2 Projection and Projectability

David Corfield

The problem of dataset shift can be viewed in the light of the more general problem of induction, in particular the question of what it is about some objects' features or properties which allow us to project correlations confidently to other times and places. We explore the varieties of background knowledge which philosophers have taken to warrant this confidence. Finally, we ask whether or not Bayesians have been inhibited by their need to encode such forms of background knowledge in the shape of probability distributions.

2.1 Introduction

Philosophers have not directly addressed the problems of *dataset shift* or *covariate shift*, but, as I shall argue, these problems are closely related to ones they have discussed. Indeed, much of the philosophical literature dealing with inductive reasoning has some bearing on these problems. For example, the "projection" of a regularity to times or places beyond those of the observed supporting instances involves a covariate shift when time or space is regarded as an independent variable.

What philosophers of science have found to be key in inductive reasoning is background knowledge. In the natural sciences this forms a highly complex web of practical and theoretic knowledge, and many have decided that little more than a qualitative description of plausible inference is possible. In the case of machine learning we still have to deal with background knowledge, or else we shouldn't be able to do any learning at all. However, this knowledge is typically simpler and thus more amenable to encoding in a learning algorithm. A point I shall consider in this chapter is whether Bayesians help or hinder themselves by requiring this knowledge to be construed in probabilistic terms.

In order to relate the thematic problem of this book to that of background knowledge, it will be convenient to establish what precisely is meant by the definition of dataset shift as a situation where the joint distribution of inputs and outputs differs between the training and test stages. What does it mean to say that the training data and test data come from a different distribution? In the first section we shall need to consider two words in this question – “data” and “distribution”.

2.2 Data and Its Distributions

Despite its etymology, data is not just something given to us. Rather, it is the result of the framing and recording of some interaction with the world. Those famous handwritten digits of the MNIST dataset need not have been collected in the way they were. They never had to have been photographed, nor once photographed to be pixellated, nor if pixellated to be pixellated the way they were. Nor did they have to be centered or size-normalized. The time and place a digit was written, who wrote it, and the type of writing implement used need not have been forgotten. But for convenience, and in the hope that sufficient usable information was captured in the representation, decisions were taken to do so. It is important to remember that decisions of this kind do not take place in a vacuum. They arise on the basis of certain expectations and knowledge. We shall need to keep the existence of this background firmly in mind.

Now, given some data-collecting protocol, what does it mean to say that the vectors of chosen attributes of the entities of a sample come from a distribution, which may be compared to another distribution generating a different sample? Does this not presuppose some relatively stable random process, like a coin-tossing device whose settings don’t change in important ways while a sample is generated? Could the MNIST dataset collection be viewed like this?

Consider the question I heard posed by Peter Grünwald as to whether we should take the writings of Tolstoy to be generated according to a probability distribution. Certainly we can use the apparatus of probability theory to model his writings, and perhaps to provide evidence whether or not some newly discovered work was written by him. We may choose to model *War and Peace* by a Markov chain of order two, then come to realize that one of order three is more accurate, but that little is gained by extending this to order four. But what does it mean to say we have the true distribution, or have come close to it?

Even those canonical coin toss samples are not viewed universally as the product of a random generating process. Certainly some prominent thinkers have had difficulty with the idea of randomness and probability distributions being out there in the world. Indeed, for people like Bruno de Finetti and Edwin Jaynes, there just is no randomness in the world, certainly not at the classical level; distributions are simply representations of our state of belief.

But de Finetti would need not have had to excuse himself politely from this book had he been asked to contribute. He could imagine himself in a situation where he would take each of the training data and test data individually as exchangeable sequences, but not both datasets taken together. Something we might seek, in his language, is a way to map these datasets such that the joint image dataset *is* exchangeable. I shall keep to the standard way of speaking throughout the chapter, as though distributions are things belonging to the world, bearing in mind that there are ways of rephrasing matters to keep the de Finettian happy.

