Chapter 5
Machine Learning Basics

MILab Undergraduate student, TaeHyeon Kim

2023. 03. 07

Overview

Chapter 5. Machine Learning Basics

- 5.1 Learning Algorithms
- 5.2 Capacity, Overfitting and Underfitting
- 5.3 Hyperparameters and Validatation Sets
- 5.4 Estimators, Bias and Variance
- 5.5 Maximum Likelihood Estimation
- 5.6 Bayesian Statistics
- 5.7 Supervised Learning Algorithms
- 5.8 Unsupervised Learning Algorithms
- 5.9 Stochastic Gradient Descent
- 5.10 Building a Machine Learning Algorithm
- 5.11 Challenges Motivating Deep Learning

5.1 Learning Algorithms

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Text 1 – Description of Machine Learning Algorithm

- Experience E : The type of input data (supervised or not)

- Task T : The types of training algorithm

  \- Classification, Classification with missing inputs, Regression, Transcription, Machine translation, Structured output, Anomaly detection, Synthesis and sampling, Imputation of missing values, Denoising, Density estimation or probability mass function estimation

- Performance P : The efficiency of learning (Accuracy, Error rate)

5.1 Learning Algorithms (Example)

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Text 1 – Description of Machine Learning Algorithm

$$\hat{y} = \boldsymbol{w}^\top \boldsymbol{x}$$

- Experience E : Supervised Learning (x and y is given.)
- Task T : Linear Regression
- Performance P : Mean Squared Error

$$\text{MSE}_{\text{test}} = \frac{1}{m} \sum_i (\hat{\boldsymbol{y}}^{(\text{test})} - \boldsymbol{y}^{(\text{test})})_i^2 .$$

5.1 Learning Algorithms (Example)

"A computer program is said to learn from $\boxed{\text{experience } E}$ with respect to some class of $\boxed{\text{tasks } T}$ and $\boxed{\text{performance measure } P}$, if its performance at tasks in T, as measured by P, improves with experience E."

Text 1 – Description of Machine Learning Algorithm

$$\hat{y} = \boldsymbol{w}^\top \boldsymbol{x}$$

You can also get the ***w*** value from the calculating 'Mean Squared Error'

$$\nabla_{\boldsymbol{w}} \mathrm{MSE}_{\mathrm{train}} = 0$$

$$\Rightarrow \nabla_{\boldsymbol{w}} \frac{1}{m} ||\hat{\boldsymbol{y}}^{(\mathrm{train})} - \boldsymbol{y}^{(\mathrm{train})}||_2^2 = 0$$

$$\Rightarrow \frac{1}{m} \nabla_{\boldsymbol{w}} ||\boldsymbol{X}^{(\mathrm{train})}\boldsymbol{w} - \boldsymbol{y}^{(\mathrm{train})}||_2^2 = 0 \tag{5.8}$$

$$\Rightarrow \nabla_{\boldsymbol{w}} \left(\boldsymbol{X}^{(\mathrm{train})}\boldsymbol{w} - \boldsymbol{y}^{(\mathrm{train})}\right)^\top \left(\boldsymbol{X}^{(\mathrm{train})}\boldsymbol{w} - \boldsymbol{y}^{(\mathrm{train})}\right) = 0 \tag{5.9}$$

$$\Rightarrow \nabla_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{X}^{(\mathrm{train})\top}\boldsymbol{X}^{(\mathrm{train})}\boldsymbol{w} - 2\boldsymbol{w}^\top \boldsymbol{X}^{(\mathrm{train})\top}\boldsymbol{y}^{(\mathrm{train})} + \boldsymbol{y}^{(\mathrm{train})\top}\boldsymbol{y}^{(\mathrm{train})}\right) = 0 \tag{5.10}$$

$$\Rightarrow 2\boldsymbol{X}^{(\mathrm{train})\top}\boldsymbol{X}^{(\mathrm{train})}\boldsymbol{w} - 2\boldsymbol{X}^{(\mathrm{train})\top}\boldsymbol{y}^{(\mathrm{train})} = 0 \tag{5.11}$$

