Bil 470 / YAP 470

Introduction to Machine Learning (Yapay Öğrenme)

Batuhan Bardak

Lecture 3: Regression, Overfitting & Underfitting, Regularization

Date: 20.09.2022 & 26.09.2022

Plan for today

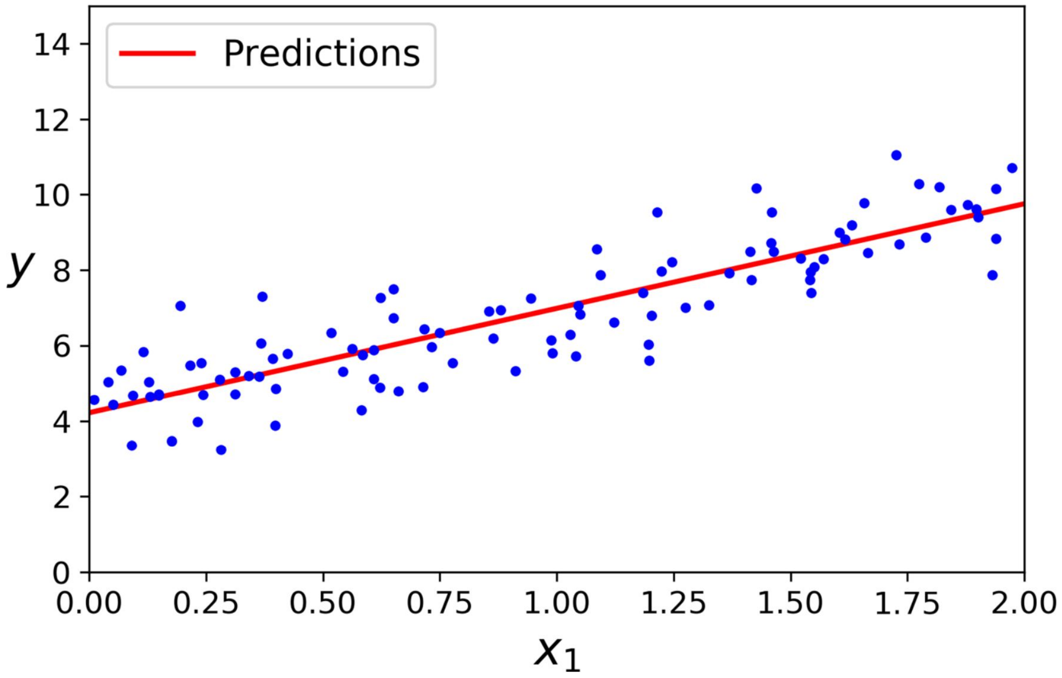
- Linear Regression
- Gradient Descent
- Overfitting & Underfitting
- Regularization

Key Terms for Simple Linear Regression

- **Response**
  - The variable we are trying to predict.
    - Dependent variable, Y variable, target, outcome
- **Independent variable**
  - The variable used to predict the response.
    - X variable, feature, attribute, predictor
- **Record**
  - The vector of predictor and outcome values for a specific individual
    - Row, case, instance, example
- **Intercept**
  - The intercept of the regression line
    - b0 , β0
- **Regression coefficient**
  - The slope of the regression line.
    - Slope, b1 , β1, parameter estimates, weights
- **Fitted values**
  - The estimates Yi obtained from the regression line.
    - Predicted values
- **Residuals**
  - **The difference between the observed values and the fitted values.**
    - Errors
- **Least Squares**
  - The method of fitting a regression by minimizing the sum of squared residuals.

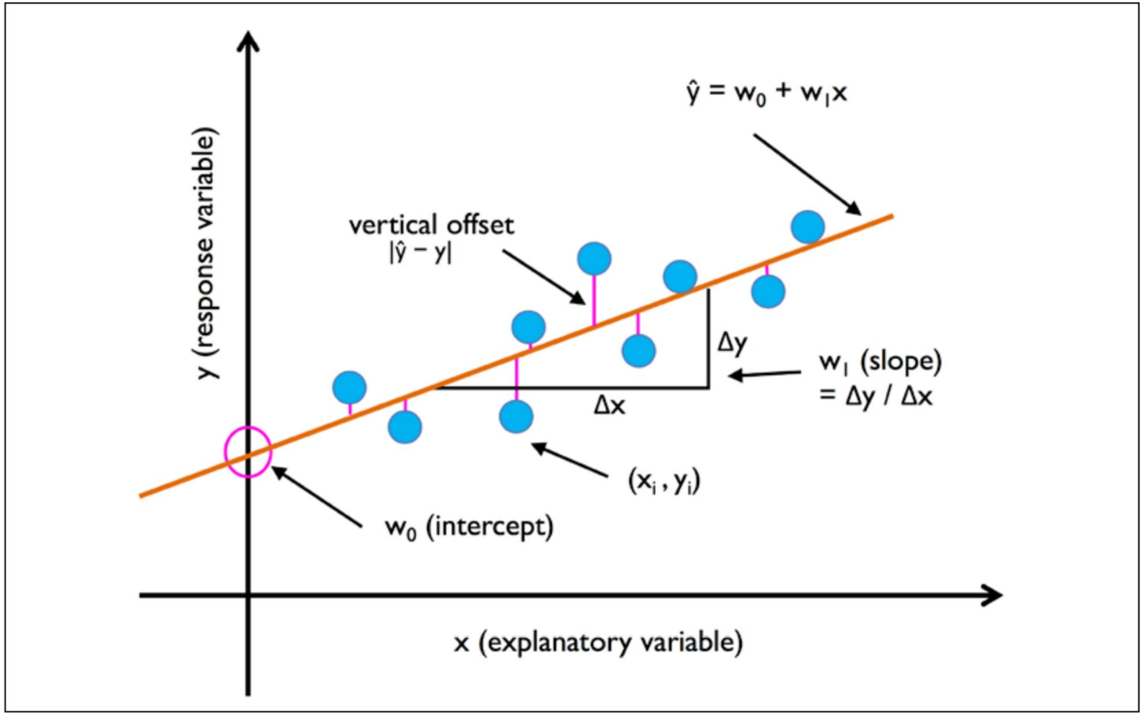
Slide Credit: Peter Bruce, Practical Statistics for Data Scientists

