G_0}(x) \wedge \psi(x, \hat{y}) \wedge h_\ell(x)\big) \sim_{\hat{y}}^{\varepsilon, D_{\mathsf{tv}}} \big(\eta_{G_1}(x) \wedge \psi(x, \hat{y}) \wedge h_\ell(x)\big) \Big).$$

Proof Let $\ell \in \mathtt{L}$ and $w_{b,\ell} = w_{\mathsf{test}}|_{\eta_{G_b}(x) \wedge \psi(x,\hat{y}) \wedge h_\ell(x)}$. It follows from (7) that:

$$\mu_{G_b,\ell}[\hat{\ell}\,] = \Pr[\,\sigma_s(\hat{y}) = \hat{\ell} \mid s \overset{\$}{\leftarrow} w_{b,\ell}\,],$$

hence $\mu_{G_b,\ell} = \sigma_{w_{b,\ell}}(\hat{y})$. Thus, by Definition 12, the separation between G_0 and G_1 is given by $\sigma_{w_{0,\ell}}(\hat{y}) = \sigma_{w_{1,\ell}}(\hat{y})$ for all $\ell \in \mathtt{L}$. Therefore, this proposition follows from Proposition 1. □

It should be noted that for $\varepsilon > 0$, $\mathsf{EqOdds}_\varepsilon(x, \hat{y})$ represents a relaxation of separation up to bias ε in terms of total variation D_{tv}.

Proposition 7 (Equal opportunity) *The equal opportunity of a label ℓ w.r.t. a group G_0 under a given test dataset w_{test} is expressed as $w_{\mathsf{test}} \models \mathsf{EqOpp}(x, \hat{y})$ where:*

$$\mathsf{EqOpp}(x, \hat{y}) \overset{def}{=} \big(\eta_{G_0}(x) \wedge \psi(x, \hat{y}) \wedge h_\ell(x)\big) \sim_{\hat{y}}^{0, D_{\mathsf{tv}}} \big(\neg\eta_{G_0}(x) \wedge \psi(x, \hat{y}) \wedge h_\ell(x)\big).$$

Proof The proof of this proposition is similar to that of Proposition 6. Let $G_1 = \mathcal{D} \setminus G_0$. By $\mu_{G_b,\ell} = \sigma_{w_{b,\ell}}(\hat{y})$, the equal opportunity of ℓ w.r.t. G_0 is given by $\sigma_{w_{0,\ell}}(\hat{y}) = \sigma_{w_{1,\ell}}(\hat{y})$. Therefore, this proposition follows from Proposition 1. □

7.4 Sufficiency (a.k.a. Conditional Use Accuracy Equality)

In this section we explain and formalize the notion of *sufficiency* [5], which is also known as *conditional use accuracy equality* [6].

While separation guarantees the equality of recall among different groups, sufficiency requires the equality of precision. More precisely, sufficiency is defined as the property that precision (positive predictive value) and negative predictive value (presented as NPV in Table 1) are the same for all the groups as follows.

Definition 13 (Sufficiency) Given a group $G_b \subseteq \mathcal{D}$ and a predicted label $\hat{\ell}$, let $\mu_{G_b,\hat{\ell}} \in \mathbb{D}\mathsf{L}$ be the probability distribution of the actual class label ℓ when an input $v \in G_b$ is sampled from a test dataset w_{test} and the classifier C outputs the predicted label $\hat{\ell}$; i.e., for each $\ell \in \mathsf{L}$,

$$\mu_{G_b,\hat{\ell}}[\ell] \overset{\text{def}}{=} \Pr[\, H(v) = \ell \mid v \xleftarrow{\$} \sigma_{w_{\mathsf{test}}}(x), \quad v \in G_b, \quad \text{and} \quad C(v) = \hat{\ell} \,]. \tag{8}$$

A classifier C satisfies *sufficiency* between two groups G_0 and G_1 if $\mu_{G_0,\hat{\ell}} = \mu_{G_1,\hat{\ell}}$ holds for all $\hat{\ell} \in \mathsf{L}$.

Then this notion can be expressed using our extension of StatEL as follows.

