EN.601.774 Theory of Replicable ML Spring 2025

Lecture 1

Instructor: *Jess Sorrell* Scribe: *Jess Sorrell*

1 Introduction

In this course, we'll discuss the challenges of replicability in machine learning and data analysis, and how we can mitigate them.

Definition 1.1 ((informal) Reproducibility)**.** An experimental result has been "reproduced" when another team of researchers obtains the same result, *using the original data*.

For example, say you and your coauthors publish a new method for image classification that you evaluated on MNIST. Another team of researchers should be able to use your code to obtain the same results on MNIST that you obtained. This seems like it should be straightforward, but failure of open-sourced code to reproduce published results is unfortunately common in the field of ML! Reproducibility already poses some technical challenges, but in this class we'll be focusing on the stronger, but related notion of replicability.

Definition 1.2 ((informal) Replicability)**.** An experimental result has been "replicated" when another team of researchers obtains the same result, *using new data* (typically from the same distribution of interest).

Replicability reflects the generalizability of published findings from the data used for experiment to new data. For instance, if we train a medical risk prediction tool to low error on patient data from a network of hospitals, we want that tool to also have low error on new patients from that same network. In fact we typically want something stronger – that the tool has low error on patients from other hospitals outside of the network from which our training data was drawn. Replication efforts in the sciences (ML included) can increase our confidence that published findings generalize beyond the experimental context; that they reflect enduring phenomena in the real world and aren't simply the result of statistical flukes in our data.

To formalize notions of replicability, we need to specify what it means to "obtain the same results" on "new data". Since this is a course on ML theory, let's start with a toy version of a familiar example: comparing the loss of a binary classifier h on a dataset S_1, drawn i.i.d. from distribution D to its loss on a new dataset S_2 drawn i.i.d. from the same the distribution.

Setup. Fix a feature domain $\mathcal{X}$ and binary label space $\mathcal{Y} = \{0, 1\}$. Let D denote a distribution over $\mathcal{X} \times \mathcal{Y}$. We'll define our loss function to be the 0-1 loss,

$$\ell(h(x), y) = \begin{cases} 0, & h(x) = y \\ 1, & h(x) \neq y \end{cases}.$$

We are given a classifier $h : \mathcal{X} \to \mathcal{Y}$ (h not trained using S_1, so S_1 and h are independent here). We go out and collect a dataset $S_1 \sim_{i.i.d} D^m$ (we draw m i.i.d. samples (x, y) from D). We determine the empirical loss of h on S_1

$$\ell_{S_1}(h) := \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i), y),$$

and publish this result. Another team of researchers performs the same experiment, using a new sample $S_2 \sim D^m$. If our result is replicable, the other team of researchers should be able to get "the same results," but what does this mean?

If our scientific hypothesis was that model h performs poorly on distribution D, then a reasonable definition would be that the second team of researchers should obtain empirical loss similar to ours. That is, for $S_1, S_2 \sim_{i.i.d.} D^m$

$$|\ell_{S_1}(h) - \ell_{S_2}(h)| < \varepsilon,$$

for some $\varepsilon \in (0, 1)$.

So to bound the probability that a replication effort fails, we want a statement of the form

$$\Pr_{S_1,S_2}[|\ell_{S_1}(h) - \ell_{S_2}(h)| \geq \varepsilon] \leq \delta$$

for some $\delta \in (0, 1)$. Using the triangle inequality, we see that it suffices to bound the probability that $\ell_{S_1}(h)$ deviates from its expectation by more than $\varepsilon/2$.

$$\begin{aligned}
\Pr_{S_1,S_2}[|\ell_{S_1}(h) - \ell_{S_2}(h)| \geq \varepsilon] &= \Pr_{S_1,S_2}[|\ell_{S_1}(h) - \ell_D(h) + \ell_D(h) - \ell_{S_2}(h)| \geq \varepsilon] \\
&\leq \Pr_{S_1,S_2}[|\ell_{S_1}(h) - \ell_D(h)| \geq \varepsilon/2] + \Pr_{S_1,S_2}[|\ell_D(h) - \ell_{S_2}(h)| \geq \varepsilon/2] \\
&= 2 \Pr_{S_1}[|\ell_{S_1}(h) - \ell_D(h)| \geq \varepsilon/2]
\end{aligned}$$

We'll now build up some very helpful techniques from probability theory to bound the probability that $\ell_{S_1}(h)$ deviates from its expectation by more than $\varepsilon/2$, as a function of m, the size of the sample S_1.

