Lecture 2 Notes

Introduction to Data Science IS360

1 Feasibility of Learning

1.1 The Feasibility of Learning

The feasibility of learning is concerned with if we can be sure that a model can generalize from a finite training set to unseen examples.

The feasibility of learning comes down to the dataset, since by definition we only know the value of the target function f within the finite set of examples $\mathcal{D}$, making the target function f an unknown. We might select a dataset containing examples that produce a hypothesis very far from the actual distribution.

The question of feasibility is how can we say with confidence (as opposed to certainty) that the learned hypothesis h represents the target function f? This turns the question of feasibility into a matter of probability.

1.2 The Hoeffding Inequality

Consider a hypothesis function h drawn from a hypothesis set $\mathcal{H}$, which maps inputs x to outputs y. We draw a sample of N independent data points from a fixed, unknown distribution (our target function). We define the true error (out of sample error) $E_{\text{out}}(h)$ as the probability that h misclassifies a randomly drawn point from the underlying distribution, and the empirical error (in-sample error) $E_{\text{in}}(h)$ as the fraction of misclassified points in the finite sample.

The Hoeffding Inequality states that for any hypothesis h

$$\Pr\left(|E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right) \leq 2\exp(-2\epsilon^2 N) \tag{1}$$

Where ϵ is a non-zero threshold.

The inequality states that the probability that the difference between $E_{\text{in}}(h)$ and $E_{\text{out}}(h)$ is greater than some threshold ϵ, decreases exponentially with the sample size N.

Note 1. With enough data points N, the empirical error $E_{\text{in}}(h)$ will be close to the true error $E_{\text{out}}(h)$ with high probability.

See lecture 2 slides for marble example and book pages 15-27 for derivation of Hoeffding inequality

1.3 Union Bound

Union Bound
The union bound states that for any collection of events $A_1, A_2, \cdots, A_M$, the probability of at least one of these events occurring is at most the sum of the probabilities of each event

$$\Pr\left(\bigcup_{i=1}^{M} A_i\right) \leq \sum_{i=1}^{M} \Pr(A_i) \tag{2}$$

Dice Roll Example
We roll a dice 3 times and we define the following events, each with a probability $= \frac{1}{6}$:

- A_1: 6 on first roll
- A_2: 6 on second roll
- A_3: 6 on third roll

To find the probability that at least one of these events occurs (that we get a 6 on one of the 3 rolls) we use union bound

$$\Pr(A_1 \cup A_2 \cup A_3) \leq \Pr(A_1) + \Pr(A_2) + \Pr(A_3)$$

$$\Pr(A_1 \cup A_2 \cup A_3) \leq \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

Note 2. The union bound is not an equation, it only provides the upper bound on what the actual probability may be. In fact for independent events like a dice roll the union bound actually overestimates the probability, the actual probability to roll at least one 6 in 3 dice rolls is $\frac{91}{216} \approx 0.421$.

1.4 The Hoeffding Inequality with Multiple Hypotheses

The inequality in (1) assumes we only look at a single hypothesis. However, we are almost always considering several hypotheses and choosing the hypothesis with the lowest in-sample error $E_{\text{in}}(h)$. To extend the Heoffding inequality to several hypotheses we apply the union bound.

Each event A_i is the event that the empirical error $E_{\text{in}}(h)$ of hypothesis h_i deviates from the true error $E_{\text{out}}(h)$ by more than ϵ

$$A_i = \{|E_{\text{in}}(h_i) - E_{\text{out}}(h_i)| > \epsilon\}$$

Applying the union bound

$$\Pr\left(\bigcup_{i=1}^{M} |E_{\text{in}}(h_i) - E_{\text{out}}(h_i)| > \epsilon\right) \leq \sum_{i=1}^{M} \Pr\left(|E_{\text{in}}(h_i) - E_{\text{out}}(h_i)| > \epsilon\right)$$

Substituting the right-hand-side for the upper bound from (1)

$$\Pr\left(\bigcup_{i=1}^{M} |E_{\text{in}}(h_i) - E_{\text{out}}(h_i)| > \epsilon\right) \leq \sum_{i=1}^{M} 2\exp(-2\epsilon^2 N)$$

Simplifying this gives us the final inequality

$$\Pr\left(\exists h \in \mathcal{H} : |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right) \leq 2M \exp\left(-2\epsilon^2 N\right) \tag{3}$$

Where M is the number of hypotheses h in $\mathcal{H}$ ($|\mathcal{H}| = M$).

The inequality states that the probability that there exists a hypothesis h in $\mathcal{H}$ for which the empirical error and true error differ by more than ϵ is at most $2M \exp\left(-2\epsilon^2 N\right)$.

This means that:

1. As the number of samples N increases, this probability $\Pr\left(\exists h \in \mathcal{H} : |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right)$ decreases exponentially.
2. When considering multiple hypotheses, due to the $2M$ term, the probability increases as N increases.

Overfitting Example
You can think of the probability on the left-hand-side of the inequality to be the probability that we are overfitting on the training data.

Note 3. The Hoeffding inequality places a bound on how well we are generalizing. We can use the inequality to say "in the worst case, the probability our model is overfitting is $2M \exp\left(-2\epsilon^2 N\right)$"

If we use a simple linear model with few parameters, our set of possible hypotheses $\mathcal{H}$ is (relatively) small. If we use a small training dataset with this model the probability that we are overfitting is high, as the size of the dataset N increases we lower this probability exponentially.

If we use a complex non-linear model with many parameters, the set of possible hypotheses is much larger, so the probability that we are overfitting also increases (by a factor of 2).

Note 4. Following this example, we can see how the size of our dataset N matters a lot more than the complexity of our model M ($\exp N$ vs. $2M$) when it comes to overfitting.

Note 5. Another thing to keep in mind is that just because were not overfitting this doesn't mean we are finding a hypothesis that captures the target function. For example a linear model will never be able to capture a non linear target function, but it can still avoid overfitting. Bias and variance, covered in another lecture, go into more detail on this point.