Proposition 8 (Sufficiency) *The sufficiency between two groups G_0 and G_1 under a given test dataset w_{test} is expressed as $w_{\mathsf{test}} \models \mathsf{Sufficency}_0(x, y)$ where:*

$$\mathsf{Sufficency}_\varepsilon(x, y) \overset{def}{=} \bigwedge_{\hat{\ell} \in \mathsf{L}} \Big(\big(\eta_{G_0}(x) \wedge \psi_{\hat{\ell}}(x) \wedge h(x, y)\big) \sim_y^{\varepsilon, D_{\mathsf{tv}}} \big(\eta_{G_1}(x) \wedge \psi_{\hat{\ell}}(x) \wedge h(x, y)\big) \Big).$$

Proof Let $\hat{\ell} \in \mathsf{L}$ and $w_{b,\hat{\ell}} = w_{\mathsf{test}}|_{\eta_{G_b}(x) \wedge \psi_{\hat{\ell}}(x) \wedge h(x,y)}$. It follows from (8) that:

$$\mu_{G_b,\hat{\ell}}[\ell] = \Pr[\, \sigma_s(y) = \ell \mid s \xleftarrow{\$} w_{b,\hat{\ell}} \,],$$

hence $\mu_{G_b,\hat{\ell}} = \sigma_{w_{b,\hat{\ell}}}(y)$. Thus, by Definition 13, the sufficiency between G_0 and G_1 is given by $\sigma_{w_{0,\hat{\ell}}}(y) = \sigma_{w_{1,\hat{\ell}}}(y)$ for all $\hat{\ell} \in \mathsf{L}$. Therefore, this proposition follows from Proposition 1. □

It should be noted that for $\varepsilon > 0$, $\mathsf{Sufficency}_\varepsilon(x, y)$ represents a relaxation of sufficiency up to bias ε in terms of total variation D_{tv}.

8 Related Work

In this section, we provide a brief overview of related work on the specification of statistical machine learning and on epistemic logic for describing specification.

Desirable properties of statistical machine learning. There have been a large number of papers on attacks and defences for deep neural networks [40, 12]. Compared to them, however, not much work has been done to explore the formal specification of various properties of machine learning. Seshia et al. [38] present a list of desirable properties of DNNs (deep neural networks) although most of the properties are presented informally without mathematical formulas. As for robustness, Dreossi et al. [13] propose a unifying formalization of adversarial input generation in a rigorous and organized manner, although they formalize and classify attacks (as optimization problems) rather than define the robustness notions themselves.

Concerning the fairness notions, Barocas et al. [5] survey various fairness notions and classify them into the three categories: independence, separation, and sufficiency. Gajane [18] surveys the formalization of fairness notions for machine learning and present some justification based on social science literature.

Epistemic logic for describing specification. Epistemic logic [44] has been studied to represent and reason about knowledge and belief [16,22], and has been applied to describe various properties of distributed systems.

The *BAN logic* [8], proposed by Burrows, Abadi and Needham, is a notable example of epistemic logic used to model and verify the authentication in cryptographic protocols. To improve the formalization of protocols' behaviors, some epistemic approaches integrate process calculi [25,11].

Epistemic logic has also been used to formalize and reason about privacy properties, including anonymity [39,20,30], receipt-freeness of electronic voting protocols [26], and privacy policy for social network services [35]. Temporal epistemic logic is used to express information flow security policies [3].

Concerning the formalization of fairness notions, previous work in formal methods has modeled different kinds of fairness involving timing by using temporal logic rather than epistemic logic. As far as we know, no previous work has formalized fairness notions of machine learning by using modal logic.

Formalization of statistical properties. In studies of philosophical logic, Lewis [32] shows the idea that when a random value has various possible probability distributions, then those distributions should be represented on distinct possible worlds. Bana [4] puts Lewis's idea in a mathematically rigorous setting. Recently, a modal logic called statistical epistemic logic [28] is proposed and is used to formalize statistical hypothesis testing and the notion of differential privacy [14]. Independently of that work, French et al. [17] propose a probability model for a dynamic epistemic logic in which each world is associated with a subjective probability distribution over the universe, without dealing with non-deterministic inputs or statistical divergence.

9 Conclusion

In this paper we proposed an epistemic approach to the modeling of supervised learning and its desirable properties. Specifically, we employed a distributional Kripke model in which each possible world corresponds to a possible dataset and modal operators are interpreted as transformation and testing on datasets. Then we formalized various notions of the classification performance, robustness, and fairness of statistical classifiers by using our extension of statistical epistemic logic (StatEL). In this formalization, we clarified relationships among properties of classifiers, and relevance between classification performance and robustness.

