2 Probably Approximate Truth

In this chapter, in Section 2.1 we first recall the classical semantics used to interpret propositional formulas, and then in Section 2.2, the fundamental problems that arise when seeking to gain knowledge in this classical sense by generalizing from specific examples. We then discuss a way in which we can appropriately "relax" our standards for knowledge to accommodate inductive generalization from examples in Section 2.4, motivated by considering statistical inference in Section 2.3. We will use this relaxed, *"Probably Approximately Correct"* semantics for the rest of the book.

2.1 The classical "possible worlds" semantics.

Knowledge is classically viewed as a collection of true statements about the real world. Breaking this down into components, we have statements that may be true or false – we call these *Boolean* values, and denote them respectively by 1 and 0 here – and our knowledge is a collection of these statements that we take to be true. Statements that have truth values are known as *propositions*.

In applications, we are usually only interested in a small part of the world called a *domain*, that is described by a finite set of propositions that we refer to as *attributes*. We'll generally use n to denote the number of attributes of the domain under consideration. The possible truth values of these n propositions gives a set of *possible worlds* $\mathcal{X}$ that are identified with the truth value settings, i.e., a Boolean valued vector of n dimensions.

As a concrete example, let's suppose we are considering a robot that moves a box between rooms joined by a hallway. We can describe the state of this limited domain by the following seven propositions:

1. "The robot is in the hallway"
2. "The robot is in the first room"
3. "The robot is in the second room"
4. "The box is in the hallway"
5. "The box is in the first room"
6. "The box is in the second room"
7. "The robot is holding the box"

For brevity, we'll refer to these as $X_1, X_2, \ldots, X_7$. Since we are supposing that there is only one box and only one robot, clearly not all of the "possible worlds" are actually possible—exactly one of X_1, X_2, and X_3 is true and exactly one of X_4, X_5, and X_6 is true. Likewise, if the robot is holding the box, then the robot and the box must be in the same place. We'll let $\mathcal{D}$ denote the actual set of possible worlds in this domain. There are twelve of these: three in which the robot is holding the box in one of the three respective locations, and then nine worlds in which the robot is not holding the box, and the robot and the box can each be in any of the three locations. This is far fewer than the $2^7 = 128$ "possible worlds" given by the possible Boolean values that seven propositions could take.

These propositions only describe the state of the world at a given moment in time: if the robot moves from one location to another, the truth values of these propositions will change. This is in contrast to propositions that are simply true or false such as "$2^7 = 128$" (true) or "I witnessed the dawn of time" (false). Propositions whose values may vary over time are sometimes called *fluents*. But we may also have a collection of statements about individual events or entities that may be true or false for the various events or entities under consideration: for example, respondents to a survey may answer questions such as "do you smoke?" or "is your annual income greater than $50,000?" and this collection of answers will be taken to describe the respondent for our purposes. Or, a record of sales made by a business may be similarly described by a collection of attributes. Again, we have a set of actual possibilities $\mathcal{D}$ for the domain that is a subset of the possible truth values $\mathcal{X}$.

We can also consider *properties* that may hold in the domain beyond the set of propositions under consideration. For example, in our robot example, the property that the robot is in the same location as the box but not holding it. This property is nevertheless determined by the truth values of the propositional variables: like the set of actual possibilities, we can identify it with a set of possible worlds $C \subset \mathcal{X}$. Now, this property C may or may not be true: there are worlds that belong to both C and $\mathcal{D}$ – say if both are in the first room and the robot is indeed not holding the box – but there are also worlds where, say, the robot is actually holding the box, that belong to $\mathcal{D}$ but not C. By contrast, as we noted earlier, the property that "either the robot is not holding the box or the robot and the box are in the same location" (that we'll call C') is true in all of the actual possible worlds: $C' \supseteq \mathcal{D}$. We say that C' is *valid* for $\mathcal{D}$ (true in the domain), but C is not.

