TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2022

Implicit Regularization

Implicit Regularization

Any stochastic learning algorithm, such as SGD, determines a stochastic mapping from training data to models.

The algorithm, especially with early stopping, can implicitly incorporate a preference or bias for models.

Implicit Regularization in Linear Regression

Linear regression (minimizing the L_2 loss of a linear predictor) where we have more parameters than data points has many solutions.

But SGD converges to the minimum norm solution (L_2-regularized solution) without the need for explicit regularization.

Implicit Regularization in Linear Regression

For linear regression SGD maintains the invariant that Φ is a linear combination of the (small number of) training vectors.

Any zero-loss (squared loss) solution can be projected on the span of training vectors to give a smaller (or no larger) norm solution.

It can be shown that when the training vectors are linearly independent any zero loss solution in the span of the training vectors is a least-norm solution.

Implict Priors

Let A be any algorithm mapping a training set Train to a probability density $p(\Phi|\text{Train})$.

For example, the algorithm might be SGD where we add a small amount of noise to the final parameter vector so that $p(\Phi|\text{Train})$ is a smooth density.

But in general we can consider any leaning algorithm that produces a smooth density $p(\Phi|\text{Train})$.

Implicit Priors

Drawing Train from Pop^N and Φ from $P(\Phi|\mathrm{Train})$ defines a joint distribution on Train and Φ. We can take the marginal distribution on Φ to be a prior distribution (independent of any training data).

$$p(\Phi) = E_{\left(\mathrm{Train}\sim\mathrm{Pop}^N\right)}\; p(\Phi\;|\mathrm{Train})$$

It can be shown that the implicit prior $p(\Phi)$ is an optimal prior for the PAC-Bayesian generalization guarantees applied to the algorithm defining $p(\Phi|\mathrm{Train})$

A PAC-Bayes Analysis of Implicit Regularization

$$\mathcal{L}(\mathrm{Train}) = E_{\langle x,y\rangle \sim \mathrm{Pop},\ \Phi \sim p(\Phi|\mathrm{Train})}\, \mathcal{L}(\Phi, x, y)$$

$$\hat{\mathcal{L}}(\mathrm{Train}) = E_{\langle x,y\rangle \sim \mathrm{Train},\ \Phi \sim p(\Phi|\mathrm{Train})}\, \mathcal{L}(\Phi, x, y)$$

A PAC-Bayes Analysis of Implicit Regularization

With probability at least $1-\delta$ over the draw of Train we have

$$\mathcal{L}(\mathrm{Train}) \;\leq\; \frac{10}{9}\left(\hat{\mathcal{L}}(\mathrm{Train}) + \frac{5L_{\max}}{N_{\mathrm{Train}}}\left(KL(p(\Phi|\mathrm{Train}), p(\Phi))\right) + \ln\frac{1}{\delta}\right)$$

$$= \;\frac{10}{9}\left(\hat{\mathcal{L}}(\mathrm{Train}) + \frac{5L_{\max}}{N_{\mathrm{Train}}}\left(I(\Phi, \mathrm{Train}) + \ln\frac{1}{\delta}\right)\right)$$

There is no obvious way to calculate this guarantee.

However, it can be shown that $p(\Phi)$ is the optimal PAC-Bayeisan prior for given algorithm run on training data data drawn from Pop^N.

END