Module 8.3 : True error and Model complexity

Using Stein's Lemma (and some trickery) we can show that

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{f}(x_i) - f(x_i)) = \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x_i)}{\partial y_i}$$

Using Stein's Lemma (and some trickery) we can show that

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{f}(x_i) - f(x_i)) = \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x_i)}{\partial y_i}$$

- When will $\frac{\partial \hat{f}(x_i)}{\partial y_i}$ be high? When a small change in the observation causes a large change in the estimation($\hat{f}$)

Using Stein's Lemma (and some trickery) we can show that

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{f}(x_i) - f(x_i)) = \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial\hat{f}(x_i)}{\partial y_i}$$

- When will $\frac{\partial\hat{f}(x_i)}{\partial y_i}$ be high? When a small change in the observation causes a large change in the estimation($\hat{f}$)
- Can you link this to model complexity?

Using Stein's Lemma (and some trickery) we can show that

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{f}(x_i) - f(x_i)) = \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x_i)}{\partial y_i}$$

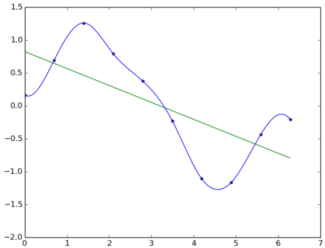
- When will $\frac{\partial \hat{f}(x_i)}{\partial y_i}$ be high? When a small change in the observation causes a large change in the estimation($\hat{f}$)
- Can you link this to model complexity?
- Yes, indeed a complex model will be more sensitive to changes in observations whereas a simple model will be less sensitive to changes in observations

Using Stein's Lemma (and some trickery) we can show that

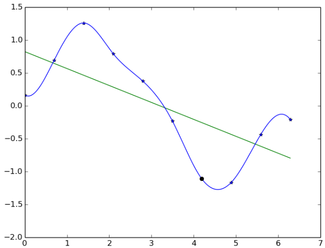
$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{f}(x_i) - f(x_i)) = \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x_i)}{\partial y_i}$$

- When will $\frac{\partial \hat{f}(x_i)}{\partial y_i}$ be high? When a small change in the observation causes a large change in the estimation($\hat{f}$)
- Can you link this to model complexity?
- Yes, indeed a complex model will be more sensitive to changes in observations whereas a simple model will be less sensitive to changes in observations
- Hence, we can say that
  true error = empirical train error + small constant + Ω(model complexity)

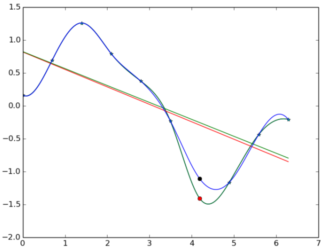
- Let us verify that indeed a complex model is more sensitive to minor changes in the data



- Let us verify that indeed a complex model is more sensitive to minor changes in the data
- We have fitted a simple and complex model for some given data



- Let us verify that indeed a complex model is more sensitive to minor changes in the data
- We have fitted a simple and complex model for some given data
- We now change one of these data points



- Let us verify that indeed a complex model is more sensitive to minor changes in the data
- We have fitted a simple and complex model for some given data
- We now change one of these data points
- The simple model does not change much as compared to the complex model

- Hence while training, instead of minimizing the training error $\mathscr{L}_{train}(\theta)$ we should minimize

$$\min_{w.r.t\ \theta} \mathscr{L}_{train}(\theta) + \Omega(\theta) = \mathscr{L}(\theta)$$

- Hence while training, instead of minimizing the training error $\mathscr{L}_{train}(\theta)$ we should minimize

$$\min_{w.r.t\ \theta} \mathscr{L}_{train}(\theta) + \Omega(\theta) = \mathscr{L}(\theta)$$

- Where $\Omega(\theta)$ would be high for complex models and small for simple models

- Hence while training, instead of minimizing the training error $\mathscr{L}_{train}(\theta)$ we should minimize

$$\min_{w.r.t\ \theta} \mathscr{L}_{train}(\theta) + \Omega(\theta) = \mathscr{L}(\theta)$$

- Where $\Omega(\theta)$ would be high for complex models and small for simple models
- $\Omega(\theta)$ acts as an approximate for $\frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x_i)}{\partial y_i}$

- Hence while training, instead of minimizing the training error $\mathscr{L}_{train}(\theta)$ we should minimize

$$\min_{w.r.t\ \theta} \mathscr{L}_{train}(\theta) + \Omega(\theta) = \mathscr{L}(\theta)$$

- Where $\Omega(\theta)$ would be high for complex models and small for simple models

- $\Omega(\theta)$ acts as an approximate for $\frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x_i)}{\partial y_i}$

- This is the basis for all regularization methods

- Hence while training, instead of minimizing the training error $\mathscr{L}_{train}(\theta)$ we should minimize

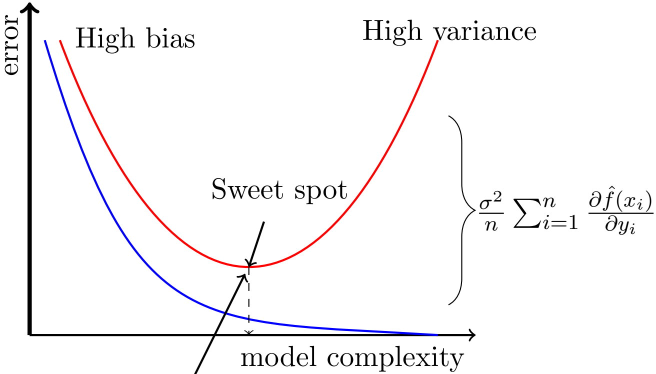
$$\min_{w.r.t\ \theta} \mathscr{L}_{train}(\theta) + \Omega(\theta) = \mathscr{L}(\theta)$$

- Where $\Omega(\theta)$ would be high for complex models and small for simple models

- $\Omega(\theta)$ acts as an approximate for $\frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x_i)}{\partial y_i}$

- This is the basis for all regularization methods

- We can show that L_1 regularization, L_2 regularization, early stopping and injecting noise in input are all instances of this form of regularization.



$\Omega(\theta)$ should ensure that model has reasonable complexity

- Why do we care about this bias variance tradeoff and model complexity?

- Why do we care about this bias variance tradeoff and model complexity?
- Deep Neural networks are highly complex models.

- Why do we care about this bias variance tradeoff and model complexity?
- Deep Neural networks are highly complex models.
- Many parameters, many non-linearities.

- Why do we care about this bias variance tradeoff and model complexity?

- Deep Neural networks are highly complex models.
- Many parameters, many non-linearities.
- It is easy for them to overfit and drive training error to 0.

- Why do we care about this bias variance tradeoff and model complexity?

- Deep Neural networks are highly complex models.
- Many parameters, many non-linearities.
- It is easy for them to overfit and drive training error to 0.
- Hence we need some form of regularization.

Different forms of regularization

- L_2 regularization

Different forms of regularization

- L_2 regularization
- Dataset augmentation

Different forms of regularization

- L_2 regularization
- Dataset augmentation
- Parameter Sharing and tying

Different forms of regularization

- L_2 regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs

Different forms of regularization

- L_2 regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs

Different forms of regularization

- L_2 regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping

Different forms of regularization

- L_2 regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods

Different forms of regularization

- L_2 regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout