2
Statistical Learning: What It Means to Learn

Students often get confused about the meaning and significance of a relationship learnt via fitting a model to data. Some of them think such relationships are analogous to, say, a law of nature like $F = ma$, which applies every time force is applied to a mass anywhere in the universe. The main goal of this chapter is to explain the statistical nature of machine learning — models are fitted on a *particular* distribution of data points, and its predictions are valid only for data points from the same distribution. (See Chapter 18.)

2.1 *A Warm-up Example*

We work through a concrete example [1] before enunciating the general properties of statistical learning. Suppose we are studying the relationship between the following quantities for the population of Princeton: *height (H), number of exercise hours per week (E), amount of calories consumed per week (C), and weight (W)*. After collecting information from 200 randomly sampled residents, and using a $80:20$ train/test split, we perform a linear regression on the training dataset to come up with the following relationship:

$$W = 50 + H + 0.1C - 4E \tag{2.1}$$

Let's also say that the average squared residual on train and test data were both 100. This means that the relationship (2.1) holds with an error of 10 lbs on a typical test data point. [2]

Question 2.1.1. *Alice was one of the Princeton residents in the study, but the prediction of the model is very off of her actual value (squared residual is 300). Does this prove the model wrong?*

The answer is no. The least squares linear regression finds the model that minimizes the *average* squared residual across all training data points. The residual could be large for a particular individual.

[1] This example is purely hypothetical, and all numbers in this section are made up.

[2] Also, the trained model exhibits *perfect* generalization: test loss is the same as training loss!

Question 2.1.2. *There was a follow-up research for every Princeton resident who is taller than 7 feet. All of them reported squared residual of* 500*. Does this prove the model wrong?*

The answer is still no. People who are taller than 7 feet make up a tiny fraction of the entire population. Their residuals have very small effect on the *average* squared residual. The residual could be large for a small subset of the population.

Question 2.1.3. *There was a follow-up survey that tested the model on every single Princeton resident. Is it possible that the average squared residue is* 200 *for the entire population?*

The answer is yes, although it is unlikely. Consider the distribution of 4-tuples (H, E, C, W) over the entire Princeton population. This is some distribution over a finite set of 4-dimensional vectors. [3] The 200 residents we surveyed were randomly drawn from this distribution. Out of these 200 data points, 40 were randomly chosen to be held-out as test data, while the remaining 160 were used as training data. We can also say that these 40 data points were chosen at random from the distribution over the entire population of Princeton. Thus when we test the model in (2.1) on held-out data, we're testing this relationship over a random sample of 40 data points drawn from the population. 40 is a large enough number to give us some confidence that the average squared residual of the test data is a good estimate of the squared residual in the population, but just as polling errors happen during elections, there is some chance that this estimate is off. In this case, we would say that the 40 test samples were *unrepresentative* of the full population.

[3] $31,000$ vectors to be more exact. The population of Princeton is $31,000$.

It is important to remember that the training and test data are sampled from the same distribution as the population. Therefore, the average squared residual of the training and test data are only a good estimate of the squared residual of the distribution they were sampled from. This also means that the relationship found from the training data only holds (with small residue) for that *particular* distribution. If the population is different, or if the distribution shifts within the same population, the relationship is not guaranteed to hold. For example, the relationship in (2.1) is not expected to hold for people from Timbuktu, Mali (a different population), or for residents of Princeton who are taller than 7 feet (a tiny subpopulation that is likely unrepresentative of the population). Now consider the following situation:

Question 2.1.4. *It becomes fashionable in Princeton to try to gain weight. Based on the relationship in* (2.1)*, everyone decides to increase their value of C and reduce their value of E. Does the model predict that many of them will gain weight?*

The answer is no. The model was fitted to and tested on the distribution obtained before everyone tried to gain weight. It has not been fitted on the distribution of data points from people who changed their values of C and E. In particular, note that if everyone reduces their E and increases their C, then the distribution has definitely changed — the average value of the E coordinate in this distribution has decreased, whereas the average value of the C coordinate has increased.

In general, a relationship learned from a fitted model illustrates *correlation* and need not imply *causation*. The values of H, C, E in (2.1) do not *cause* W to take a specific value. The equation only shows that the values are connected via this linear relationship on average (with some bounded square residuals).

2.2 *Summary of Statistical Learning*

The above discussion leads us to summarize properties of Statistical Learning. Note that these apply to most methods of machine learning, not just linear regression.

Training/test data points are sampled from some distribution $\mathcal{D}$: In the above example, 200 residents were randomly sampled from the entire population of Princeton residents.

The learnt relationship holds only for the distribution $\mathcal{D}$ that the data was sampled from. The performance of the model on test data is an estimate of the performance of the model on the full distribution $\mathcal{D}$.

There is a small probability that the estimate using test data is off. This is analogous to polling errors in opinion polls. The computation of "confidence bounds" is discussed in Chapter 18.

2.3 *Implications for Applications of Machine Learning*

The above framework and its limitations have real-life implications.

1. *Results of medical studies may not apply to minority populations.* This can happen if the minority population is genetically distinct and constitutes only a small fraction of the population. Then test error could be large on the minority population even if it is small on average. In fact, there have been classic studies about heart disease in the 1960s whose conclusions and recommendations fail to apply well to even a group that is half of the population: females! In those days heart disease was thought to largely strike males (which subsequently turned out to be quite false) and so

the studies were done primarily on males. It turns out that heart diseases in female patients behave differently. Many practices that came out of those studies turned out to be harmful to female patients. [4]

2. *Classifiers released by tech companies in the recent past were found to have high error rates on certain minority populations.* It was quickly recognized that relying on test error alone can lead to adverse outcomes on subpopulations. [5]

3. *Creating interactive agents is difficult.* In an interactive setting (*e.g.*, an online game), a decision-making program is often called an *agent*. When an agent has to enter an extended number of interactions [6] with a human (or another agent designed by a different group of researchers, as happens in Robocup soccer [7]), then statistical learning requires that the agent to have been exposed to similar situations/interactions during training (*i.e.*, from a fixed distribution). It is quite unclear if this is true.

[4] See https://www.theatlantic.com/health/archive/2015/10/heart-disease-women/412495/.

[5] See https://time.com/5520558/artificial-intelligence-racial-gender-bias/.

[6] Later in the book we encounter Reinforcement Learning, which deals with such settings.

[7] See https://www.robocup.org/.