Linear Regression



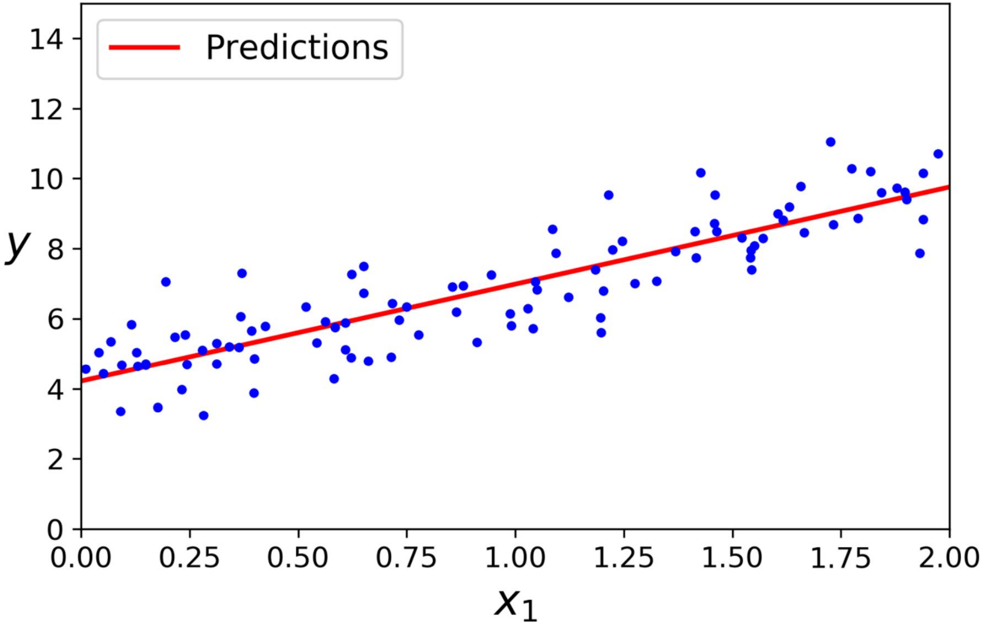
Slide Credit: Amuller, COMS4995

Linear Regression



Slide Credit: Amuller, COMS4995

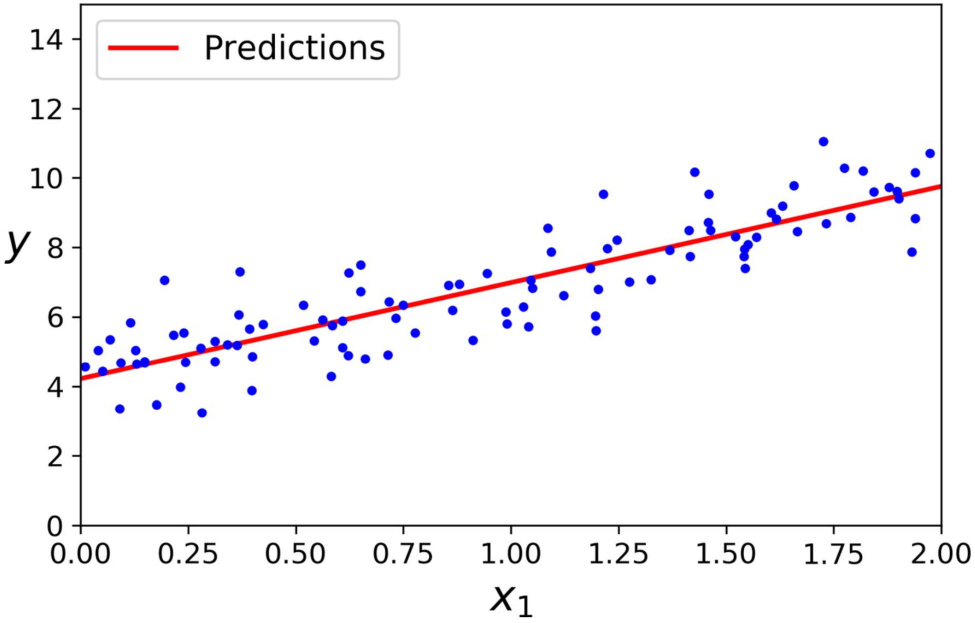
Linear Regression



$$\hat{y} = w^T \mathbf{x} + b = \sum_{i=1}^{p} w_i x_i + b$$

Slide Credit: Amuller, COMS4995

Linear Regression



$$\hat{y} = w^T\mathbf{x} + b = \sum_{i=1}^{p} w_i x_i + b$$

$$\min_{w\in\mathbb{R}^p, b\in\mathbb{R}} \sum_{i=1}^{n} (w^T\mathbf{x}_i + b - y_i)^2$$

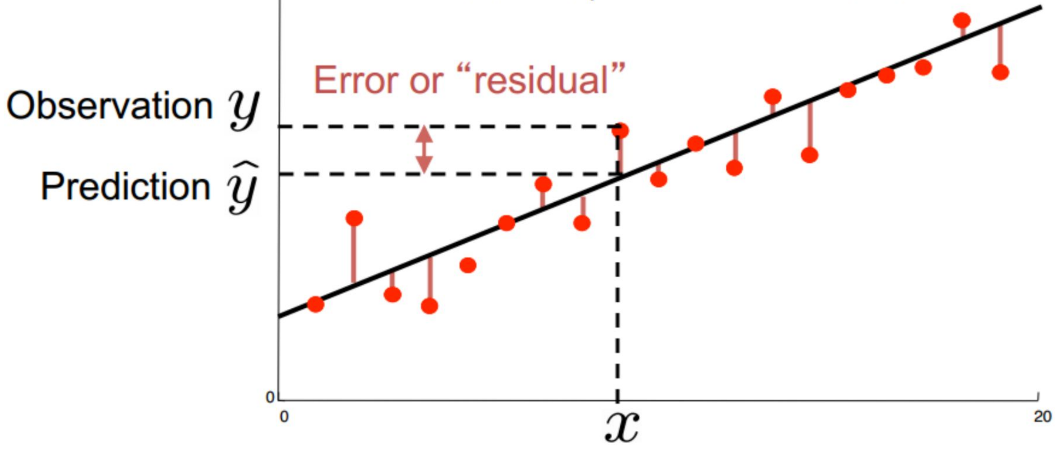
Slide Credit: Amuller, COMS4995

Ordinary Least Squares (OLS)