2.2 Hume's challenge to induction.

In *learning*, we may be interested in determining whether or not a particular property C is valid or not in our domain $\mathcal{D}$. If C is a property of *all* of the possible worlds, i.e., $C = \mathcal{X}$, then since $\mathcal{D}$ is a set of possible worlds, $\mathcal{D} \subseteq \mathcal{X}$, we have $\mathcal{D} \subseteq C$ and so C is trivially valid for $\mathcal{D}$.

But now suppose that C is not true of all of the possible worlds. A natural approach to determine if C in fact holds is to observe various *example* states of the world – elements

of $\mathcal{D}$ – and seeing if these confirm C. If C indeed holds in all of these examples, we might seek to make an *inductive generalization* of C to be true in all of the worlds in $\mathcal{D}$.

There is an obvious difficulty with such an approach, noted by the philosopher David Hume: namely, C being true in some worlds of $\mathcal{D}$ tells us nothing about whether or not it is true in other worlds. In other words, until we have observed the entirety of $\mathcal{D}$, it is still possible that $\mathcal{D}$ contains possible worlds in which C fails to hold. We have no grounds for making such an inductive leap from the examples to all of $\mathcal{D}$. Induction is simply not a sound principle for acquiring knowledge in the classical sense.

2.3 Inference from surveys.

Hume's observation tells us that only *tautologies* – statements that are true in all possible worlds – can be known to be valid with certainty. But, there are ways of determining that statements are true in *most* possible worlds by only examining a few examples. Surveys concretely illustrate how this is possible.

Indeed, let's suppose for example that we want to establish that *most* households in San Francisco either rent or have an annual household income of more than $300,000. Suppose we send 20 surveys out to households independently chosen uniformly at random from the population, and we manage to get responses to all of them; plausibly, suppose that they all either rent or meet the income threshold. It turns out that from this outcome, we can conclude that it is unlikely that more than 10% of households both do not rent and have less than $300,000 annual income.

How do we see this? Recall that we say that a collection of events $A_1, \ldots, A_m$ are *mutually independent* if the probability that they all hold simultaneously is equal to the product of the probabilities of the individual events, i.e.,

$$\Pr[A_1 \text{ and } A_2 \text{ and } \ldots \text{ and } A_m] = \prod_{i=1}^{m} \Pr[A_i].$$

In particular, let's consider the event A_i that the ith survey went to a household that meets our criteria. Since we suppose that each survey was sent to an independent, uniform random household, the probability that each given ith household meets the criteria is the fraction of such households. What if this were less than 90%? Then the probability that all 20 surveys would come back meeting our criteria would be less than

$$0.9^{20} \approx 0.11$$

or in other words, we would have to have been rather unlucky to fail to encounter a household that does not rent and has less than $300,000 annual income. We usually state this as: with 89% confidence, at least 90% of households meet the criteria. The claim that "households either rent or have more than $300,000 in annual income" is Probably Approximately Correct.

What may be surprising about this calculation is that the actual size of the population does not enter anywhere. No matter how large it is, the same calculation can be carried

out, and we can conclude that we only need 20 responses to obtain 89% confidence about 90% of the population. This is in contrast to establishing that *all* members of the population satisfy some criteria, which we observed would require investigating the entire population.

2.4 Approximate validity.

Our survey did not establish that "households either rent or have more than $300,000 in annual income" is valid for the set of actual possible worlds, only that it was true of *most* of the possible worlds. For the purposes of our survey, we chose a respondent uniformly at random from the population, and so the guarantee we have corresponds to a proportion of the population.

But, suppose we had represented the responses to the survey in terms of attributes X_1 = "household rents", X_2 = "more than $10,000 annual income", X_3 = "more than $20,000 annual income", X_4 = "more than $30,000 annual income", and so on. With respect to the answers to these questions, certain relationships necessarily hold: for example, if X_2 is true, X_3 and X_4 must also be true. X_4 is presumably thus somewhat less likely to be true than X_2, and so on. Thus, if we are thinking of our possible worlds in terms of the answers we receive on such a survey, even though we selected a member of the population uniformly at random, the distribution on possible worlds given by survey responses that we observe is *not* uniformly distributed over the responses. Indeed, if it were, the relationships we described among attributes $X_2, X_3, \ldots$ would generally not hold. These relationships are captured by some classically valid properties on the possible worlds captured by the survey.