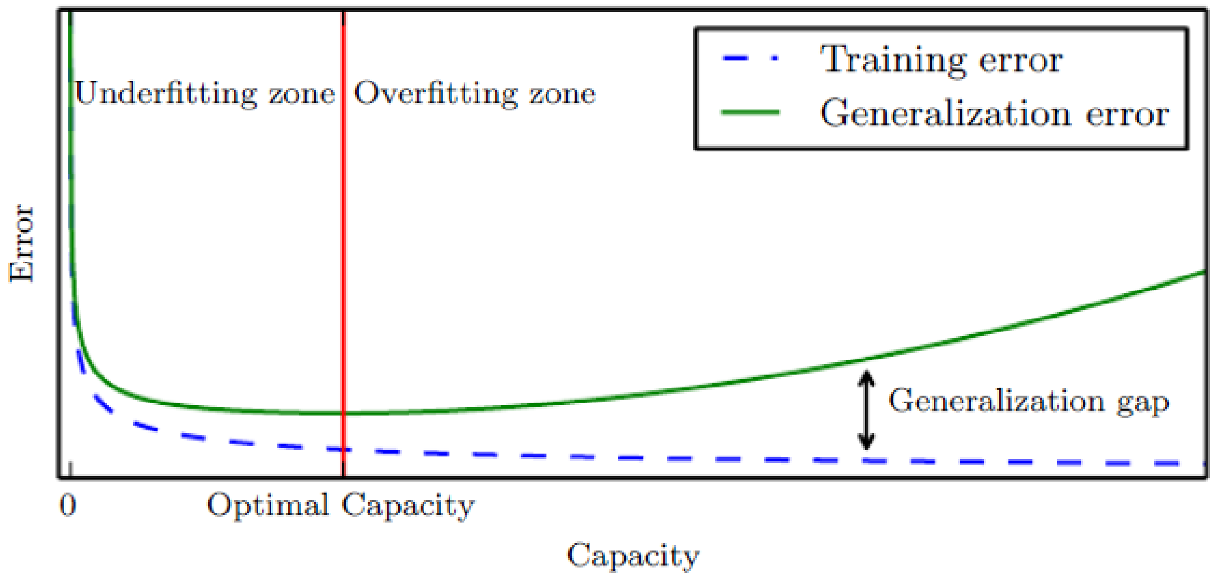
$$\Rightarrow \boldsymbol{w} = \left(\boldsymbol{X}^{(\mathrm{train})\top}\boldsymbol{X}^{(\mathrm{train})}\right)^{-1} \boldsymbol{X}^{(\mathrm{train})\top}\boldsymbol{y}^{(\mathrm{train})} \tag{5.12}$$

5.2 Capacity, Overfitting and Underfitting

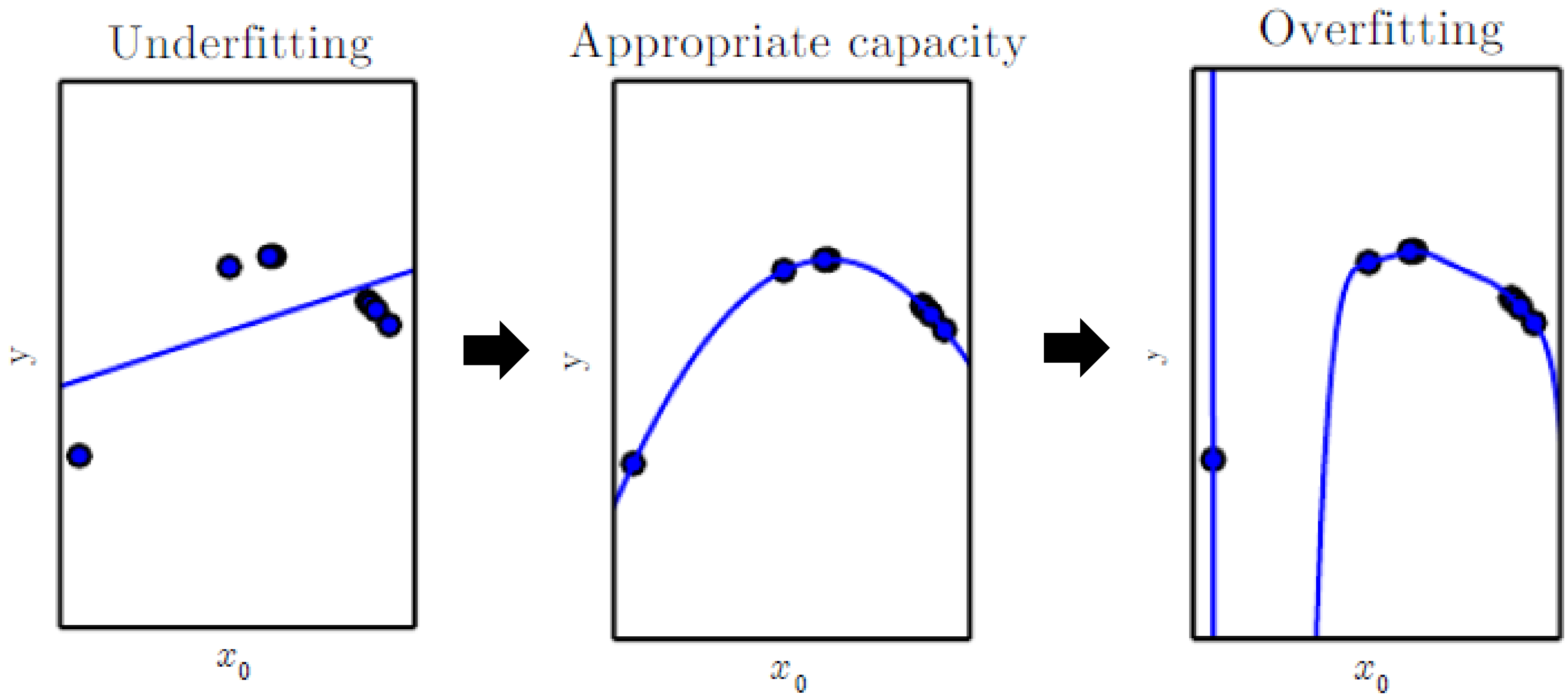
- Capacity : How could treat many variables?
- Underfitting : Training Error ↑
- Overfitting : Generalization Error ↑

