11 | Regularization

Ivan Corneillet

Data Scientist

GENERAL ASSEMBLY

Learning Objectives

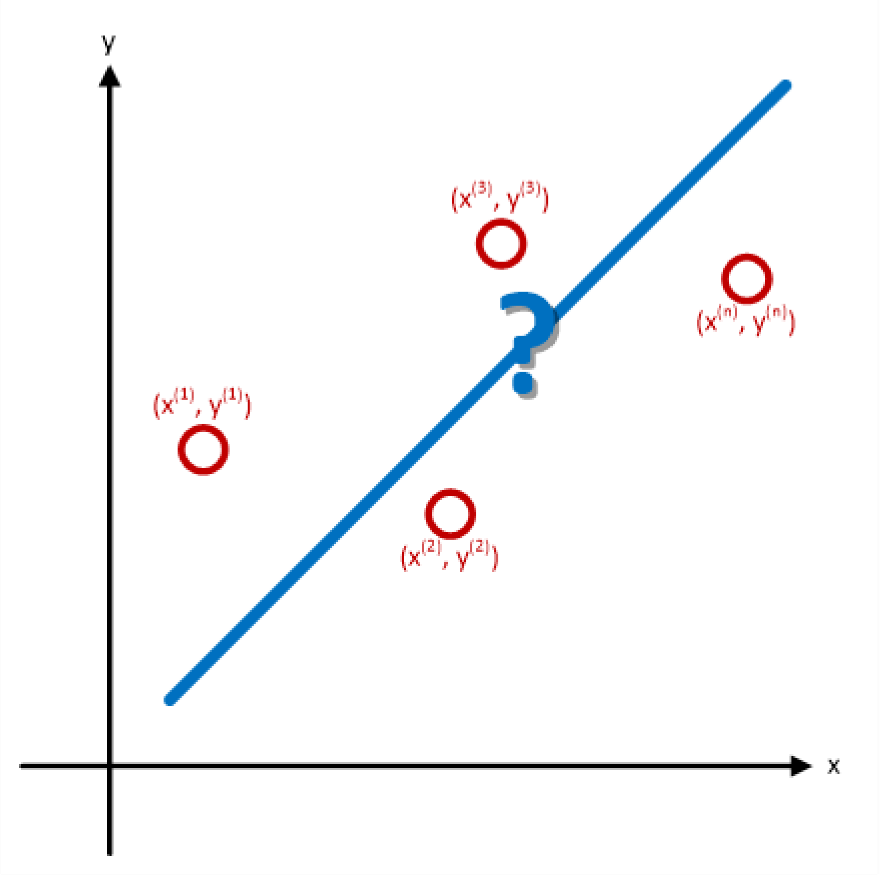
After this lesson, you should be able to:

- Understand the closed-form solution of the regression coefficients for linear regression models
- Use Ordinary Least Squares (OLS) and Loss Functions to also derive estimations for the coefficients
- Understand the Regularization Bias-Variance Trade-Off

DS

How to fit a linear regression model on a dataset?

How do we estimate $\hat{\beta}$?



How to fit a linear regression model on a dataset?

Closed-form solution for $\hat{\beta}$

Closed-form solution for $\hat{\beta}$

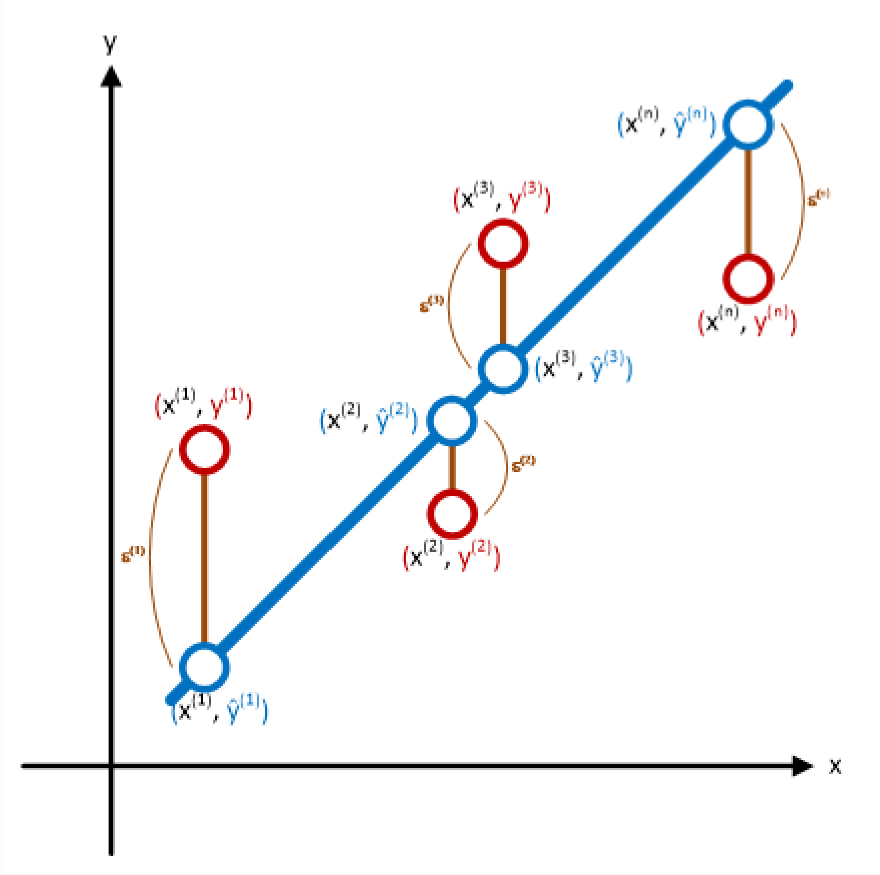
$$\hat{\beta} = \left(X_{train}^{T} \cdot X_{train}\right)^{-1} \cdot X_{train}^{T} \cdot y_{train}$$