We stress, moreover, that we don't expect the distribution on the survey responses to simply be uniformly distributed over the set of actual possible worlds subject to these classically valid relationships. For example, there may exist some households with low income that nevertheless own their home; but the results of our survey suggest that these are rare. So even though there may exist households with all possible income levels that own a home, the distribution on responses that we observe may have the probability of owning a home be markedly higher for the responses with higher income levels. By contrast, if we simply had a uniform distribution on possible worlds, 50% of the responses at each income level would rent, since that is one of the two possibilities for each level.

In the framework we study here, we therefore make *no* assumptions on the distribution on possible worlds. We take the stance that the attributes describing the possible worlds may obey certain relationships, and it is such relationships we hope to learn from observations. We capture these relationships in general with the following notion of approximate validity:

Definition 2.1 Let $\mathcal{D}$ be a probability distribution on possible worlds $\mathcal{X}$. We say that a property $C \subseteq \mathcal{X}$ is $1 - \epsilon$-*valid* if the probability that C holds of a possible

world drawn from $\mathcal{D}$ is at least $1 - \epsilon$. We denote the probability of C holding of a possible world $\vec{x} \in \mathcal{X}$ sampled from $\mathcal{D}$ by $\Pr_{\vec{x} \in \mathcal{D}}[C(\vec{x})]$, so we can equivalently write $\Pr_{\vec{x} \in \mathcal{D}}[C(\vec{x})] \geq 1 - \epsilon$

When we study learning, as with our surveys, we will generally suppose that we have sampled many possible worlds $\vec{x}^{(1)}, \vec{x}^{(2)}, \ldots, \vec{x}^{(m)}$ independently from the same unknown probability distribution $\mathcal{D}$. The situation is typically described as: $\vec{x}^{(1)}, \vec{x}^{(2)}, \ldots, \vec{x}^{(m)}$ are independent and identically distributed, or "i.i.d." This definition of approximate validity captures what we hope to obtain from such example observations. As illustrated by our survey example, in this terminology, with 89% confidence, the property C ="households either rent or have more than \$300,000 in annual income" is 90%-valid for $\mathcal{D}$.

We stress once again, even though two of the examples $\vec{x}^{(i)}$ and $\vec{x}^{(j)}$ are independently sampled from $\mathcal{D}$, it does not mean that the attributes x_i and x_j within a single draw $\vec{x}$ from $\mathcal{D}$ are independently distributed. We generally assume i.i.d. examples from an arbitrary distribution $\mathcal{D}$, that may not feature any independence relationships among the attributes.

Next, let's return to the example of the box-moving robot from Section 2.1. Let's suppose that we are recording the state of the environment as the robot is set to various tasks, to move the box from one room to another, move about the environment and so on. Our possible worlds correspond to "snapshots" of the environment that capture its current configuration at a somewhat low level of detail, where we ignore the precise location of the robot and box within the rooms and so on. During the collection of this data, the state of the environment between two consecutive snapshots is *not* independent. Nevertheless, over the long run, we can consider a distribution $\mathcal{D}_{\text{time avg.}}$ on settings of the possible worlds, where each configuration has a probability corresponding to the fraction of the snapshots featuring that precise configuration. Equivalently, we can think of the possible worlds as being sampled by choosing a point in time at random from the recordings, and reporting the configuration at that time. Approximate validity here captures the fraction of recorded points in time in which some property was true of the scene.

Again, this distribution on worlds features some rules that we may wish to learn, for example: when the robot is holding the box, the robot and the box are in the same room. That is, in our attribute encoding, either X_7 is false, or X_1 and X_4 are both true, or X_2 and X_5 are both true, or X_3 and X_6 are both true. But, because we are choosing points in time to observe independently from one another, we cannot learn the rules governing how the environment changes from this distribution. Our notion of approximate validity for $\mathcal{D}_{\text{time avg.}}$ only captures the relationships that hold among the attributes at single points in time.