$$\text{total error} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i \left(y_i - \sum_k w_k x_k^{(i)} \right)^2$$

sum over data points

features



Slide Credit: Bilkent, CS464

Ordinary Least Squares (OLS)

- Regression Line is the estimate that minimizes the sum of squared residual values, also called the *residual sum of squares* or *RSS*:

$$RSS = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$
$$= \sum_{i=1}^{n} \left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right)^2$$

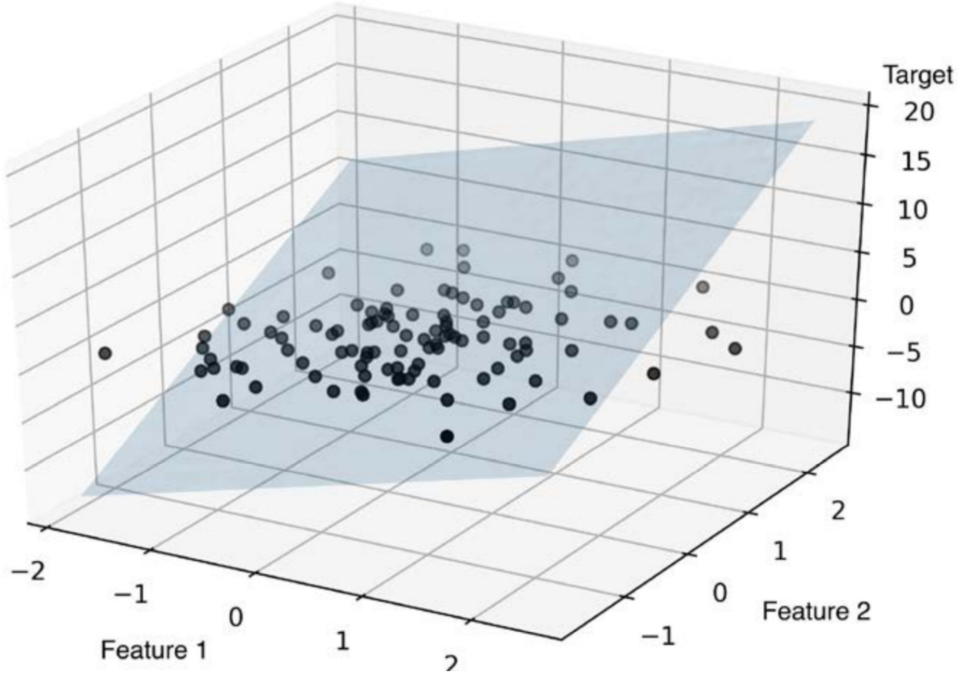
- The method of minimizing the sum of the squared residuals is termed *least squares regression,* or *ordinary least squares (OLS)* regression.