2.3 Data Attributes and Projection

There are many reasons why training data might have a different distribution from test data. As I said in the previous section, the data has already been chosen to take on a specific form. Perhaps we included a variable which made no difference to the output, but in whose presence we detect covariate shift. It is possible then that with a better choice of variables, and a projection onto them, the shift would disappear.

Viewing matters the other way around, had a different choice been made and had we been less restrictive, what was not a case of dataset or covariate shift becomes one. In other words, if the term "input variable" were taken broadly enough, all learning confronts the problem of covariate shift, as we may consider in the case of, say, classification,

$$\text{Pr (class | features, place of observation, time of observation,...).}$$

Sometimes it is obvious that some of the dimensions of the input space are irrelevant. Usually we ignore them without thinking about it. If we wish to model the running speed of individuals we may consider their age, height, and weight, but we typically won't think to count the number of whirls on the fingerprint of their left ring finger. But we can consider that already a projection has taken place onto the chosen set of input variables.

Leaving major projections of time and space to one side, imagine that on a digit recognition task I find the top left pixel doesn't matter to the classification. Yet the training data has 80% dark pixels in that position, while the test data has 20% dark pixels there. If the projected training distribution matches the projected test distribution, won't we proceed happily, unless we suspect a relevant reason for the incidence of this pixel's darkness?

Or, I want to predict people's preference for a film Y. In my training data I have an input distribution skewed to the extremes for the likes/dislikes of another film, X, but these are not found to be significant for the assessment of film Y. Now in the test data, many people are indifferent to film X. Do I confidently project out from that film preference feature?

Clearly our confidence must rely on background knowledge. Set the same problem, i.e., exactly the same numerical values, in an artificial setting with meaningless predicates, you will be much less confident. This must be due to prior beliefs about the length scales in such a direction, which could be encoded in an algorithm such as Gaussian process classification.

Dataset shift problems may come about from mistakenly ignoring some feature of the collection of data. Had we included this feature we would just have the covariate shift problem. Of course we will have to model this extra dimension somehow. If all we have is the extra tag of either being in the test sample or training sample, then if we devise a kernel which makes differently tagged members be seen as far apart in feature space, we won't be able to learn from the training data. If X is the input space and $P(X)$ differs for test and training samples, then we're looking to find a function $f : X \to X'$, such that the induced $P(X')$ is sufficiently similar for test and training data. But clearly the closeness of projected distributions is not sufficient for a projection to be counted as good, otherwise a mapping with X' a single point might qualify. We would also like $P(Y|X = x)$ to be reasonably uniform in the neighborhoods of training points with the same image in X'. Furthermore, we would need to control the class of such mappings, f, or, in other words, we would need to regularize the projection algorithm.

How much evidence we require to convince ourselves that we have a good projection naturally depends upon our prior expectations. And indeed in the whole process of inductive inference, background knowledge is vital. Even though I had never been to Canada before the 2006 conference on neural information processing systems, I had a good idea whether and when it mattered. I expected the snow to be like it is Europe, but the number of times per hour I heard "dude" to be different. Now the term philosophers have used for the activity of extending the domain of some observed regularity is precisely the same as the one I have been using, namely, *projection*. This is no mere coincidence.

2.4 The New Riddle of Induction

Nelson Goodman [1955] devised the "new riddle of induction" in the 1950s in the heyday of logical empiricism, when philosophers tried to formulate scientific reasoning with the resources of predicate logic and a Bayesian inductive logic. What he was challenging was a position that all could be achieved merely syntactically, that is, in terms of the logical form of the relevant statements. He famously presented the riddle by way of a paradox.

Grue Paradox Suppose that at time t we have observed many emeralds to be green. We thus have evidence statements

emerald a is green,
emerald b is green,
etc.

and these statements support the generalization

All emeralds are green.

But now define the predicate "grue" to apply to all things observed before t just in the case that they are green, and to other things just in the case that they are blue. Then we have also the evidence statements

emerald a is grue,
emerald b is grue,
etc.

and these evidence statements support the hypothesis

All emeralds are grue.

Hence the same observations support incompatible hypotheses about emeralds to be observed for the first time in the future after t; that they will be green and that they will be blue. What is it about green, Goodman asks, which warrants our projecting it to as yet unobserved cases, which does not hold for grue?

