CMSE 890-002: Mathematics of Deep Learning, MSU, Spring 2020

Lecture 02: Prediction versus estimation

January 13, 2019

Lecturer: Matthew Hirn

1 Prediction versus estimation

Inspired by [5, Section I.A].

1.1 Introduction

Prediction versus estimation; correlation versus causation. When you hear these phrases in the context of machine learning, what do you think of? Maybe one thinks of the difference between classifying new data points and generating new data points. Or perhaps one considers that correlation is a symmetric assessment (e.g., if A is correlated with B, then B is correlated with A), but causation is directional (e.g., if A causes B, B does not necessarily cause A)[1].

These concepts are in some sense the difference between machine learning and statistics. In machine learning and prediction based tasks, we are often interested in developing algorithms that are capable of learning patterns from given data in an automated fashion, and then using these learned patterns to make predictions or assessments of newly given data. In many cases, our primary concern is the quality of the predictions or assessments, and we are less concerned about the underlying patterns that were learned in order make these predictions. Neural networks are, in some sense, the epitome of this point of view. In various contexts they are incredibly good at making predictions, but they are often referred to as "black box" methods due to the difficulty in understanding the model by which they make such predictions. For example, the most powerful convolutional neural networks are incredibly good at classifying natural images (sometimes even better than humans), but it is very difficult to understand the mechanisms by which they make such predictions.

In (classical) statistics and estimation, one is more concerned with the underlying model that makes the prediction. In other words, are the parameters of the model that makes the prediction statistically significant? Or could several other models (i.e., different parameter choices) have made the same prediction? This is the correlation versus causation issue. It comes up, for example and perhaps most notably, in medical trials and studies, in which one must not only find correlations and patterns in the data, but one must find the causal factors of a disease, so that one may develop and prescribe treatment.

[1] Thanks to Cullen Haselby and Bashir Sadeghi for these comments in class.

1.2 Making things more precise with probability

Let us try to make this difference a bit more precise. To do so we will use the language of probability. Consider a given data set

$$T = \{(x_1, y_1), \ldots, (x_N, y_N)\} ,$$

consisting of data points $\{x_1, \ldots, x_N\} \subset \mathcal{X}$ and associated scalar valued labels $\{y_1, \ldots, y_N\} \subset \mathcal{Y}$. Let us assume that each data point x_i was sampled from $\mathcal{X}$ according to a probability distribution $\mathbb{P}_X$. This means, more precisely, we have a *probability space* $(\mathcal{X}, \Sigma, \mathbb{P}_X)$, which consists of:

- $\mathcal{X}$: The set of all possible outcomes, i.e., data points.
- Σ: The space of all possible events, i.e., collections of data. This is a set of sets, which has additional structure (see below).
- $\mathbb{P}_X$: The probability measure. For each event (set / collection of data points) $A \in \Sigma$, $\mathbb{P}_X(A)$ is the probability of the event A occuring.

The set Σ is a σ-algebra, meaning it has the following properties:

1. $\mathcal{X} \in \Sigma$.
2. If $A \in \Sigma$, then the complement of A, $A^c = \mathcal{X} \setminus A$, is also in Σ. Note this means $\emptyset \in \Sigma$.
3. If $A_1, A_2, \ldots$ are all in Σ, then their union is also in Σ, i.e.,

$$\bigcup_{i=1}^{\infty} A_i \in \Sigma .$$

Note that these properties also imply that if $A_1, A_2, \ldots$ are in Σ, then

$$\bigcap_{i=1}^{\infty} A_i \in \Sigma .$$

The probability measure $\mathbb{P}_X$ satisfies the following properties:

1. $\mathbb{P}_X(\mathcal{X}) = 1$.
2. Whenever $A_1, A_2, \ldots$ is a sequence of disjoint sets in Σ, then

$$\mathbb{P}_X\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}_X(A_i) .$$

From these properties we can also conclude that $\mathbb{P}_X(\emptyset) = 0$ and $\mathbb{P}_X(A^c) = 1 - \mathbb{P}_X(A)$.

Example 1.1. A simple example, that is not too relevant to our future discussions but which illustrates the idea, is the following. Consider flipping a coin twice, with the probability of heads being p and the probability of tails being q. There are four possible outcomes:

$$\mathcal{X} = \{HH\,,\,HT\,,\,TH\,,\,TT\}\,.$$

We also know the probabilities of each of these outcomes are p^2, pq, pq, and q^2, respectively. We thus set

$$\mathbb{P}_X(HH) = p^2\,,\ \mathbb{P}_X(HT) = \mathbb{P}_X(TH) = pq\,,\ \mathbb{P}_X(TT) = q^2\,. \tag{1}$$

This information is enough to define a probability space $(\mathcal{X}, \Sigma, \mathbb{P}_X)$, but it does not specify all the sets in Σ or their probabilities. Indeed, the event $A = \{HH\,,\,HT\}$ = "the first coin toss is a head," should be in Σ since it is the union of HH and HT. We assign A the probability $\mathbb{P}_X(A) = p^2 + pq = p$. In fact, the easiest way to ensure Σ is a σ-algebra is to include every subset of $\mathcal{X}$, including $\emptyset$ and $\mathcal{X}$ itself, and to assign probabilities using the rules of (1)[2].