Theorem 1.3 (Markov's Inequality)**.** *Let X be a non-negative random variable. Then for any $a > 0$,*

$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}.$$

Proof. Let p denote the PDF of X. Since X is non-negative, we have

$$\begin{aligned}
\mathbb{E}[X] &= \int_{v=0}^{\infty} v \cdot p(v)dv && \text{by definition} \\
&= \int_{v=0}^{a} v \cdot p(v)dv + \int_{a}^{\infty} v \cdot p(v)dv \\
&\geq \int_{v=a}^{\infty} v \cdot p(v)dv && \text{by non-negativity of } X \\
&\geq \int_{v=a}^{\infty} a \cdot p(v)dv && v \geq a \text{ for } v \in [a, \infty] \\
&= a \int_{v=a}^{\infty} p(v)dv && a \text{ is a constant} \\
&= a \cdot \Pr[X \geq a]
\end{aligned}$$

□

We'll apply Markov's inequality to the r.v. $X = (\ell_{S_1}(h) - \ell_D(h))^2$ with $a = \varepsilon^2$.

$$\begin{aligned}
\Pr_{S_1}[|\ell_{S_1}(h) - \ell_D(h)| \geq \varepsilon] &= \Pr_{S_1}[(\ell_{S_1}(h) - \ell_D(h))^2 \geq \varepsilon^2] \\
&\leq \frac{\mathbb{E}[(\ell_{S_1}(h) - \ell_D(h))^2]}{\varepsilon^2} \\
&= \frac{\mathsf{Var}(\ell_{S_1}(h))}{\varepsilon^2} && \text{by def. of } \mathsf{Var} \\
&= \frac{\mathsf{Var}(\frac{1}{m}\sum_{i=1}^{m} \ell(h(x_i), y_i))}{\varepsilon^2} && \text{unpack } \ell_{S_1}(h) \\
&= \frac{\mathsf{Var}(\sum_{i=1}^{m} \ell(h(x_i), y_i))}{m^2\varepsilon^2} && \mathsf{Var}(cX) = c^2\mathsf{Var}(X) \\
&= \frac{\sum_{i=1}^{m} \mathsf{Var}(\ell(h(x_i), y_i))}{m^2\varepsilon^2} && \mathsf{Var}(X_1 + X_2) = \mathsf{Var}(X_1) + \mathsf{Var}(X_2) \\
&= \frac{m\mathsf{Var}(\ell(h(x), y))}{m^2\varepsilon^2} && \mathsf{Var}(\ell(h(x_i), y_i)) = \mathsf{Var}(\ell(h(x_j), y_j)) \\
&= \frac{\mathsf{Var}(\ell(h(x), y))}{m\varepsilon^2} \\
&\leq \frac{1}{4m\varepsilon^2} && \mathsf{Var}(B(p)) = p(1-p) \leq \frac{1}{4}
\end{aligned}$$

So if we want $\Pr_{S_1}[|\ell_{S_1}(h) - \ell_D(h)| \geq \varepsilon] < \delta$, we can take $m > \frac{1}{4\varepsilon^2\delta}$. As an aside, we have just proven Chebyshev's inequality along the way.

Theorem 1.4 (Chebyshev's Inequality)**.** *Let X be a random variable with non-zero variance $\sigma^2 = \mathsf{Var}(X)$. Then for any $\lambda > 0$*

$$\Pr[|X - \mathbb{E}[X]| \geq \lambda\sigma] \leq \frac{1}{\lambda^2}.$$

Great! So now we have some guarantee that, so long as we take our sample large enough (and so does the other team of researchers), replication efforts will be successful with good probability! Both research teams will end up with an empirical loss $\ell_S(h)$ that is close to its expectation $\ell_D(h)$, and therefore close to the other team's, except with probability 2δ. But we can do much, much better!

