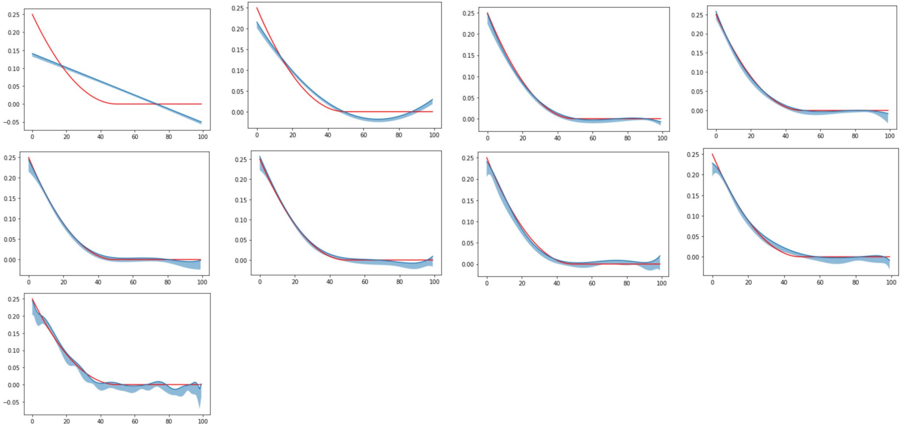
Deep Learning

2.3 Bias-Variance trade-off

Dr. Konda Reddy Mopuri
kmopuri@iittp.ac.in
Dept. of CSE, IIT Tirupati
Aug-Dec 2021

Visualize overfitting



1. If we formalize the observations

1. If we formalize the observations
2. Let x be fixed, y be the true value associated with it, $f*$ is what we leaned from dataset $\mathcal{D}$, and $Y = f*(x)$ is the predicted value

1. If we formalize the observations
2. Let x be fixed, y be the true value associated with it, $f*$ is what we leaned from dataset $\mathcal{D}$, and $Y = f*(x)$ is the predicted value
3. Let's consider that $\mathcal{D}$ is a random variable, then $f*$ and Y get random

Consider

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}((Y-y)^2) =& \mathbb{E}_{\mathcal{D}}(Y^2 - 2Yy + y^2) \\
=& \mathbb{E}_{\mathcal{D}}(Y^2) - 2\mathbb{E}_{\mathcal{D}}(Y)y + y^2 \\
=& \mathbb{E}_{\mathcal{D}}(Y^2) - \mathbb{E}_{\mathcal{D}}(Y)^2 + \mathbb{E}_{\mathcal{D}}(Y)^2 - 2\mathbb{E}_{\mathcal{D}}(Y)y + y^2 \\
=& \color{red}{\mathbb{E}_{\mathcal{D}}(Y^2) - \mathbb{E}_{\mathcal{D}}(Y)^2} + \color{blue}{\mathbb{E}_{\mathcal{D}}(Y)^2 - 2\mathbb{E}_{\mathcal{D}}(Y)y + y^2} \\
=& \color{blue}{(\mathbb{E}_{\mathcal{D}}(Y) - y)^2} + \color{red}{\mathbb{V}_{\mathcal{D}}(Y)}
\end{aligned}
$$

Bias-Variance Decomposition

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}((Y-y)^2) =& \mathbb{E}_{\mathcal{D}}(Y^2 - 2Yy + y^2) \\
=& \mathbb{E}_{\mathcal{D}}(Y^2) - 2\mathbb{E}_{\mathcal{D}}(Y)y + y^2 \\
=& \mathbb{E}_{\mathcal{D}}(Y^2) - \mathbb{E}_{\mathcal{D}}(Y)^2 + \mathbb{E}_{\mathcal{D}}(Y)^2 - 2\mathbb{E}_{\mathcal{D}}(Y)y + y^2 \\
=& {\color{red}\mathbb{E}_{\mathcal{D}}(Y^2) - \mathbb{E}_{\mathcal{D}}(Y)^2} + {\color{blue}\mathbb{E}_{\mathcal{D}}(Y)^2 - 2\mathbb{E}_{\mathcal{D}}(Y)y + y^2} \\
=& {\color{blue}(\mathbb{E}_{\mathcal{D}}(Y) - y)^2} + {\color{red}\mathbb{V}_{\mathcal{D}}(Y)}
\end{aligned}
$$

1. This is known as Bias-Variance decomposition

Bias-Variance Decomposition

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}((Y-y)^2) =& \mathbb{E}_{\mathcal{D}}(Y^2 - 2Yy + y^2) \\
=& \mathbb{E}_{\mathcal{D}}(Y^2) - 2\mathbb{E}_{\mathcal{D}}(Y)y + y^2 \\
=& \mathbb{E}_{\mathcal{D}}(Y^2) - \mathbb{E}_{\mathcal{D}}(Y)^2 + \mathbb{E}_{\mathcal{D}}(Y)^2 - 2\mathbb{E}_{\mathcal{D}}(Y)y + y^2 \\
=& \mathbb{E}_{\mathcal{D}}(Y^2) - \mathbb{E}_{\mathcal{D}}(Y)^2 + \mathbb{E}_{\mathcal{D}}(Y)^2 - 2\mathbb{E}_{\mathcal{D}}(Y)y + y^2 \\
=& (\mathbb{E}_{\mathcal{D}}(Y) - y)^2 + \mathbb{V}_{\mathcal{D}}(Y)
\end{aligned}$$

1. This is known as Bias-Variance decomposition
2. Bias term quantifies how much the model fits the data on average

Bias-Variance Decomposition

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}((Y-y)^2) =& \mathbb{E}_{\mathcal{D}}(Y^2 - 2Yy + y^2) \\ =& \mathbb{E}_{\mathcal{D}}(Y^2) - 2\mathbb{E}_{\mathcal{D}}(Y)y + y^2 \\ =& \mathbb{E}_{\mathcal{D}}(Y^2) - \mathbb{E}_{\mathcal{D}}(Y)^2 + \mathbb{E}_{\mathcal{D}}(Y)^2 - 2\mathbb{E}_{\mathcal{D}}(Y)y + y^2 \\ =& \mathbb{E}_{\mathcal{D}}(Y^2) - \mathbb{E}_{\mathcal{D}}(Y)^2 + \mathbb{E}_{\mathcal{D}}(Y)^2 - 2\mathbb{E}_{\mathcal{D}}(Y)y + y^2 \\ =& (\mathbb{E}_{\mathcal{D}}(Y) - y)^2 + \mathbb{V}_{\mathcal{D}}(Y)\end{aligned}$$

1. This is known as Bias-Variance decomposition
2. Bias term quantifies how much the model fits the data on average
3. Variance term quantifies how much the model changes across datasets

Bias-Variance Trade-off

1. Reducing the capacity makes f^* fit the data less on average, which increases the bias term

Bias-Variance Trade-off

1. Reducing the capacity makes f^* fit the data less on average, which increases the bias term
2. Increasing the capacity makes f^* vary a lot with the training data, which increases the variance term