We emphasize that this is the first attempt to use epistemic models and logical formulas to describe statistical properties of machine learning, and would be a starting point to develop theories of formal specification of machine learning.

In future work, we are planning to extend our framework to formally reason about system-level properties of learning-based systems. We are also interested in developing a more general framework for the formal specification of machine learning associated with testing methods. Our future work will also include an extension of StatEL to formalize unsupervised learning and reinforcement learning.

Acknowledgements I am grateful to Gergei Bana for his useful comments on part of a preliminary manuscript.

References

1. Angell, R., Johnson, B., Brun, Y., Meliou, A.: Themis: automatically testing software for discrimination. In: Proc. ESEC/SIGSOFT FSE, pp. 871–875. ACM (2018). DOI 10.1145/3236024.3264590
2. Athalye, A., Engstrom, L., Ilyas, A., Kwok, K.: Synthesizing robust adversarial examples. In: Proc. ICML, pp. 284–293 (2018)
3. Balliu, M., Dam, M., Guernic, G.L.: Epistemic temporal logic for information flow security. In: Proc. of PLAS, p. 6 (2011). DOI 10.1145/2166956.2166962
4. Bana, G.: Models of objective chance: An analysis through examples. In: Making it Formally Explicit, pp. 43–60. Springer International Publishing (2017). DOI 10.1007/978-3-319-55486-0_3
5. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning. fairmlbook.org (2019). `http://www.fairmlbook.org`
6. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research (2018). DOI 10.1177/0049124118782533
7. Blackburn, P., de Rijke, M., Venema, Y.: Modal Logic. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press (2001). DOI 10.1017/CBO9781107050884
8. Burrows, M., Abadi, M., Needham, R.M.: A logic of authentication. ACM Trans. Comput. Syst. **8**(1), 18–36 (1990). DOI 10.1145/77648.77649
9. Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. Data Min. Knowl. Discov. **21**(2), 277–292 (2010). DOI 10.1007/s10618-010-0190-x
10. Carlini, N., Wagner, D.A.: Towards evaluating the robustness of neural networks. In: Prc. S&P, pp. 39–57 (2017). DOI 10.1109/SP.2017.49
11. Chadha, R., Delaune, S., Kremer, S.: Epistemic logic for the applied pi calculus. In: Proc. of FMOODS/FORTE, pp. 182–197 (2009). DOI 10.1007/978-3-642-02138-1_12
12. Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., Mukhopadhyay, D.: Adversarial attacks and defences: A survey. CoRR **abs/1810.00069** (2018). URL `http://arxiv.org/abs/1810.00069`
13. Dreossi, T., Ghosh, S., Sangiovanni-Vincentelli, A.L., Seshia, S.A.: A formalization of robustness for deep neural networks. In: Proc. VNN (2019)
14. Dwork, C.: Differential privacy. In: Proc. of ICALP, pp. 1–12 (2006)
15. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.S.: Fairness through awareness. In: Proc. of ITCS, pp. 214–226. ACM (2012)
16. Fagin, R., Halpern, J., Moses, Y., Vardi, M.: Reasoning about Knowledge. The MIT Press (1995)
17. French, T., Gozzard, A., Reynolds, M.: Dynamic aleatoric reasoning in games of bluffing and chance. In: Proc. AAMAS, pp. 1964–1966 (2019)
18. Gajane, P.: On formalizing fairness in prediction with machine learning. CoRR **abs/1710.03184** (2017). URL `http://arxiv.org/abs/1710.03184`
19. Galhotra, S., Brun, Y., Meliou, A.: Fairness testing: testing software for discrimination. In: Proc. ESEC/FSE, pp. 498–510. ACM (2017). DOI 10.1145/3106237.3106277
20. Garcia, F.D., Hasuo, I., Pieters, W., van Rossum, P.: Provable anonymity. In: Proc. of FMSE, pp. 63–72 (2005). DOI 10.1145/1103576.1103585
21. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Proc. of ICLR (2015)
22. Halpern, J.Y.: Reasoning about uncertainty. The MIT press (2003)
23. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: proc. NIPS, pp. 3315–3323 (2016)
24. Huang, X., Kwiatkowska, M., Wang, S., Wu, M.: Safety verification of deep neural networks. In: Proc. CAV, pp. 3–29 (2017). DOI 10.1007/978-3-319-63387-9_1
25. Hughes, D., Shmatikov, V.: Information hiding, anonymity and privacy: a modular approach. J. of Comp. Security **12**(1), 3–36 (2004)
26. Jonker, H.L., Pieters, W.: Receipt-freeness as a special case of anonymity in epistemic logic. In: Proc. Workshop On Trustworthy Elections (WOTE'06) (2006)
27. Katz, G., Barrett, C.W., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: An efficient SMT solver for verifying deep neural networks. In: Proc. CAV, pp. 97–117 (2017). DOI 10.1007/978-3-319-63387-9_5