< How to Solve >

1. Make the training error small.
2. Make the gap between training and test error small



5.2 Capacity, Overfitting and Underfitting



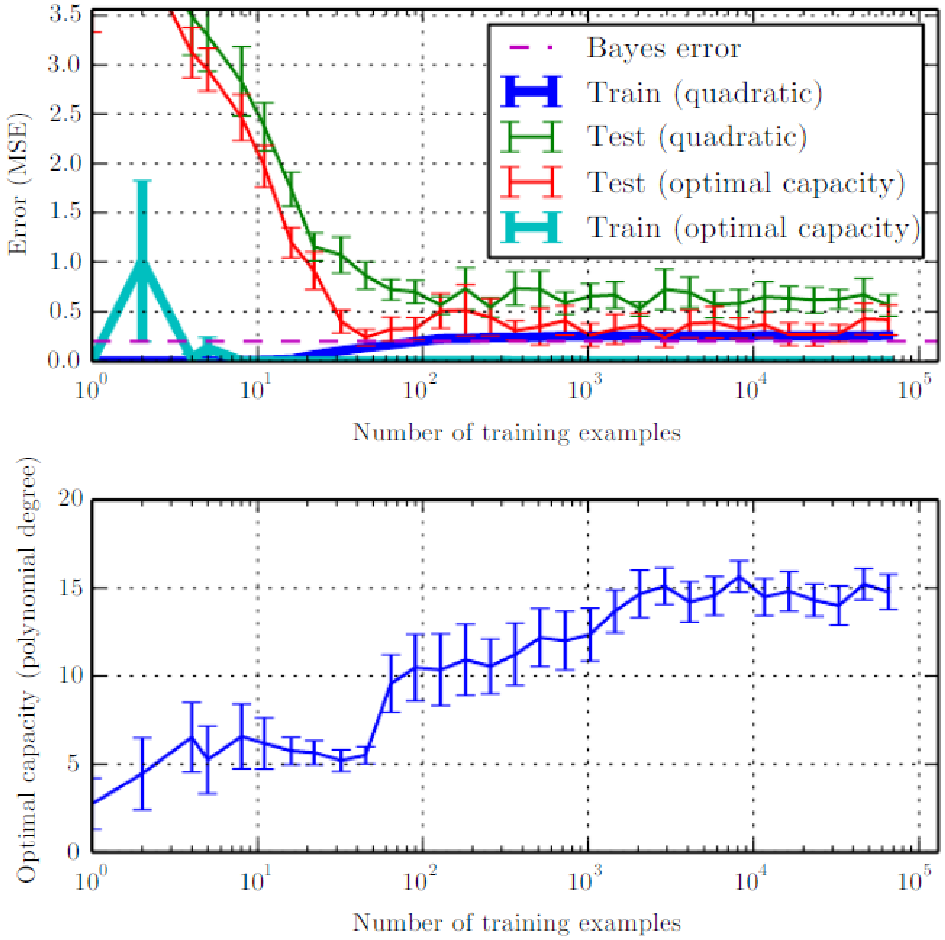
$$\hat{y} = b + wx.$$

$$\hat{y} = b + w_1 x + w_2 x^2.$$

$$\hat{y} = b + \sum_{i=1}^{9} w_i x^i.$$

5.2 Capacity, Overfitting and Underfitting

- The No Free Lunch Theorem



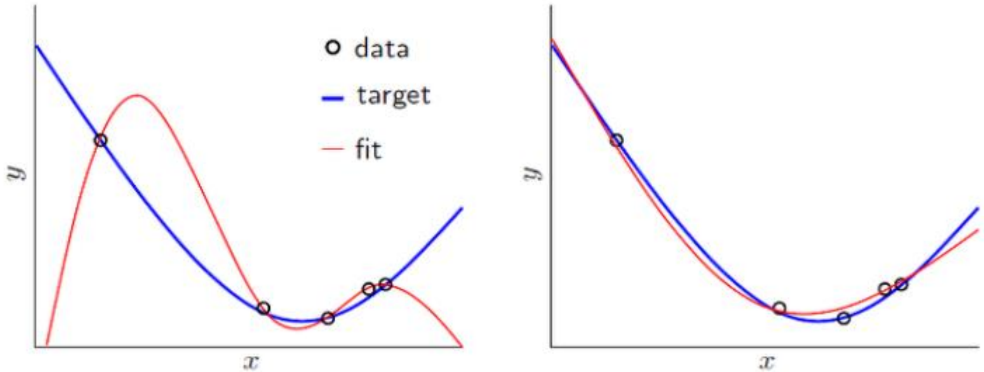
5.2 Capacity, Overfitting and Underfitting

- Use Regularization to prevent overfitting using L2 norm

$$J(w) = \mathrm{MSE}_{\mathrm{train}} + \lambda w^\top w$$

- Regularizer : $\Omega(w) = w^\top w$

- Make the value of W(weight) lower changing cost function

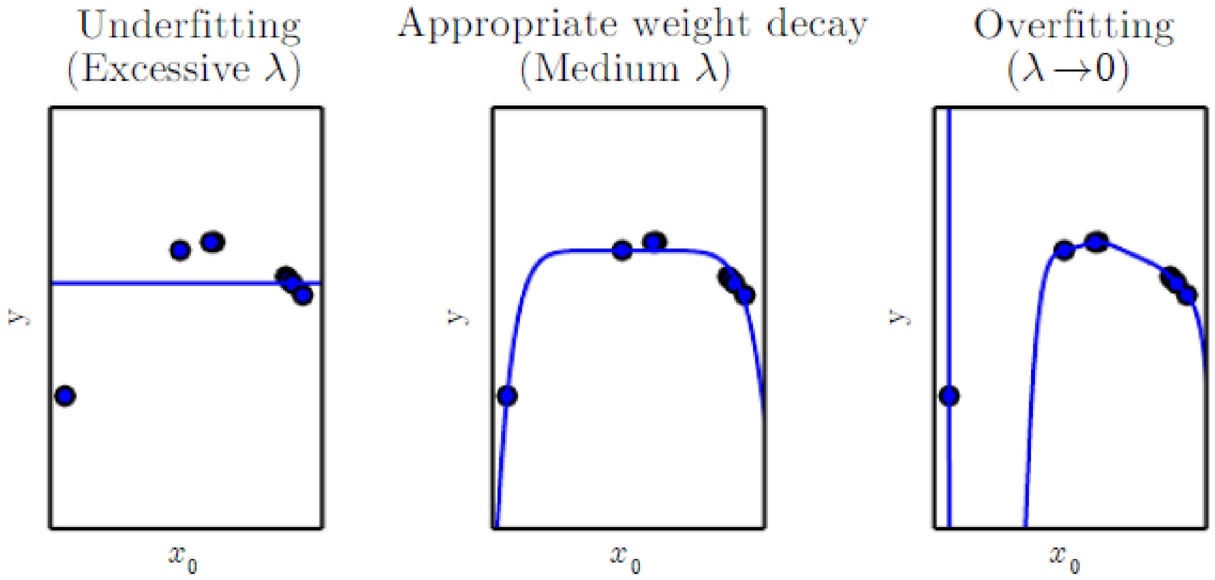


(a) without regularization (b) with regularization

5.3 Hyperparameters and Validatation Sets

- The definition of the ‘Model Parameter’
  - The ‘Model Parameter’ is a configuration variable that is internal to the model and whose value can be estimated from data. (ex. Avg, Stdev)

- The definition of the ‘Model Hyperparameter’
  - The ‘Model Hyperparameter’ is a configuration that is external to the model and whose value cannot be estimated from data. (ex. λ)



- Cross-Validation

5.4 Estimators, Bias and Variance

- Point Estimation

$$\hat{\boldsymbol{\theta}}_m = g(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}).$$

- Bias

$$\text{bias}(\hat{\boldsymbol{\theta}}_m) = \mathbb{E}(\hat{\boldsymbol{\theta}}_m) - \boldsymbol{\theta},$$

- Standard Error Mean

$$\text{SE}(\hat{\mu}_m) = \sqrt{\text{Var}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right]} = \frac{\sigma}{\sqrt{m}},$$