To study how the environment changes, at minimum we need to be able to formulate statements about its state at two points in time. For example, we might have *two* copies of each attribute, one labeled "pre" and the other labeled "post," which, respectively, are the set of attributes describing two consecutive snapshots. Our "possible worlds" now have $2 \times 7 = 14$ attributes. With this extended formulation, we can now consider

properties such as that the in the pre- and post-states, the robot and box may only be in certain pairs of locations (that are adjacent), for example, $X_{2,\text{pre}}$ is false or $X_{2,\text{post}}$ or $X_{1,\text{post}}$ are true—meaning, if the robot was in the first room, it may subsequently only be in the first room or the hallway. Similarly, if $X_{7,\text{pre}}$ is false (the robot was not holding the box) and $X_{4,\text{pre}}$ is true (the box was in the hallway) than $X_{4,\text{post}}$ is true (the box will remain in the hallway).

The natural distribution corresponding to this formulation on pairs of consecutive states is to generalize $\mathcal{D}_{\text{time avg.}}$ to include two consecutive states, rather than just one: for the distribution $\mathcal{D}_{\text{pre/post}}$ we choose any point in time except for the final point uniformly at random, and report the values describing the chosen state in the $X_{i,\text{pre}}$ attributes and the values describing the following state in the $X_{i,\text{post}}$ attributes. We remark that this distribution now depends not only on the frequency with which states occurred, but also on the frequency with which the robot took various actions in the different states. If the robot is acting completely at random, this may make little difference, but if the robot was only ever tasked with picking up boxes in the first room and delivering them to the second room, then the observed examples may satisfy properties such as "if the robot was holding the box in the first room, the robot will be holding the box in the post state" or "if the robot is not holding the box in the second room, the robot will not be holding the box in the post state," even if the robot is actually capable of such things. Indeed, the examples we obtain in the environment in this case are indistinguishable from examples of an environment in which the robot is actually not capable of such actions. This is an inherent limitation of the "passive" learning model we consider.

Although this formulation only captures a single transition, we can draw conclusions about a sequence of several steps by using the *union bound*:

Proposition 2.2 (Union bound) *Let A_1 and A_2 be two events. The probability that A_1 or A_2 holds satisfies*

$$\Pr[A_1 \text{ or } A_2] \leq \Pr[A_1] + \Pr[A_2].$$

More generally, for any collection of events $A_1, \ldots, A_m$,

$$\Pr[A_1 \text{ or } A_2 \text{ or } \cdots \text{ or } A_m] \leq \sum_{i=1}^{m} \Pr[A_i].$$

Recall that the union bound holds because the expression on the left counts the contribution of the probability of each possible world to the union exactly once, whereas on the right it counts each possible world in the union *at least* once, and probabilities are always nonnegative.

To apply the union bound correctly here, we need to first consider a probability distribution on longer sequences of actions, so that we can properly define an event "A_1 or A_2 or …or A_m" that refers to a long sequence of points in time. Here each A_i only refers to the ith pair from the sequences of states, and it represents the event that at least one of the rules we're using to describe the environment's dynamics *does not hold* – i.e., its probability is the ϵ "gap" in our approximate validity for this

collection of rules. For any number of steps T, we can consider the distribution $\mathcal{D}_T$ on sequences of T states, represented by attributes indexed additionally by $t = 1, 2, \ldots, T$, e.g., $X_{1,1}, X_{1,2}, \ldots, X_{1,T}$. Suppose that each A_i refers only to the attributes for some time indices t and $t + 1$. Then the union bound tells us that for our collection of events, $\Pr_{\mathcal{D}_T}[A_1 \text{ or } A_2 \text{ or } \cdots \text{ or } A_m] \leq \sum_{i=1}^{m} \Pr_{\mathcal{D}_T}[A_i]$. Now, if we have enough snapshots relative to T, $\Pr_{\mathcal{D}_T}[A_i] \approx \Pr_{\mathcal{D}_{\text{pre/post}}}[A_i]$, since the only difference is that $\Pr_{\mathcal{D}_{\text{pre/post}}}[A_i]$ considers all pairs of consecutive points, whereas $\Pr_{\mathcal{D}_T}[A_i]$ only considers the ith pair from sequences of states up to the final T recorded snapshots. The total difference between the probabilities is at most $T/\#$ snapshots recorded. We can then use $\sum_{i=1}^{m} \Pr_{\mathcal{D}_{\text{pre/post}}}[A_i]$(+$m \cdot T/\#$ snapshots recorded) as an estimate for $\Pr_{\mathcal{D}_T}[A_1 \text{ or } A_2 \text{ or } \cdots \text{ or } A_m]$. By collecting enough data, we can make this approximation as accurate as we require.