Slide Credit: Peter Bruce, Practical Statistics for Data Scientist

Ordinary Least Squares (OLS)

https://phet.colorado.edu/sims/html/curve-fitting/latest/curve-fitting_en.html

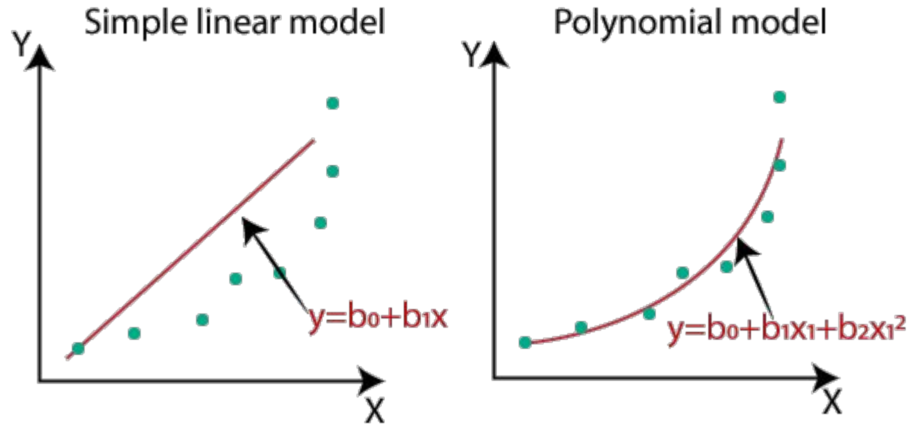
Multiple Linear Regression

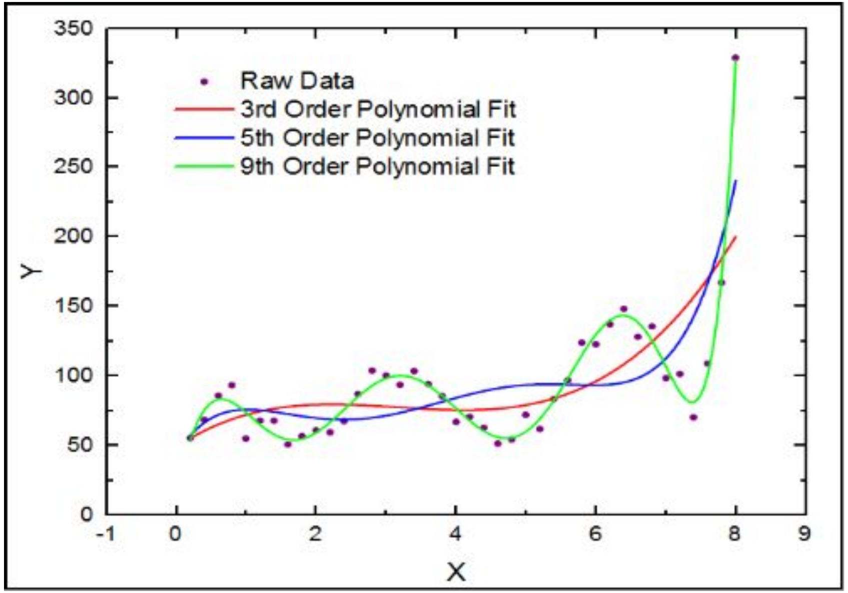


$$y = w_0 x_0 + w_1 x_1 + \ldots + w_m x_m = \sum_{i=0}^{n} w_i x_i = w^T x$$

Slide Credit: Rascka, Python Machine Learning

Polynomial Regression





Evaluation Metrics

- Pearson correlation
- R-squared, or R^2
- Adjusted- R-squared
- Mean Squared Error
- Mean Absolute Error
- Root Mean Squared Error

Pearson Correlation

- *Pearson's r*
- Is a measure of linear correlation between two sets of data.
- It is the ratio between the *covariance* of two variables and the product of their *standard deviations.*
- The results always has a value between -1 and 1.