5.4 Estimators, Bias and Variance

- Get bias value of Mean and Get Variance at Bernoulli

$$P(x^{(i)};\theta) = \theta^{x^{(i)}}(1-\theta)^{(1-x^{(i)})}$$

$$\hat{\theta}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\begin{aligned}
\mathrm{bias}(\hat{\theta}_m) &= \mathbb{E}[\hat{\theta}_m] - \theta \\
&= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right] - \theta \\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[x^{(i)}\right] - \theta \\
&= \frac{1}{m}\sum_{i=1}^{m}\sum_{x^{(i)}=0}^{1}\left(x^{(i)}\theta^{x^{(i)}}(1-\theta)^{(1-x^{(i)})}\right) - \theta \\
&= \frac{1}{m}\sum_{i=1}^{m}(\theta) - \theta \\
&= \theta - \theta = 0
\end{aligned}$$

Bias value about Mean

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}_m) &= \mathrm{Var}\left(\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right) \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\mathrm{Var}\left(x^{(i)}\right) \\
&= \frac{1}{m^2}\sum_{i=1}^{m}(\theta(1-\theta)) \\
&= \frac{1}{m^2} m\theta(1-\theta) \\
&= \frac{1}{m}\theta(1-\theta)
\end{aligned}$$

Variance

5.4 Estimators, Bias and Variance

- Get bias value of Mean and Get Variance at Gaussian

$$p(x^{(i)}; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x^{(i)} - \mu)^2}{\sigma^2}\right).$$

$$\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\begin{aligned}
\text{bias}(\hat{\mu}_m) &= \mathbb{E}[\hat{\mu}_m] - \mu \\
&= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right] - \mu \\
&= \left(\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[x^{(i)}\right]\right) - \mu \\
&= \left(\frac{1}{m}\sum_{i=1}^{m} \mu\right) - \mu \\
&= \mu - \mu = 0
\end{aligned}$$

Bias value about Mean

$$\hat{\sigma}_m^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2,$$

$$\text{bias}(\hat{\sigma}_m^2) = \mathbb{E}[\hat{\sigma}_m^2] - \sigma^2.$$

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}_m^2] &= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2\right] \\
&= \frac{m-1}{m}\sigma^2
\end{aligned}$$

$$\tilde{\sigma}_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2$$

Variance

$$\begin{aligned}
\mathbb{E}[\tilde{\sigma}_m^2] &= \mathbb{E}\left[\frac{1}{m-1}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2\right] \\
&= \frac{m}{m-1}\mathbb{E}[\hat{\sigma}_m^2] \\
&= \frac{m}{m-1}\left(\frac{m-1}{m}\sigma^2\right) \\
&= \sigma^2.
\end{aligned}$$

Variance Estimator

5.4 Estimators, Bias and Variance

- Bias-Variance Trade-off

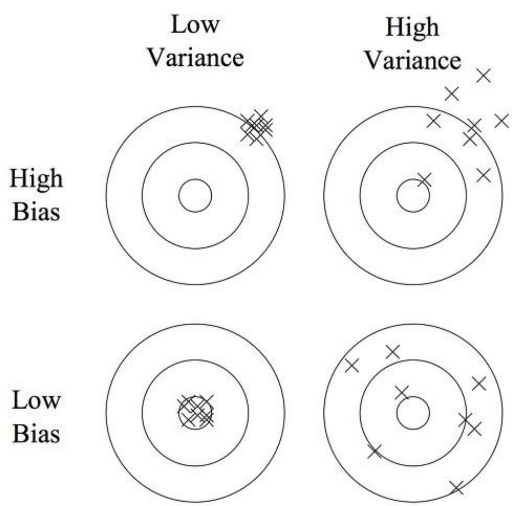
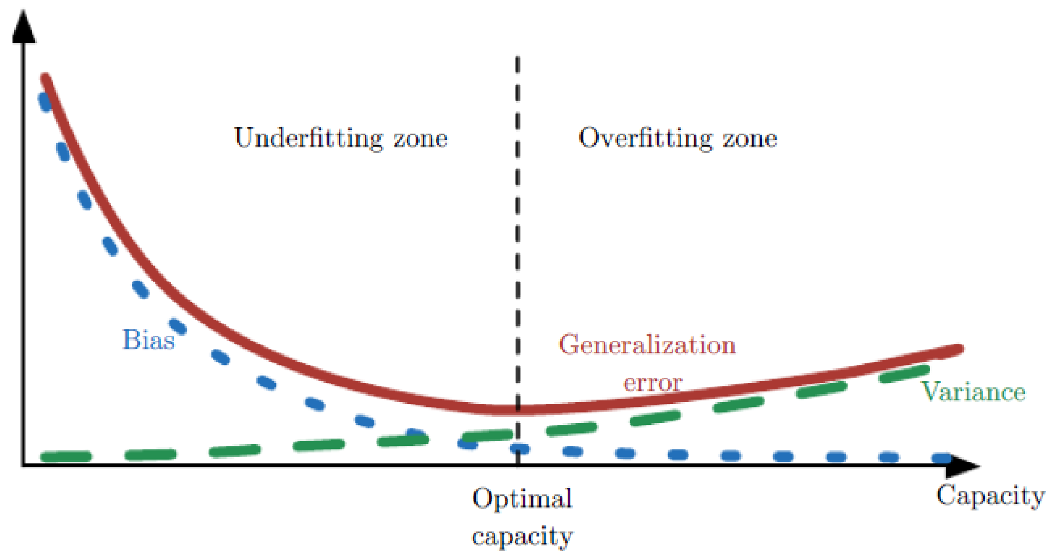


