CSC413: Homework 2

Tianyu Du (1003801647)

2020/02/06 at 16:34:19

1 Optimization

1.1 Stochastic Gradient Descent (SGD)

1.1.1 Minimum Norm Solution

Answer. Recall from the previous homework that the solution found by gradient descent in the over-parameterized situation was

$$\mathbf{w}^* = X^T(XX^T)^{-1}\mathbf{t} \tag{1.1}$$

Moreover, the solution $\mathbf{w}^*$ reached by gradient descent was in the row space of X (i.e., the span of rows of X), and $\mathbf{w}^*$ is a zero loss solution.
For one single updating using $\mathbf{x}_i$ in the t^{th} iteration in the SGD process, the gradient is

$$\begin{align}
\nabla_{\mathbf{w}_t}\mathcal{L}_i &= \frac{\partial}{\partial \mathbf{w}_t}||\mathbf{w}_t^T\mathbf{x}_i - t_i||_2^2 \tag{1.2}\\
&= \underbrace{2(\mathbf{w}_t^T\mathbf{x}_i - t_i)}_{\in\mathbb{R}}\mathbf{x}_i \tag{1.3}\\
\implies -\eta\nabla_{\mathbf{w}_t}\mathcal{L}_i &\in Row(X) \tag{1.4}
\end{align}$$

Provided that the starting point $\mathbf{w}_0 = \mathbf{0} \in Row(X)$, for every t, $\mathbf{w}_t \in Row(X)$ by an inductive argument.
Because $\mathbf{w}_t \in Row(X)$, let

$$\begin{align}
\mathbf{w}_t &= X^T\mathbf{r}_t \quad \mathbf{r}_t \in \mathbb{R}^n \tag{1.5}\\
\hat{\mathbf{w}} &= X^T\hat{\mathbf{r}} \quad \hat{\mathbf{r}} \in \mathbb{R}^n \tag{1.6}
\end{align}$$

Suppose the solution $\hat{\mathbf{w}}$ reached by SGD is a zero loss solution, that is, $X\hat{\mathbf{w}} = \mathbf{t}$. Then

$$XX^T\hat{\mathbf{r}} = \mathbf{t} \tag{1.7}$$

Because $\hat{\mathbf{w}}$ is a global minimum and SGD converges to this point, then it must be the case that

$$\nabla_{\hat{\mathbf{w}}}\mathcal{L}_i(\mathbf{x}_i, \hat{\mathbf{w}}) = 0 \quad \forall i \in \{1, \cdots, n\} \tag{1.8}$$

That is, one additional iteration of SGD does not improve performance no matter which sample is chosen. This is the same as

$$
\begin{aligned}
\nabla_{r_i} ||XX^T\hat{\mathbf{r}} - \mathbf{t}||_2^2 = 0 \quad \forall i \in \{1, \cdots, n\} && (1.9)\\
\iff \nabla_{\mathbf{r}} ||XX^T\hat{\mathbf{r}} - \mathbf{t}||_2^2 = 0 && (1.10)\\
\implies (XX^T\hat{\mathbf{r}} - \mathbf{t})^T XX^T = 0 && (1.11)\\
\implies XX^T(XX^T\hat{\mathbf{r}} - \mathbf{t}) = 0^T && (1.12)\\
\implies XX^TXX^T\hat{\mathbf{r}} = XX^T\mathbf{t} && (1.13)\\
\implies \hat{\mathbf{r}} = (XX^T)^{-1}\mathbf{t} && (1.14)\\
\implies \hat{w} = X^T\hat{\mathbf{r}} = X^T(XX^T)^{-1}\mathbf{t} && (1.15)\\
= \mathbf{w}^* && (1.16)
\end{aligned}
$$

■

2 Gradient-Based Hyper-Parameter Optimization

3 Convolutional Neutral Networks

3.1 Convolutional Filters

Answer.

$$
\mathbf{I} * \mathbf{J} = \begin{bmatrix} -1 & 2 & 2 & -2 & 0 \\ -2 & 1 & 0 & 2 & -1 \\ 3 & 0 & 0 & 1 & -1 \\ -2 & 2 & 0 & 2 & -1 \\ 0 & -2 & 3 & -2 & 0 \end{bmatrix} \tag{3.1}
$$

This convolutional filter detect edges. ■

3.2 Size of ConvNets

Answer. ■