Lecture 3: Error Analysis

CS 182 Spring 2021 – Taught by Sergey Levine

Notes by Michael Zhu

Empirical Risk vs True Risk

Risk: probability that your output is wrong

This is quantified by **expected value of loss** under the distribution that your data comes from

$$\text{True risk} = E_{x\sim p(x),\ y\sim p(y|x)}[L(x, y, \theta)]$$

NOT THE SAME AS TRAINING LOSS. During training, we can't sample $x \sim p(x)$. We just have dataset D and can't generate new samples during training.

$$\text{Empirical risk (from training)} = \frac{1}{n}\sum_{i=1}^{n} L(x_i, y_i, \theta)$$

Empirical risk minimization

Supervised learning is usually *empirical* risk minimization.

Question: Is this the same as *true* risk minimization?

$$\frac{1}{n}\sum_{i=1}^{n} L(x_i, y_i, \theta) \approx E_{x\sim p(x),\ y\sim p(y|x)}[L(x, y, \theta)]$$

Not always. Since we are selecting θ based on the empirical risk, the θ we get from training will be **biased** to the empirical risk. This creates a potential issue where the empirical risk is no longer a good approximation of true risk. It possible that we end up with a low empirical risk but a high true risk after training (aka overfitting).

Overfitting: when empirical risk is low, but true risk is high

- training data fits well
- true function fits poorly
- learned function is very different for different training sets, even if the training sets come from the same distribution

Potential causes:

- can happen if dataset is to small
- can happen if the model is too high capacity (i.e. there are many possible function approximations that can match the data)

Underfitting: when empirical risk is high, and true risk is high

- traing data fits poorly
- true function fits poorly
- learned function stays the same for different training sets, even if you increase dataset size

Potential causes:

- can happen if the model is too low capacity (i.e. there are no function approximations that match the data well)
- can happen if optmizer is not configured well enough

Mathematical Derivation of Bias and Variance

$$\text{Reminder: } p(D) = \prod_i p(x_i)p(y_i|x_i)$$

Consider the **expected error** of the algorithm with respect to the **distribution of datasets**.

$$E_{D\sim p(D)}[||f_D(x) - f(y)||^2]$$

Where $f_D(x)$ is the function found by the learning algorithm for dataset D and $f(y)$ is the true function.

$$E_{D\sim p(D)}[||f_D(x) - f(y)||^2] = \sum_D p(D)||f_D(x) - f(y)||^2$$

Why is this value useful?

We want to understand how well our **algorithm** does independently of the **particular (random) choice of dataset**.

Note: this is a theoretical exercise, it's not practical to compute this in the real world

$$\text{Let } \bar{f}(x) := E_{D\sim p(D)}[f_D(x)] \quad \text{the average function learned by our algorithm}$$

$$\begin{gathered}
E_{D\sim p(D)}[||f_D(x) - f(y)||^2] \\
= E_{D\sim p(D)}[||f_D(x) - \bar{f}(x) + \bar{f}(x) - f(y)||^2] \\
= E_{D\sim p(D)}[||(f_D(x) - \bar{f}(x)) + (\bar{f}(x) - f(y))||^2] \\
= E_{D\sim p(D)}[||f_D(x) - \bar{f}(x)||^2] + E_{D\sim p(D)}[||\bar{f}(x) - f(y)||^2] + E_{D\sim p(D)}[2(f_D(x) - \bar{f}(x))^T(\bar{f}(x) - f(y))] \\
= E_{D\sim p(D)}[||f_D(x) - \bar{f}(x)||^2] + E_{D\sim p(D)}[||\bar{f}(x) - f(y)||^2] + E_{D\sim p(D)}[2(0)(\bar{f}(x) - f(y))] \\
= E_{D\sim p(D)}[||f_D(x) - \bar{f}(x)||^2] + E_{D\sim p(D)}[||\bar{f}(x) - f(y)||^2] + 0 \\
= E_{D\sim p(D)}[||f_D(x) - \bar{f}(x)||^2] + E_{D\sim p(D)}[||\bar{f}(x) - f(y)||^2]
\end{gathered}$$

$$\begin{gathered}
= E_{D\sim p(D)}[||f_D(x) - \bar{f}(x)||^2] + ||\bar{f}(x) - f(y)||^2 \\
\text{Expected Error} = \text{Variance} + \text{Bias}^2
\end{gathered}$$

$$\begin{gathered}
E_{D\sim p(D)}[||f_D(x) - \bar{f}(x)||^2] \rightarrow \text{Variance} \\
||\bar{f}(x) - f(y)||^2 \rightarrow \text{Bias}^2
\end{gathered}$$

Expected Error = Variance + Bias^2

Variance:

How much does the algorithm's predicted function change with the dataset (difference from the average function)?

If variance is too high $\rightarrow$ overfitting

Bias^2:

How far off is the algorithm's average function to the true function?

If bias is too high $\rightarrow$ underfitting

Key Question: how to regulate the tradeoff between variance and bias?

- Usually when you decrease one the other increases