Image from Quora

$$\text{MSE} = \mathbb{E}\left[(\hat{\theta}_m - \theta)^2\right] = \text{Bias}(\hat{\theta}_m)^2 + \text{Var}(\hat{\theta}_m)$$



- Consistency
  - When training data set be greater, point estimates will go to original.

$$\text{plim}_{m\to\infty}\hat{\theta}_m = \theta$$

$$P(|\hat{\theta}_m - \theta| > \epsilon) \to 0$$

5.5 Maximum Likelihood Estimation

- Maximum Likelihood Estimation
  - Estimator could be analyzed using by **bias** and **variance**
  - Also, we could know how to extract some function for good estimator
  - One of the general principle is the **Maximum Likelihood Estimation**

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} p_{\mathrm{model}}(\mathbb{X}; \boldsymbol{\theta}),$$

$$= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p_{\mathrm{model}}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}).$$

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log p_{\mathrm{model}}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}).$$

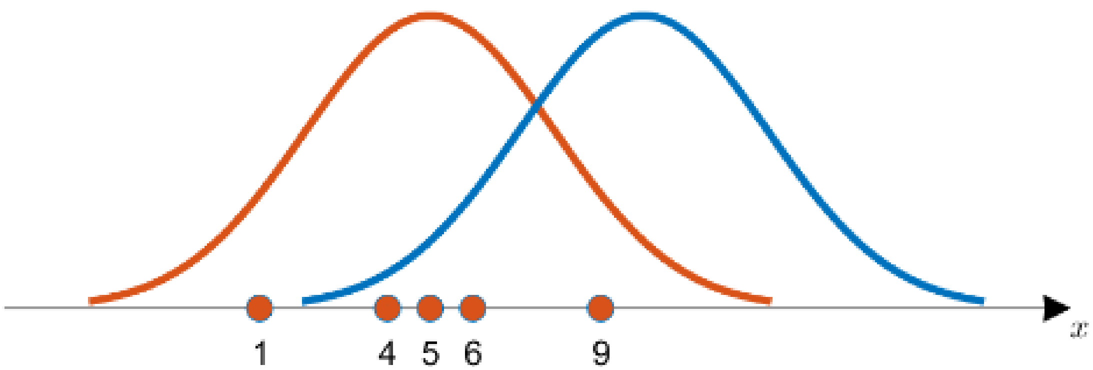
$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x}\sim\hat{p}_{\mathrm{data}}} \log p_{\mathrm{model}}(\boldsymbol{x}; \boldsymbol{\theta}).$$

$$D_{\mathrm{KL}}\left(\hat{p}_{\mathrm{data}} \| p_{\mathrm{model}}\right) = \mathbb{E}_{\mathbf{x}\sim\hat{p}_{\mathrm{data}}}\left[\log \hat{p}_{\mathrm{data}}(\boldsymbol{x}) - \log p_{\mathrm{model}}(\boldsymbol{x})\right].$$

$$-\mathbb{E}_{\mathbf{x}\sim\hat{p}_{\mathrm{data}}}\left[\log p_{\mathrm{model}}(\boldsymbol{x})\right],$$

5.5 Maximum Likelihood Estimation

- Maximum Likelihood Estimation
  - Estimator could be analyzed using by **bias** and **variance**
  - Also, we could know how to extract some function for good estimator
  - One of the general principle is the **Maximum Likelihood Estimation**



Which distribution represents the points as well? Red or Blue?

5.6 Bayesian Statistics

- Bayesian Statistics
  - In Frequentist Statistics, True Parameter θ is fixed and unknown, and usually estimate some data using variable the hat of θ.
  - But, in Bayesian Statistics, the probability means certainty. True parameter is uncertain, and it is used to random variable.
  - Prior Probability Distribution is represented to $p(\theta)$.

$$p(\boldsymbol{\theta} \mid x^{(1)}, \ldots, x^{(m)}) = \frac{p(x^{(1)}, \ldots, x^{(m)} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(x^{(1)}, \ldots, x^{(m)})}$$

$$p(x^{(m+1)} \mid x^{(1)}, \ldots, x^{(m)}) = \int p(x^{(m+1)} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid x^{(1)}, \ldots, x^{(m)}) d\boldsymbol{\theta}$$