To summarize, we have seen that even though classical properties cannot be inferred through anything less than an exhaustive consideration of all possible worlds, we can infer the approximate validity of a property with high confidence by examining example possible worlds that have been sampled from a probability distribution. We can therefore obtain a simple kind of inductive generalization in this framework, that will be our focus of study. Even though the collection the possible worlds and the application of the knowledge we obtain is generally for worlds that are *not* "independent and identically distributed" (henceforth, *"i.i.d."*), the assumption that our data consists of i.i.d. examples is relatively benign—provided that we make no a priori assumptions about the probabilities of the possible worlds in our formulation. We stress that this agnostic stance on the world's structure will allow us to apply our framework to complex worlds that may feature only subtle forms of structure, that we were not aware of a priori. Although the passive learning framework does impose some significant restrictions on what we can accomplish, it nevertheless is faithful to many realistic scenarios in which, at least at first, we only have the ability to observe a system of interest.

Chapter notes

Propositional logic, as we recall here, is ancient, originating in Aristotle's *Prior Analytics* (ca. 350 BC). The more modern framing of logic as the study of the validity of inferences in an abstracted formal language can be traced to De Morgan (1847). Hume's objection to inductive generalization is of course due to Hume (1748). "Probably approximately correct" semantics is an example of what statisticians call a "confidence interval": that is, we are formally claiming the probability that the statement in question holds lies in the interval $[1 - \epsilon, 1]$ with confidence $1 - \delta$ (or with "p-value"/"significance" δ); our example inference here is a particularly simple special case of this framework that was first rigorously treated by Bernoulli (1713). It is challenging to assign credit for the development of these ideas because they were largely communicated orally in this period. Hacking (1975) provides a thorough broader history of early work on probability, that encompasses such inferences.

Many philosophers desired a kind of logic that could capture inductive inferences, based on probabilistic reasoning. This topic was later a major focus of logical positivists, see for example Carnap's work (Carnap, 1950). The prior approach that is most directly similar to what we consider is Adams' ϵ-semantics (Adams, 1975): the difference is that for Adams, ϵ was generally infinitesimal so that countable union bounds again yield an infinitesimally small ϵ. Adams, unlike us, thus had no need to track the magnitude of ϵ; unfortunately for the proposal, there is no way to empirically reach inferences with such infinitesimally small ϵ. In the appendix, we review the amount of data required to reach a given bound on ϵ, and we see that it grows polynomially with $1/\epsilon$—and hence, an infinitesimally small ϵ cannot be reached by observing any finite number of examples. The version we present here, where ϵ is nonnegligible, is from Valiant (2000b).

Exercise 2.1 is based on an example used by McCarthy.

Exercises

2.1 *Setting up attribute encodings of a domain.* Consider a simplified car-diagnosis domain, where we can observe the status of the battery, the fuel tank, the ignition key, as well as whether the engine is running, and whether there is a potato in the tailpipe.
(a) Give a Boolean attribute encoding for this domain. Your encoding should use at least six attributes, and each attribute should have a clear informal meaning.
(b) If your encoding uses n attributes, how many Boolean vectors are there in the full space $X = \{0, 1\}^n$?
(c) Describe at least three constraints that distinguish the actual possible worlds $D \subseteq X$ from arbitrary Boolean vectors in X. These constraints may express physical, mechanical, or diagnostic relationships among the attributes.
(d) Write one property $C \subseteq X$ that is classically valid for the actual domain D, but not true in all of X. Write this property both informally and as a propositional formula over your attributes.
(e) Write one property $C' \subseteq X$ that is sometimes true and sometimes false in the actual domain D. Again, write it both informally and as a propositional formula.
(f) Suppose examples are sampled from a distribution D_{obs} over diagnostic states of cars. Explain how the property