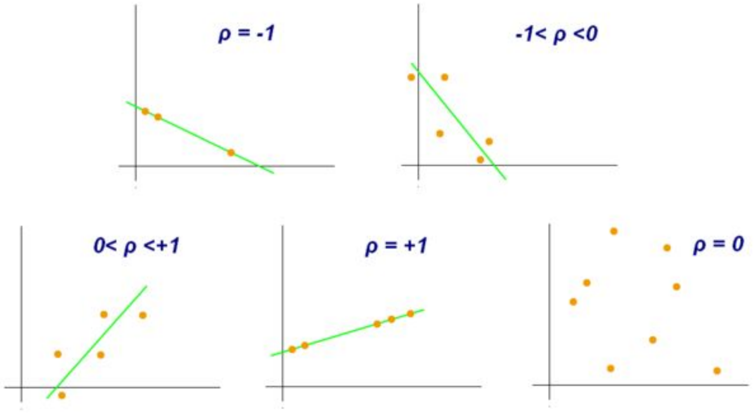
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples $\quad$ y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\quad$ $\bar{y}$ =mean of values in y variable



Slide Credit: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Coefficient of Determination

- R-squared, or R^2
- R-squared ranges from 0 to 1 and measures the proportion of variation in the data that is accounted for in the model.
- It is useful mainly in explanatory uses of regression where you want to assess how well the model fits the data.
- **Note**: It does not take into consideration of overfitting problem. If your model has many independent variables, due to model is too complicated, it may fit very well to the training data but performs badly for testing data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

Slide Credit: Peter Bruce, Practical Statistics for Data Scientist

Adjusted R^2

- The disadvantage of the R_square score is while adding new features in data the R_square score starts increasing or remains constant but never decreases because it assumes that while adding more data variance of data increases.

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Where

R^2 Sample R-Squared

N Total Sample Size

p Number of independent variable

- Where where n is the total number of observations and p is the number of predictors. Adjusted R^2 will always be less than or equal to R^2.

Slide Credit: Peter Bruce, Practical Statistics for Data Scientist

Adjusted R^2

Case 1		Case 2			Case 3		
Var1	Y	Var1	Var2	Y	Var1	Var2	Y
x1	y1	x1	2*x1	y1	x1	2*x1+0.1	y1
x2	y2	x2	2*x2	y2	x2	2*x2	y2
x3	y3	x3	2*x3	y3	x3	2*x3 + 0.1	y3
x4	y4	x4	2*x4	y4	x4	2*x4	y4
x5	y5	x5	2*x5	y5	x5	2*x5 + 0.1	y5

	Case 1	Case 2	Case 3
R_squared	0.985	0.985	0.987
Adj_R_squared	0.981	0.971	0.975

Slide Credit: https://medium.com/usf-msds/choosing-the-right-metric-for-machine-learning-models-part-1-a99d7d7414e4

Mean Square Error (MSE)

- MSE is calculated by the sum of square of prediction error which is real output minus predicted output and the divide by the number of data points.
- It is hard to interpret many insights from one single result but it gives you a real number to compare against other model results and help you select the best regression model.
- If you have outliers in the dataset then it penalizes the outliers most and the calculated MSE is bigger (-).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Slide Credit: Songhao Wu

Mean Absolute Error (MAE)

- Same unit as the output variable (+)
- Robust to outliers (+)

Divide by the total number of data points

Predicted output value

Actual output value

$$MAE = \boxed{\frac{1}{n}} \underset{\text{Sum of}}{\Sigma} \underbrace{\left| \boxed{y} - \boxed{\hat{y}} \right|}_{\text{The absolute value of the residual}}$$

Root Mean Square Error (MSE)

-

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Choosing the right metric

- A nice blog post that you can read more about metrics for evaluation regression models.
  - https://medium.com/usf-msds/choosing-the-right-metric-for-machine-learning-models-part-1-a99d7d7414e4

Case 1: Actual Values = [2,4,6,8] , Predicted Values = [4,6,8,10]