5.6 Bayesian Statistics

- Bayesian Statistics
  - In contrast to MLE estimating point using θ, predict **using all over the distributions**
  - Prior Distribution : in 'roll the dice' case

$$p(\boldsymbol{\theta} \mid x^{(1)}, \dots, x^{(m)}) = \frac{p(x^{(1)}, \dots, x^{(m)} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(x^{(1)}, \dots, x^{(m)})}$$

$$p(x^{(m+1)} \mid x^{(1)}, \dots, x^{(m)}) = \int p(x^{(m+1)} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid x^{(1)}, \dots, x^{(m)}) d\boldsymbol{\theta}$$

5.6 Bayesian Statistics

- Maximum a Posteriori (MAP) Estimation

$$p(\boldsymbol{\theta} \mid x^{(1)}, \ldots, x^{(m)}) = \frac{p(x^{(1)}, \ldots, x^{(m)} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(x^{(1)}, \ldots, x^{(m)})}$$

$$p(x^{(m+1)} \mid x^{(1)}, \ldots, x^{(m)}) = \int p(x^{(m+1)} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid x^{(1)}, \ldots, x^{(m)}) d\boldsymbol{\theta}$$

$$\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

Use only one estimation using posterior distribution, and the others are not.

5.7 Supervised Learning Algorithms

- Supervised Learning Algorithms
  - The learning algorithm means that learns the relationship between the input and output data
  - Answer Label is contained in train data set

- Probabilistic Supervised Learning

5.8 Unsupervised Learning Algorithms

- Unsupervised Learning Algorithms
  - Learning without answer label

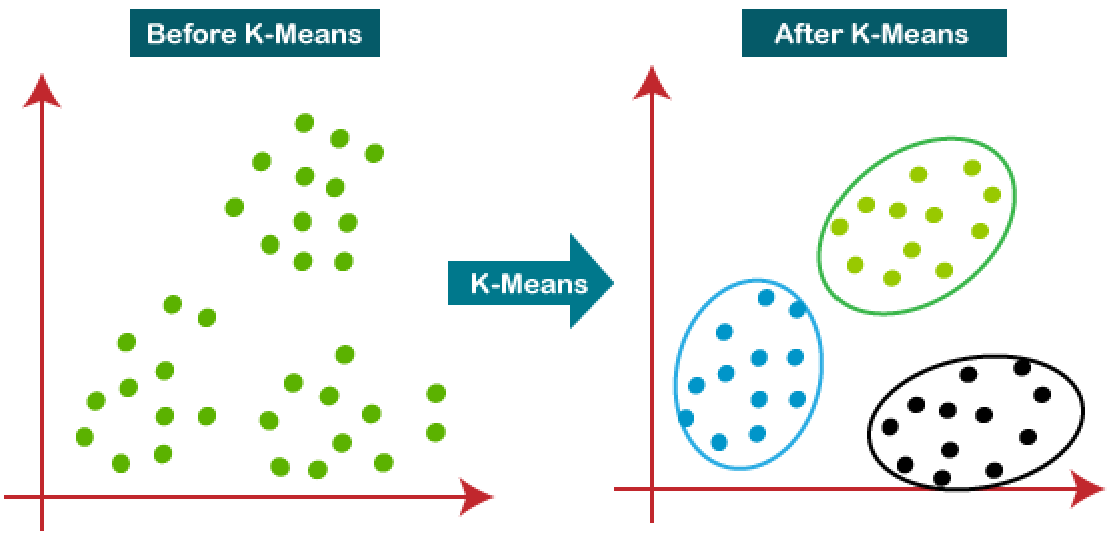
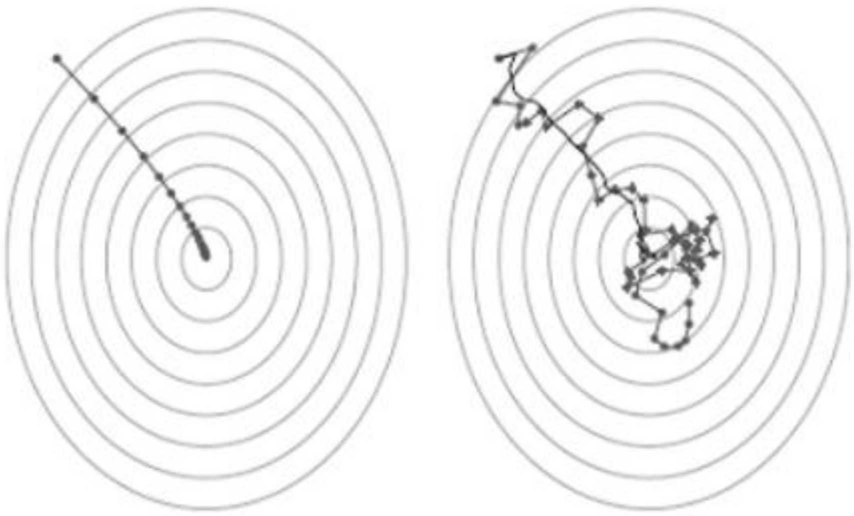