> "If the battery is charged, there is fuel, and the ignition key is turned on,
> then the engine is running"

could be $1 - \epsilon$-valid with respect to D_{obs}, even though it is not classically valid for D. In particular, explain how rare abnormal conditions, such as a potato in the tailpipe, can make the property fail in some actual possible worlds.

2.2 *Inference with counterexamples.* In Section 2.3 we saw that by surveying just 20 members of any size population, we could get $\approx 89\%$ confidence that a claim was true of at least 90% of the population. But, this calculation relied on the *absence* of any counterexamples to the claim among the 20 surveyed; indeed, if the claim were actually true of, say, 95% of the population, then we are somewhat likely to find such

a counterexample within a sample of size 20, and this calculation would do us no good. In such a case, we can instead use Chernoff/Hoeffding bounds: for independent $X_1, \ldots, X_m \in [0,1]$ and any $\gamma \in [0,1]$,

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m} X_i \geq \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} X_i\right] + \gamma\right] \leq e^{-2m\gamma^2} \qquad \text{Hoeffding (additive)}$$

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m} X_i \geq (1+\gamma)\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} X_i\right]\right] \leq e^{-\frac{1}{3}\mathbb{E}[\sum_{i=1}^{m} X_i]\gamma^2} \qquad \text{Chernoff upper bound}$$

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m} X_i \leq (1-\gamma)\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} X_i\right]\right] \leq e^{-\frac{1}{2}\mathbb{E}[\sum_{i=1}^{m} X_i]\gamma^2} \qquad \text{Chernoff lower bound}$$

but this comes at a price. Use the additive form of the bound to determine a sample size such that by counting the fraction of counterexamples to the claim, with 89% confidence, we obtain an estimate of the true fraction of the population for which the claim holds to within an additive 10%.

2.3 *Classical entailment in PAC-Semantics.* In this problem, we will relate classical entailment to entailment in PAC-Semantics.
(a) Let $A \subseteq \mathcal{X}$ be any property. Suppose we know that A is $(1-\epsilon)$-valid with respect to D, i.e., $\Pr_{x\in D}[A(x) = 1] \geq 1-\epsilon$ and that $A \models \varphi$ for some representation $\varphi(x)$. Prove that φ is therefore also $(1-\epsilon)$-valid with respect to D.
(b) Now suppose that for representations $\psi_1(x), \ldots, \psi_k(x)$, we know that each $\psi_i(x)$ is $(1-\epsilon_i)$-valid with respect to a common distribution D for some $\epsilon_i \in [0,1]$, and for the knowledge base $KB = \{\psi_1, \ldots, \psi_k\}$, $KB \models \varphi$ for some representation $\varphi(x)$. Prove that then $\varphi(x)$ is $(1-\epsilon')$-valid with respect to D where $\epsilon' = \sum_{i=1}^{k} \epsilon_i$.
(c) Finally, suppose that for $KB = \{\psi_1, \ldots, \psi_k\}$ where $\psi_1, \ldots, \psi_k$ are consistent, $\sum_{i=1}^{k} \epsilon_i = \epsilon' < 1$ and for each i, ψ_i is *not* entailed by $KB - \{\psi_i\}$. Show then that there *exists* a distribution D on $\mathcal{X}$ under which, for $\varphi = \psi_1 \wedge \cdots \wedge \psi_k$,

$$\Pr_{x\in D}[\psi_i(x) = 1] = 1-\epsilon_i \text{ for each } i = 1, \ldots, k \qquad \text{and}$$
$$\Pr_{x\in D}[\varphi(x) = 1] = 1-\epsilon'.$$

Thus: classical reasoning can be used freely in PAC-Semantics, and the "error" of the conclusions may depend additively on the error in the premises in general.