28. Kawamoto, Y.: Statistical epistemic logic. In: The Art of Modelling Computational Systems: A Journey from Logic and Concurrency to Security and Privacy - Essays Dedicated to Catuscia Palamidessi on the Occasion of Her 60th Birthday, *LNCS*, vol. 11760, pp. 344–362. Springer (2019). DOI 10.1007/978-3-030-31175-9_20
29. Kawamoto, Y.: Towards logical specification of statistical machine learning. In: Proc. SEFM, *LNCS*, vol. 11724, pp. 293–311. Springer (2019). DOI 10.1007/978-3-030-30446-1_16
30. Kawamoto, Y., Mano, K., Sakurada, H., Hagiya, M.: Partial knowledge of functions and verification of anonymity. Transactions of the Japan Society for Industrial and Applied Mathematics **17**(4), 559–576 (2007). DOI 10.11540/jsiamt.17.4_559
31. Kripke, S.A.: Semantical analysis of modal logic i normal modal propositional calculi. Mathematical Logic Quarterly **9**(5-6), 67–96 (1963)
32. Lewis, D.: A subjectivist's guide to objective chance. In: Studies in Inductive Logic and Probability, Volume II, pp. 263–293. Berkeley: University of California Press (1980)
33. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: Proc. ICLR (2018)
34. Moosavi-Dezfooli, S., Fawzi, A., Frossard, P.: Deepfool: A simple and accurate method to fool deep neural networks. In: Proc. CVPR, pp. 2574–2582 (2016). DOI 10.1109/CVPR.2016.282
35. Pardo, R., Schneider, G.: A formal privacy policy framework for social networks. In: Proc. SEFM, pp. 378–392 (2014). DOI 10.1007/978-3-319-10431-7_30
36. Pei, K., Cao, Y., Yang, J., Jana, S.: Deepxplore: Automated whitebox testing of deep learning systems. In: Proc. SOSP, pp. 1–18 (2017). DOI 10.1145/3132747.3132785
37. Prior, A.N.: Time and modality (1957)
38. Seshia, S.A., Desai, A., Dreossi, T., Fremont, D.J., Ghosh, S., Kim, E., Shivakumar, S., Vazquez-Chanlatte, M., Yue, X.: Formal specification for deep neural networks. In: Proc. ATVA, pp. 20–34 (2018). DOI 10.1007/978-3-030-01090-4_2
39. Syverson, P.F., Stubblebine, S.G.: Group principals and the formalization of anonymity. In: World Congress on Formal Methods (1), pp. 814–833 (1999). DOI 10.1007/3-540-48119-2_45
40. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: Proc. ICLR (2014)
41. Tian, Y., Pei, K., Jana, S., Ray, B.: Deeptest: automated testing of deep-neural-network-driven autonomous cars. In: Proc. ICSE, pp. 303–314 (2018). DOI 10.1145/3180155.3180220
42. Udeshi, S., Arora, P., Chattopadhyay, S.: Automated directed fairness testing. In: Proc. ASE, pp. 98–108. ACM (2018). DOI 10.1145/3238147.3238165
43. Vaserstein, L.: Markovian processes on countable space product describing large systems of automata. Probl. Peredachi Inf. **5**(3), 64–72 (1969)
44. von Wright, G.H.: An Essay in Modal Logic. Amsterdam: North-Holland Pub. Co. (1951)
45. Zafar, M.B., Valera, I., Gomez-Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proc. WWW, pp. 1171–1180 (2017). DOI 10.1145/3038912.3052660