Image from Javatpoint

5.9 Stochastic Gradient Descent

- Stochastic Gradient Descent
  - Use the Mini-Batch not using all Batch
  - The one step of BGD takes a long time
  - SGD is faster than BGD normally, but it has uncertainty more than BGD



<BGD, SGD>

5.10 Building a Machine Learning Algorithm

- Building a Machine Learning Algorithm

1. Link the dataset and cost function
2. Decide the model and the type of optimization

In Linear Regression,

$$J(\boldsymbol{w}, b) = -\mathbb{E}_{\mathrm{x,y} \sim \hat{p}_{data}} \log p_{model}(y \mid \boldsymbol{x})$$

$$J(\boldsymbol{w}, b) = \lambda \|\boldsymbol{w}\|_2^2 - \mathbb{E}_{\mathrm{x,y} \sim \hat{p}_{data}} \log p_{model}(y \mid \boldsymbol{x})$$

$$J(\boldsymbol{w}) = \mathbb{E}_{\mathrm{x} \sim \hat{p}_{data}} \|\boldsymbol{x} - r(\boldsymbol{x}; \boldsymbol{w})\|_2^2$$

5.11 Challenges Motivating Deep Learning

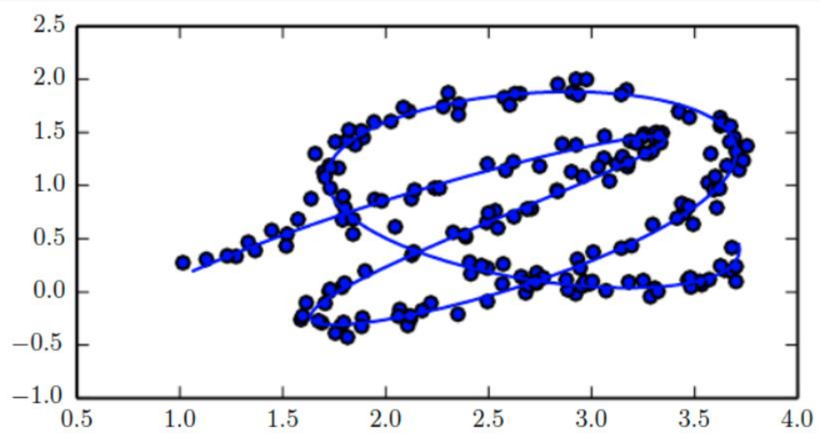
- Challenges Motivating Deep Learning
  - The Curse of Dimensionality

5.11 Challenges Motivating Deep Learning

- Challenges Motivating Deep Learning

$$f^*(\boldsymbol{x}) \approx f^*(\boldsymbol{x} + \epsilon)$$

  - Local Constancy and Smoothness Regularization
  - Manifold Learning



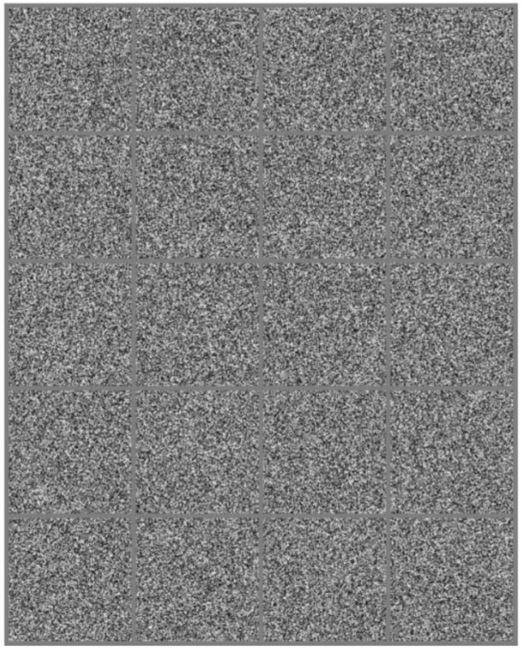


Figure 5.12: Sampling images uniformly at random (by randomly picking each pixel according to a uniform distribution) gives rise to noisy images. Although there is a nonzero probability of generating an image of a face or of any other object frequently encountered in AI applications, we never actually observe this happening in practice. This suggests that the images encountered in AI applications occupy a negligible proportion of the volume of image space.

The End.
Thank you for watching!

MILab Undergraduate student, TaeHyeon Kim

2023. 03. 07

+) Set rules on our study
(for example, contents, …)

+) How about to put PDFs together?