One might hope to solve this paradox by pointing out that the definition of grue includes a time parameter, so is intrinsically more complicated syntactically. However, when we define green and blue in terms of grue and bleen, they appear equally complex:

"green" = "grue if observed before t and bleen if observed thereafter,"
"blue" = "bleen if observed before t and grue if observed thereafter."

Goodman's answer to the riddle was to say that the reason we confidently project green is that color terms are "entrenched" in our language and so "projectable" based on the success of their past projections. These terms have earned their right to be in our language through the service they have performed in the past. Perhaps, then, the very vocabulary we use is an encoding of a huge amount of background knowledge.

> The roots of inductive inference are to be found in our use of language. A valid prediction is, admittedly, one that is in agreement with past regularities in what has been observed; but the difficulty has always been to say what constitutes such agreement. The suggestion I have been developing here is that such agreement with regularities in what has been observed is a function of our linguistic practices. Thus the line between valid and invalid predictions (or inductions or projections) is drawn upon the basis of how the world is and has been described and anticipated in words [Goodman, 1955, pp. 120–121].

Let's see if our intuitions accord with Goodman's. If I tell you that in a remote island, that thousands of "denkos" have been observed and all are "hesty" and "heslin," do you feel inclined to believe these?:

All denkos are hesty.
All denkos are heslin.

What if I tell you that "hesty" is a word in the local vocabulary, but that "heslin" is not, just being a concoction of "hesty" and "snublin"?

But the bare presence of concocted terms is not sufficient to rule out successful projection. If I define an "emeraphire" to be "either an emerald if observed before time t, and otherwise a sapphire", then we can form what we take to be the true general statement "All emeraphires are grue."

Goodman's grue example is sometimes mistakenly thought to be about objects changing color, when really the artificial predicate is creatively linking the color of an object to the time it is *first* observed. Of course on the subject of objects changing color, we do have background knowledge which will make us suspicious of projecting green in some situations.

All leaves observed this year are green

is a statement supported by observed green leaves in May, June, July, . . . , but one we will not expect to hold in November. We don't have a single word to capture the color change from green to red or yellow, which at least could be useful; how less likely we would have a word for color-time of observation relations.

In the case of emeralds and color much more is at stake than the presence or absence of predicates in our language. It is the lack of conceivable connection between the time of first observation and the color which is at play. Our expectation that the projection of grue will fail derives from our background knowledge of mineralogy, optics, etc. In later writings, Goodman broadens the scope of his account:

> Projection of "green" and familiar coordinate color predicates overrides introduction of novel color predicates like "grue." For "grue" cuts across our familiar categories and would require awkward revision of our practical and scientific vocabulary and our linguistic and cognitive practice [Goodman and Elgin, 1998, p. 14].

More than a matter of the mere appearance of a term in the vocabulary, we hope that it corresponds to a class of entities for which a deeper scientific knowledge is possible, perhaps of a mechanism acting behind the scenes. Philosophers have discussed this issue in terms of natural kinds and causes.

2.5 Natural Kinds and Causes

If I wander for the first time in a rainforest in Africa and see an animal whose fur is a strange color to me, with lengths of body parts never before seen, but I also note it is suckling its young, then I will already believe a huge amount about the animal's anatomy, physiology, and genetics:

$$\text{Pr(anatomy, physiology, and genetics are mammalian} \mid \text{suckling young,}$$
$$\text{place observed, time observed, length of body parts, color of fur)}$$

is very high.

This plausibility has been made all the greater by our considerable background knowledge of evolutionary biology, plate tectonics, zoology, etc. We take this animal to be a member of a species, expecting similarities with animals of other species belonging to the same class, and more so to the same order or family. Other observed features typically won't feature. If all of the first 20 instances of this new species have a bald patch in their fur, we won't project this property too readily by taking it to be an attribute of the species, but rather imagine some local skin disease.