Theorem 1.5 (Hoeffding's Inequality)**.** *Let $X_1, X_2, \ldots, X_m$ be independent, bounded random variables with $X_i \in [a_i, b_i]$. Let $S_m = \sum_{i=1}^m X_i$. Then*

$$\Pr_{X_1, X_2, \ldots, X_m}[S_m \geq \mathbb{E}[S_m] + t] \leq e^{-\frac{2t^2}{\sum_{i=1}^m (b_i - a_i)^2}}.$$

Note this also implies

$$\Pr_{S \sim D^m}[|\ell_S(h) - \ell_D(h)| \geq t/m] \leq 2e^{-\frac{2t^2}{\sum_{i=1}^m (b_i - a_i)^2}}$$

and so

$$\Pr_{S \sim D^m}[|\ell_S(h) - \ell_D(h)| \geq t] \leq 2e^{-2t^2 m}$$

We'll prove this theorem is 2 parts. We'll assume the following lemma (to be proved later).

Lemma 1.6 (Hoeffding's Lemma)**.** *Let X be a random variable such that $X \in [a, b]$. Then for any $\lambda > 0$,*

$$\mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq e^{\frac{\lambda^2 (b-a)^2}{8}}$$

Proof. (Hoeffding's Inequality) From Markov's inequality, we know that for all $\lambda, t > 0$,

$$\begin{aligned}
\Pr[S_m - \mathbb{E}[S_m] \geq t] &= \Pr[e^{\lambda(S_m - \mathbb{E}[S_m])} \geq e^{\lambda t}] \\
&\leq \frac{\mathbb{E}[e^{\lambda(S_m - \mathbb{E}[S_m])}]}{e^{\lambda t}} && \text{Markov's inequality} \\
&= \frac{\mathbb{E}[e^{\lambda(\sum_{i=1}^m X_i - \mathbb{E}[X_i])}]}{e^{\lambda t}} && \text{def of } S_m \text{ and linearity of } \mathbb{E} \\
&= \frac{\mathbb{E}[\prod_{i=1}^m e^{\lambda(X_i - \mathbb{E}[X_i])}]}{e^{\lambda t}} \\
&= \frac{\prod_{i=1}^m \mathbb{E}[e^{\lambda(X_i - \mathbb{E}[X_i])}]}{e^{\lambda t}} && \text{Independence of } X_i s \\
&\leq \frac{\prod_{i=1}^m e^{\frac{\lambda^2 (b_i - a_i)^2}{8}}}{e^{\lambda t}} && \text{Hoeffding's lemma}
\end{aligned}$$

We showed this is true for all $\lambda > 0$, so in particular is must be true for $\lambda = \frac{4t}{\sum_{i=1}^m (b_i - a_i)^2}$.

Then we have

$$\begin{aligned}
\Pr[S_m - \mathbb{E}[S_m] \geq t] &\leq \frac{\prod_{i=1}^{m} e^{\frac{\lambda^2 (b_i - a_i)^2}{8}}}{e^{\lambda t}} \\
&= \frac{e^{\frac{\lambda^2}{8} \sum_{i=1}^{m} (b_i - a_i)^2}}{e^{\lambda t}} \\
&= e^{\frac{\lambda t}{2} - \lambda t} \\
&= e^{-\frac{\lambda t}{2}} \\
&= e^{-\frac{2t^2}{\sum_{i=1}^{m} (b_i - a_i)^2}}
\end{aligned}$$

□

Now it remains to prove the lemma.

Lemma 1.7 (Hoeffding's Lemma)**.** *Let X be a random variable such that $X \in [a, b]$. Then for any $\lambda > 0$,*

$$\mathbb{E}[e^{(\lambda X - \mathbb{E}[X])}] \leq e^{\frac{\lambda^2 (b-a)^2}{8}}$$