$$\hat{y}_{predict} = X_{predict} \cdot \hat{\beta}$$

How to fit a linear regression model on a dataset?

Ordinary Least Squares (OLS) and Loss Functions

We can also estimate $\hat{\beta}$ with Ordinary Least Squares (OLS)



- Hypothesis

$$\hat{y}(x) = x \cdot \hat{\beta}$$

- Parameters

$$\hat{\beta}$$

- Goal

$$\underset{\hat{\beta}}{\operatorname{argmin}} \underbrace{\left\| y_{train} - \overbrace{X_{train} \cdot \hat{\beta}}^{\hat{y}_{train}} \right\|^2}_{L(\hat{\beta})}$$

(i.e., minimizing the least square errors)

How to fit a linear regression model on a dataset?

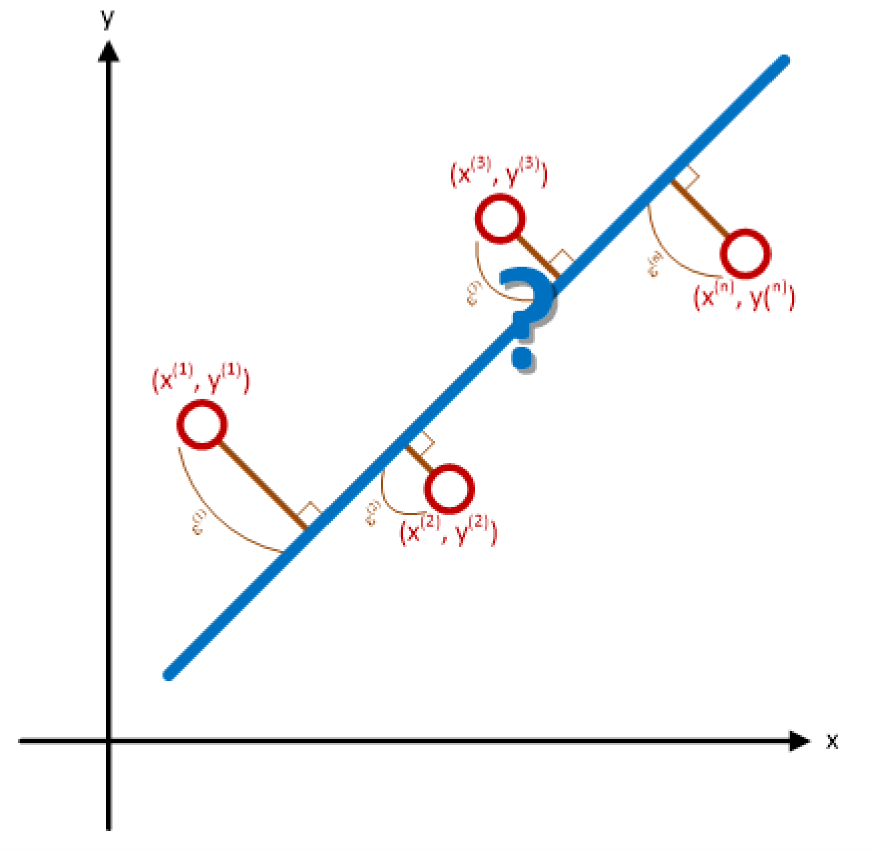
Ordinary Least Squares (OLS) and the closed-form solution for $\hat{\beta}$

Ordinary Least Squares (OLS) and the closed-form for $\hat{\beta}$

- Minimizing $L(\hat{\beta}) = \left\| y_{train} - X_{train} \cdot \hat{\beta} \right\|^2$ yields our previous closed-form solution for $\hat{\beta} = \left(X_{train}^T \cdot X_{train}\right)^{-1} \cdot X_{train}^T \cdot y_{train}$