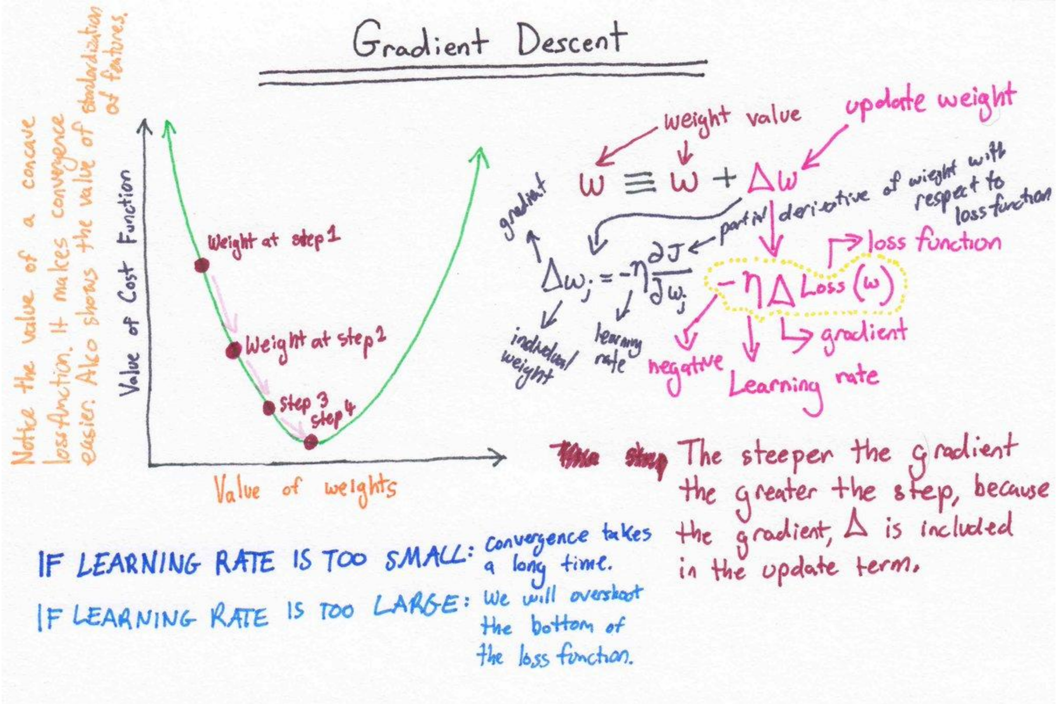
Case 2: Actual Values = [2,4,6,8] , Predicted Values = [4,6,8,12]