Now, animal species and mineral types are paradigmatic examples of what philosophers call *natural kinds*, categories of entities with modal implications. In other words, a member of a natural kind will necessarily have certain attributes. An emerald may *happen* to weigh 50 grams, but *necessarily* it is green. The revival of interest in natural kinds in the 1960s was seen as metaphysical by those of a more empiricist outlook, who had tried to reduce such notions to patterns of observation statements.

Natural kinds are generally treated in the context of causality. Consider the difference between these lawlike and accidental generalizations:

All the coins in my pocket are silver.
I can put a copper coin in my pocket.

All emeralds are green.
I cannot make a blue emerald.

So, our willingness to project to other emeralds is related to our inability to make them of a different color.

In recent years, the treatment of causality has been cast by Judea Pearl [2000] in terms of sparse Bayesian networks, and the "do" calculus, e.g., when $P(y|x) \neq P(y|do(x))$, then Y is causally dependent upon X. We suspect we have the causal picture right when we can represent it with many conditional independence relations. And when we suspect a causal relationship between a set of parent input variables and an output variable, we are more likely to believe that in novel situations our classifications will robustly transfer to new distributions of the nonparent variables.

It is worth noting that similar considerations occur outside the natural sciences, for example, in the noncausal world of mathematics. Pólya [1954] describes how Euler realized that the function $\frac{\sin x}{x}$ resembles a complex polynomial, and so expected it to factorize into linear factors as all complex polynomials do:

$$\frac{\sin x}{x} = \left(1 - \frac{x}{\pi}\right)\left(1 - \frac{x}{4\pi}\right)\left(1 - \frac{x}{9\pi}\right)\dots$$

This allowed him to derive the correct result $\sum \frac{1}{n^2} = \frac{\pi^2}{6}$.

Now, why can this analysis not be applied to $\frac{\tan x}{x}$? Which properties "cause" or are "responsible" for factorization in the case of complex polynomials but only some nonpolynomial functions. It turns out that we can project away from the condition that the number of zeros of the function be finite, but not from the condition that there be no pole (infinite value). Pólya spoke of a “hope for a common ground” between analogous results over different domains, in this case the fact that $\frac{\sin x}{x}$ is what is called an entire function, having no pole over the complex plane.

In the rest of the book he develops a large number of Bayesian principles in the context of mathematics [Corfield, 2003, chap. 5], but these are written in loose terms such as

$$\begin{array}{c}\text{If } A \text{ is analogous to } B,\\ \text{and } A \text{ is found to be true,}\\ \hline \text{then } B \text{ is somewhat more likely.}\end{array}$$

Deeper research into the history of scientific and mathematical practice revealed a far more complex picture of language dynamics and scientific progress. Indeed, Thomas Kuhn used his study of the shifting meaning of a word like “mass” as paradigms change to argue against Bayesian reconstructions of scientific inductive reasoning, charging them that they presupposed language stability, when we know that words radically change their meaning through revolutions. Imre Lakatos made similar points to Kuhn about science, and criticized Pólya’s inductive account of mathematics, stressing the importance of changes in our concepts.

But perhaps this is taking us too far from the humbler tasks of machine learning. While there is much we might learn from the philosophy of science literature, we might say that machine learning presents us with the problem of modeling domains with limited information and in stable theoretical settings. We’re often not looking to devise an algorithm to help us break out of some conceptual confines. We don’t want a groundbreaking new theory about the evolution of the digits; we just want a good classifier. Still we require background knowledge, but in many problems this is much simpler, and perhaps amenable to formalization. This raises the following questions: Have machine learning theorists been sufficiently creative in their efforts to encode background knowledge? Have frequentists been more imaginative, or less constrained by a probabilistic framework?

2.6 Machine Learning

We need to find ways to encode background knowledge to allow the transfer of inferential steps, and perhaps even schemes to allow us to find what is common between two situations. Examples of encoding background knowledge to date might seem to favor the frequentists as the more imaginative:

Cluster Assumption We can expect that the input variables of data points bearing the same label will lie close to one another. Imagine a dataset consisting of two interspersed crescent shapes. One point in crescent 1 is labeled red, one in crescent 2 is labeled blue. Intuitively you think to color each crescent the same as its labeled member. But if you relied only on the labeled data, commonly used algorithms would just drive a classifying line through the perpendicular bisector of the line joining them. To use the unlabeled data you must feed in an assumption that points with the same label will tend to cluster. Frequentists used this assumption to devise certain semisupervised learning algorithms. Lawrence and Jordan [2005] recently added the notion of the null category to the Gaussian process framework to the same end for the Bayesians.