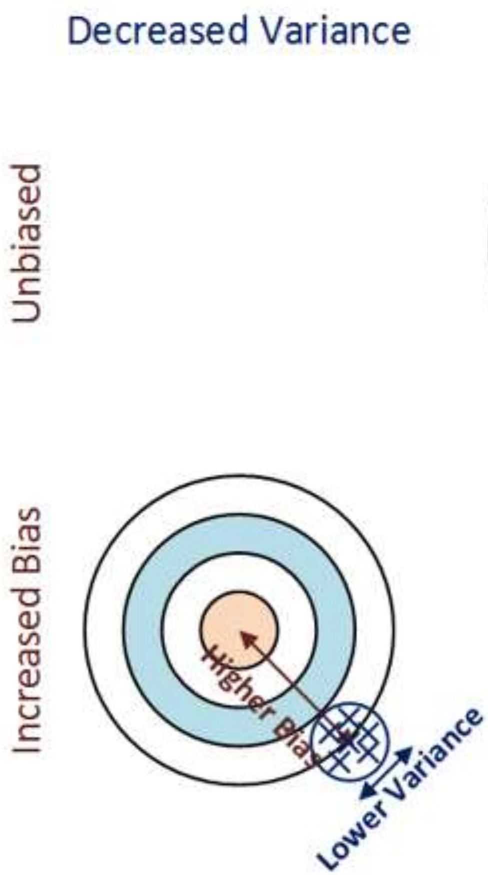
There are many ways to fit a line…

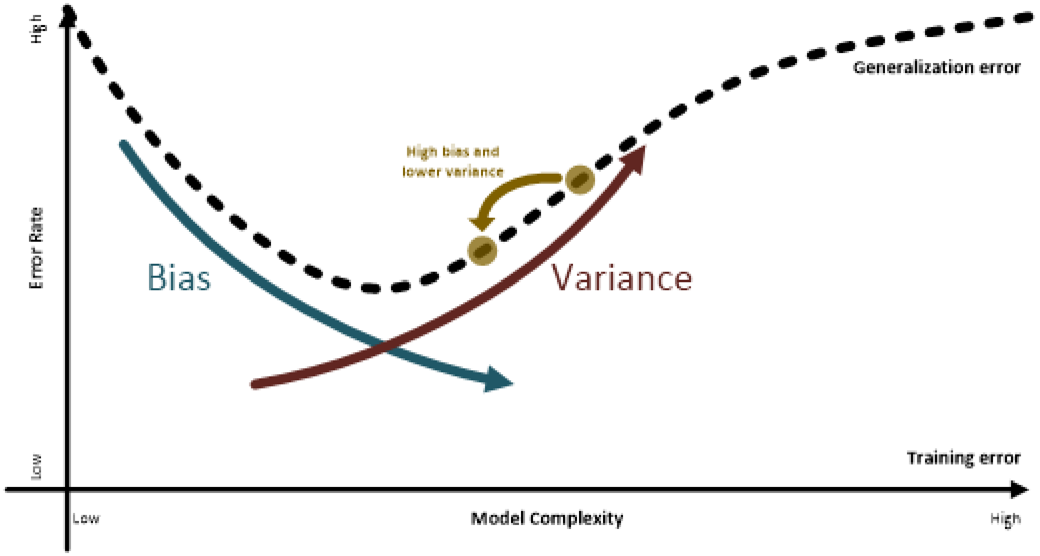
- It can be shown that $\hat{y}$ is unbiased for y
  - I.e., $E[\hat{y}] = y$
- E.g., if $\hat{\beta} = (\hat{\beta}_0)$, then $\hat{y} = \bar{y}$ is unbiased
  - In the SF housing dataset, if you don't anything about the houses (e.g., size, etc.) then the best estimation of the sale price you can give is the mean of the training set's sale price



Regularization

OLS yields Unbiased Estimators at the cost of High Variance. Can we trade some (Higher) Bias for Lower Variance and get ahead on the Bias-Variance Trade-off?





Revisiting Complexity

- E.g., as a function of the size of the coefficients
  - $\|\beta\|_p = \left(\sum_{j=0}^{k} \left|\beta_j\right|^p\right)^{1/p}$ (Lp-norm)
  - $\|\beta\|_1 = \sum_{j=0}^{k} \left|\beta_j\right|$ (L1-norm)
  - $\|\beta\|_2 = \left(\sum_{j=0}^{k} \left|\beta_j\right|^2\right)^{1/2}$(L2-norm)

Regularization helps against overfitting by explicitly controlling model complexity

- These definitions of complexity lead to the following regularization techniques
  - $\underset{\hat{\beta}}{\operatorname{argmin}}\left(\underbrace{\left\|y_{train} - X_{train} \cdot \hat{\beta}\right\|^2}_{OLS\ term} + \underbrace{\lambda\left\|\hat{\beta}\right\|_1}_{regularization\ term}\right)$ (Lasso regularization using the L1 norm)
  - $\underset{\hat{\beta}}{\operatorname{argmin}}\left(\left\|y_{train} - X_{train} \cdot \hat{\beta}\right\|^2 + \lambda\left\|\hat{\beta}\right\|_2^2\right)$ (Ridge regularization using the L2 norm)
    - (note that in the loss function the term $\hat{\beta}_0$ isn't regularized and is in fact excluded from the norm here)
- This formulation reflects the fact that there is a cost associated with regularization that we want to minimize

About Loss Functions

- Loss functions are a powerful tool to optimize the fit of machine learning algorithms
- Loss functions are not limited to linear regression- and regularization-based models.
  - E.g., training a logistic regression algorithm (while also leveraging linear regression) is also modeled and fitted with loss functions