泛化理论

第六章 Implicit Bias

§6.1.1 Overparameterization

@ 滕佳烨

Recall:

- Uniform Convergence: $L(\hat{f}) - \hat{L}(\hat{f}) \leq \sup_{f \in \mathcal{F}} \left| L(f) - \hat{L}(f) \right|$.
- Stability-based bound: similar dataset returns similar models
- PAC-Bayesian bound: Bayesian training (parameter following a distribution), $\sqrt{\frac{KL(P||Q)}{n}}$ (P: prior distribution, Q: posterior distribution)
- Information-based bound: mutual information $\sqrt{\frac{2\sigma^2}{n} I(S, W(S))}$

Today's topic:

Implicit bias: another view for generalization

Implicit Bias

One sentence: Algorithm may prefer some type of simple solutions.

Modern neural networks are usually overparameterized.

- Overparameterization: # parameters >> # samples

Therefore, there are infinitely many solutions.

However, these solutions are **not all** good to generalization.
Fortunately, for a given algorithm, it has some preference on a given type of solution.
And usually, these preferred solutions are **simple** solutions.

What does "**simple**" mean?

- low-norm; low gradient; max-margin; etc.

A simple case: overparameterized linear regression with gradient descent

Theorem (LR & GD). For **linear regression** with MSE loss $\ell(x, y; \theta) = \left(y - x^\top \theta\right)^2$, if we use full-batch **GD** with proper stepsize and zero initialization, the trained parameter converges to its **min-norm solution**.

Remark:

1. Training loss: $L(\theta; S) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i; \theta) = \frac{1}{n}\|Y - X\theta\|^2$
2. GD iterate: $\theta^{(t+1)} = \theta^{(t)} - \lambda \nabla_\theta L(\theta; S) = \theta^{(t)} - \frac{\lambda}{n} X^\top (Y - X\theta^{(t)})$
3. Proper stepsize: λ which can minimize the training loss to zero ($\lambda < \sigma_{min}(XX^\top)$)
4. Min-norm solution: $\min\|\theta\|$ such that $X\theta = Y$

Model & Data: **linear regression**
Algorithm: **gradient descent**
Simple form: **min-norm solutions**

A simple case: overparameterized linear regression with gradient descent

Theorem (LR & GD). For **linear regression** with MSE loss $\ell(x, y; \theta) = \left(y - x^\top \theta\right)^2$, if we use full-batch **GD** with proper stepsize and zero initialization, the trained parameter converges to its **min-norm solution**.

Proof: $\theta^{(t+1)} = \theta^{(t)} - \frac{\lambda}{n} X^\top \left(Y - X\theta^{(t)}\right)$

- For min-norm solution, we can write it as $\theta_{mm} = X^\top \left(XX^\top\right)^{-1} Y$ (assume $XX^\top$ has full rank for simplicity)
- Note that we can always write it as $\theta^{(t)} = X^\top v$ where v is a vector (see the iterate)
- Therefore, when $X\theta^{(\infty)} = Y = XX^\top v$, $v = \left(XX^\top\right)^{-1} Y$, therefore $\theta^{(\infty)} = \theta_{mm}$.

Note: when $XX^\top$ is not full rank, one can reduce it to the full rank case.

Take-away messages

(a) Implicit Bias: when there are infinitely many solutions, algorithm prefers some simple solutions.
(b) Simple Case: Linear regression + gradient descent → min-norm solutions

All the slides will be available at www.tengjiaye.com/generalization soon.

@ 滕佳烨

Thanks!