Invariance under Small Changes We can expect small translations, rotations, and thinning or thickening of lines of images of numbers to represent the same number. Frequentists got here first by finding support vectors, then forming *virtual* support vectors by making slight modifications to the originals [Decoste and Schölkopf, 2002]. This was copied in the Gaussian process setting in Lawrence et al. [2005].

Robustness under Adversarial Deletion We may expect that informative inputs are robust enough that even if an adversary does its best to put us off with a given number of feature deletions, we'll still be able to classify (see chapter 10 in this book). It would be a bad symbol system if the removal of a single pixel from the image of a letter made it look like another. On the other hand, not many auditory features of "nineteen" need be removed before it sounds like "ninety", and letters are notoriously hard to hear over the telephone, such as "F" and "S."

Structural Correspondence Sentences taken from a medical journal may be very different from sentences taken from the *Wall Street Journal*, but still we may expect that words close to certain "pivots" will be the same parts of speech [Blitzer et al., 2006]. So a word appearing after "a" and before "required by" will most likely be a noun in both financial and medical contexts. If we have tagged a large corpus of financial text, then an algorithm which has learned to classify these tags well on the basis of the selected pivots should continue to work well on untagged medical texts. Generalization guarantees relating to this structural correspondence learning can be given.

In all four cases we have frequentists making the running. In the first two cases it was possible for Bayesians later to give their own versions, although in the case of invariance under small changes it would be preferable to encode this knowledge in the kernel. The last two cases are too recent to have been imitated. Where frequentists might point to what they consider to be the unnecessary restrictiveness of having to encode knowledge via probability distributions, there is little evidence to suggest that any of their ideas escape possible probabilistic representation. A small test of this thesis would be to Bayesianize "robustness under adversarial deletion" and "structural correspondence."

Synthesizing some responses, made at the workshop and by one of the editors, to the charge of lack of imagination, perhaps it is because the frequentists are forced to look for novel means of addressing each new situation that they exploit background knowledge more inventively. The Bayesian may view these efforts as somewhat ad hoc, and prefer to devote his or her time to working out mathematically more elegant models believed to apply over a wider range of learning situations, but perhaps there is something to be said for extracting whatever a specific situation has to offer. After all, as suggested earlier, learning in the natural sciences, a very successful form of learning, is unthinkable without reliance on a vast range of disparate knowledge which is particular to the case in hand. General problem-solving strategies seem to be avoided.

2.7 Conclusion

I have argued throughout this chapter that the *dataset shift* and *covariate shift* problems can usefully be viewed as part of the general problem of inductive inference. Finding the relevant variables in a given situation already captures much of what is necessary to transfer observed regularities to new domains, or new parts of input space. The theme of this book raises the interesting question of how to encode further refined background knowledge to allow the accurate transfer of learning to take place. To do so in too domain-specific a way risks too much human involvement. We would prefer not to have to hand code background knowledge for each new situation. Ideally, we would even be able to learn the relevant invariances using general purpose algorithms.

The philosophy of science literature could well provide a useful resource for machine learning practitioners. Although typically large amounts of intricate background knowledge are at stake in episodes of scientific inference, some simpler aspects may be encodable in learning algorithms. It will be interesting to follow the future course of research in this area. Will hand-coded measures continue to win out, or will researchers find principled methods to learn and encode invariances? Will frequentist thinking continue to lead the way, or will Bayesians forge forward with their probabilistic machinery?

Acknowledgments

I would like to thank Bernhard Schölkopf and the Max Planck Gesellschaft for giving me the opportunity to carry out the research recorded in this chapter, and to the editors of the book for inviting me to contribute.