TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Early Stopping meets Shrinkage

L_1 Regularization and Sparsity

Ensembles

Shrinkage meets Early Stopping

Early stopping can limit $||\Phi||$.

But early stopping more directly limits $||\Phi - \Phi_{\mathrm{init}}||$.

It seems better to take the prior on Φ to be

$$p(\Phi) \propto \exp\left(-\frac{||\Phi - \Phi_{\mathrm{init}}||^2}{2\sigma^2}\right)$$

giving

$$\Phi_{t+1} = \Phi_t - \eta\hat{g} - \gamma(\Phi_t - \Phi_{\mathrm{init}})$$

L_1 Regularization

$$p(\Phi) \propto e^{-\lambda||\Phi||_1} \qquad ||\Phi||_1 = \sum_i |\Phi_i|$$

$$\Phi^* = \underset{\Phi}{\text{argmax}} \quad p(\Phi) \prod_i P_\Phi(y_i|x_i)$$

$$\Phi^* = \underset{\Phi}{\text{argmin}} \quad \left(\sum_i -\ln P_\Phi(y_i|x_i) \right) + \lambda||\Phi||_1$$

$$\Phi^* = \underset{\Phi}{\text{argmin}} \quad \hat{\mathcal{L}}(\Phi) + \frac{\lambda}{N_{\text{Train}}}||\Phi||_1$$

L_1 Regularization

$$\Phi^* \;=\; \underset{\Phi}{\mathrm{argmin}} \quad \hat{\mathcal{L}}(\Phi) \;+\; \frac{\lambda}{N_{\mathrm{Train}}}||\Phi||_1$$

$$\Phi_i \;\texttt{-=}\; \eta\left(\hat{g}_i + \frac{\lambda}{N_{\mathrm{Train}}}\,\mathrm{sign}(\Phi_i)\right)$$

$$\eta \;=\; (1-\mu)B\eta_0$$

Sparsity

$$\Phi_i \; \texttt{-=} \; \eta \left(\hat{g}_i + \frac{\lambda}{N_{\mathrm{Train}}} \, \mathrm{sign}(\Phi_i) \right)$$

For Φ^* the gradient of the objective, and hence the average update, must be zero:

$$\begin{aligned} \Phi_i^* &= 0 & \text{if } |g_i| < \lambda/N_{\mathrm{Train}} \\ g_i &= -(\lambda/N_{\mathrm{Train}})\mathrm{sign}(\Phi_i) & \text{otherwise} \end{aligned}$$

But in practice Φ_i will never be exactly zero.

Ensembles

Train several models $\mathrm{Ens} = (\Phi_1, \ldots, \Phi_K)$ from different initializations and/or under different meta parameters.

We define the ensemble model by

$$P_{\mathrm{Ens}}(y|x) = \frac{1}{K} \sum_k P_{\Phi_k}(y|x) = E_k \; P_k(y|x)$$

Ensemble models almost always perform better than any single model.

Ensembles Under Cross Entropy Loss

$$
\begin{aligned}
\mathcal{L}\,(P_{\mathrm{Ens}}) &= E_{\langle x,\,y\rangle\sim\mathrm{Pop}}\ \ -\ln P_{\mathrm{Ens}}(y|x) \\
&= E_{\langle x,\,y\rangle\sim\mathrm{Pop}}\ \ -\ln E_k P_k(y|x) \\
&\leq E_{\langle x,\,y\rangle\sim\mathrm{Pop}}\ E_k\ \ -\ln P_k(y|x) \\
&= E_k\ \mathcal{L}(P_k)
\end{aligned}
$$

Ensembles Under Cross Entropy Loss

It is important to note that

$$-\ln E_k \; P_k(y|x) \quad \leq E_k - \ln P_k(y|x)$$

for each individual pair $\langle x, y \rangle$.

$\forall z \; f(z) \leq g(z)$ is stronger than $(E_z \; f(z)) \leq (E_z \; g(z))$.

This may explain why in practice an ensemble model is typically better than any single component model.

END