Example 1.2. Another example, more relevant to our studies in this course, is inspired by the MNIST database of handwritten digits; see Figure 6. In this case $\mathcal{X}$ is the infinite set of all possible handwritten digits, and the σ-algebra Σ and the probability measure $\mathbb{P}_X$ are unknown to us, but we assume the training and testing set in the database are sampled according to $\mathbb{P}_X$, whatever it may be[3].

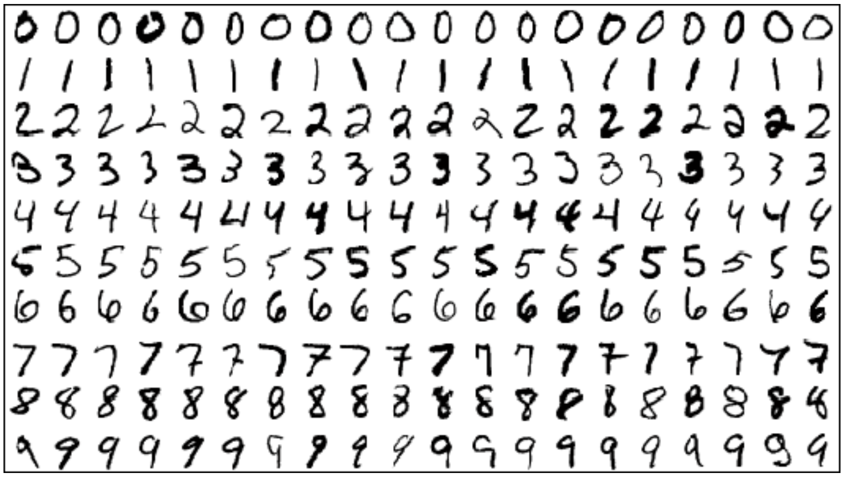

Figure 6: *Examples from the MNIST database of handwritten digits.*

Since the data points x_i are randomly sampled, and the label y_i depends on x_i, the labels y_i are sampled from $\mathcal{Y}$ according to a probability distribution that encodes the dependence of y_i on x_i. We model this as a conditional probability distribution $\mathbb{P}_{Y|X}(B \mid X = x)$, which

[2] As pointed out by ???, in this case Σ is the power set of the set of outcomes.

[3] As pointed out by Cullen Haselby, if we assume that each image is of a fixed resolution and has a finite grayscale gradient, then the number of possible images is very large, but finite.

measures the probability of an event $B \subset \mathcal{Y}$ (i.e., a set of labels) given an outcome $x \in \mathcal{X}$. Together these two distributions induce a joint distribution $\mathbb{P}_{X,Y}$ on $\mathcal{X} \times \mathcal{Y}$, from which we draw the training samples.

In particular, suppose we draw the x_i's independently from $\mathbb{P}_X$. This means we "run the experiment" of drawing a point N times, each time independent from the others, and we obtain a random data point x_i. An analogy is flipping the coin, i.e. Example 1.1. Suppose we carry out the experiment of flipping the coin twice N times, each time independent from the other times. Then each time we will get a "data point," which corresponds to one of the four outcomes HH, HT, TH, TT, with the probabilities calculated earlier. Drawing a training sample (that is, a data point and a label) first entails drawing a point $x \in \mathcal{X}$ according to the distribution $\mathbb{P}_X$, and then drawing a point $y \in \mathcal{Y}$ according to $\mathbb{P}_{Y|X}(\cdot \mid X = x)$. When we want to think of the data point as being determined (say by an experiment, or a draw of a training point) we will write (x, y). On the other hand, when we want to think of the training point as a pair of random variables, one X taking values in $\mathcal{X}$ and the other Y taking values in $\mathcal{Y}$, we will write (X, Y).

Example 1.3. An example to keep in mind, that we will come back to later, is the following. Suppose $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$, and that a label $y_i \in \mathcal{Y}$ is generated from an underlying deterministic function $F : \mathbb{R}^d \to \mathbb{R}$ plus random noise:

$$y_i = F(x_i) + \varepsilon_i\,, \quad 1 \leq i \leq N\,.$$

Often we will assume that the ε_i are independently and identically distributed (i.i.d.) according to the normal distribution with mean zero and variance σ^2, i.e. $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. In this case, if X is the random variable that takes values in $\mathbb{R}^d$ according to the probability distribution $\mathbb{P}_X$, and Y is the random variable that takes values in $\mathbb{R}$ according to $\mathbb{P}_{Y|X}(\cdot \mid X = x)$, then we have that

$$Y \sim \mathcal{N}(F(x), \sigma^2)\,,$$

where $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2. In other words, given that $X = x$, the label Y is a normal random variable with mean $F(x)$ and variance σ^2.

References

[1] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, 2015.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 2009.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012.

[4] Larry Greenemeier. AI versus AI: Self-taught AlphaGo Zero vanquishes its predecessor. *Scientific American*, October 18, 2017.

[5] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G.R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to Machine Learning for physicists. arXiv:1803.08823, 2018.