MAE for case 1 = 2.0, RMSE for case 1 = 2.0

MAE for case 2 = 2.5, RMSE for case 2 = 2.65

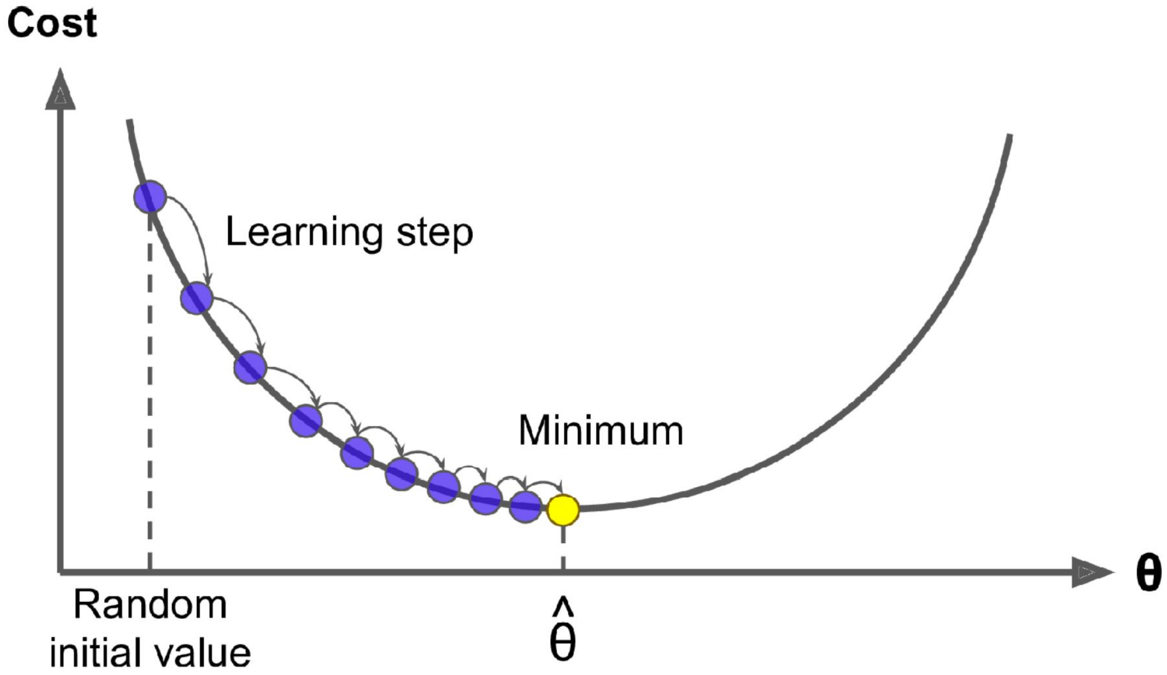
Slide Credit: Songhao Wu

Gradient Descent



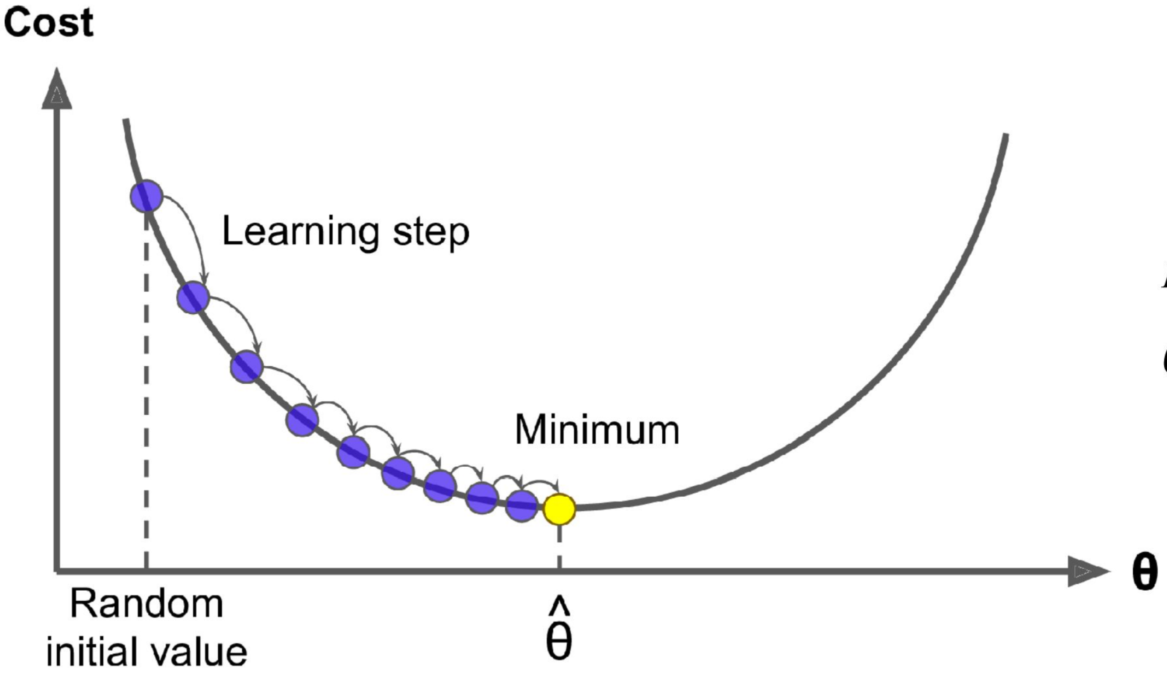
Slide Credit: Machine Learning Flashcards

Gradient Descent



Slide Credit: Aurelien, Hands-on Machine Learning

Gradient Descent



Equation 4-7. Gradient Descent step

$$\boldsymbol{\theta}^{(\text{next step})} = \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \, \text{MSE}(\boldsymbol{\theta})$$

Slide Credit: Aurelien, Hands-on Machine Learning

Gradient Descent

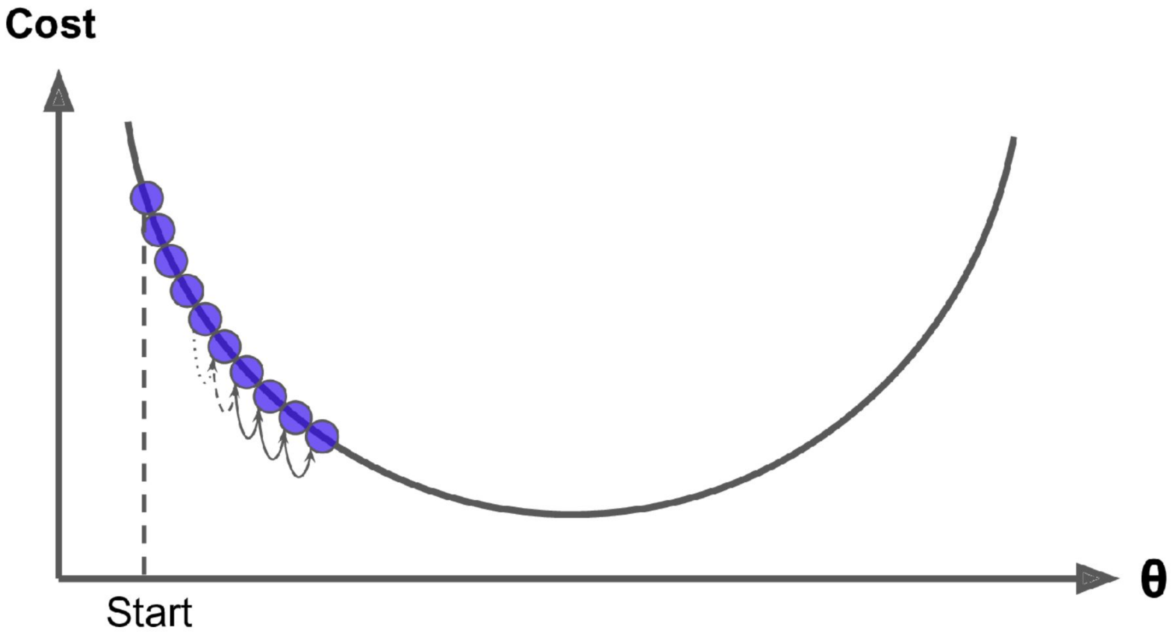


Figure 4-4. The learning rate is too small

Slide Credit: Aurelien, Hands-on Machine Learning

Gradient Descent

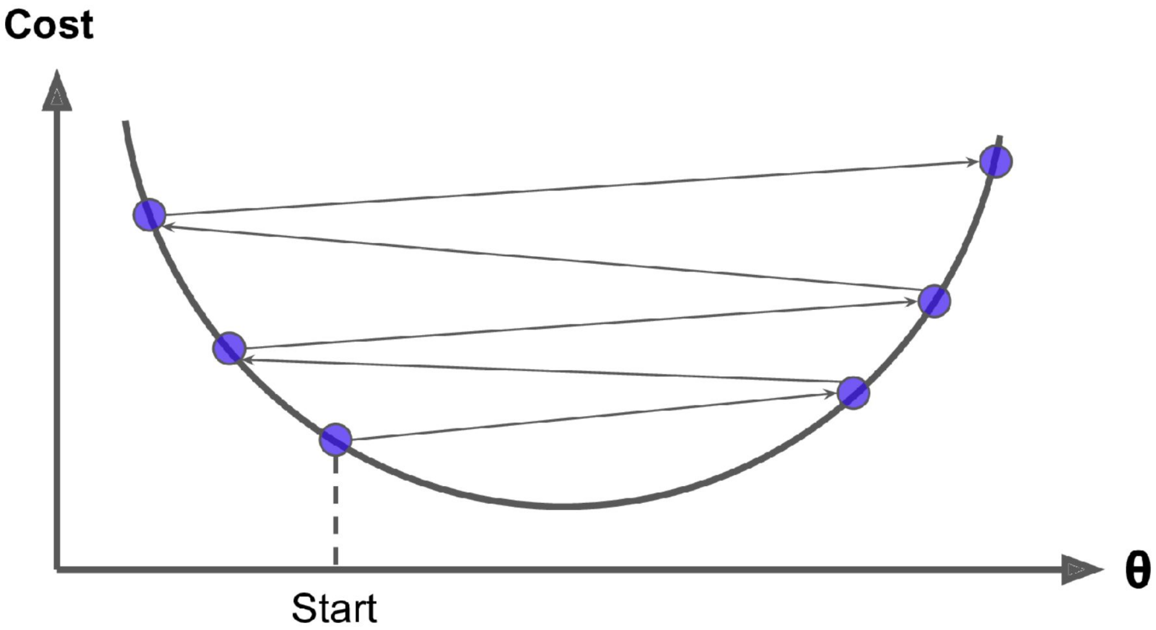


Figure 4-5. The learning rate is too large

Slide Credit: Aurelien, Hands-on Machine Learning

Gradient Descent

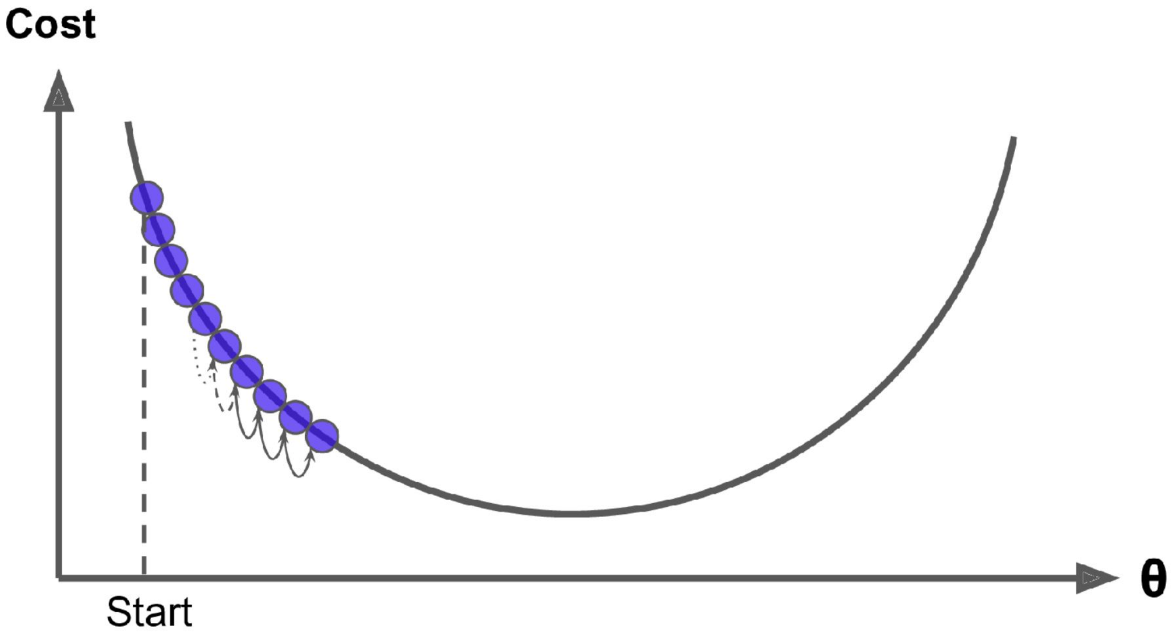


Figure 4-4. The learning rate is too small

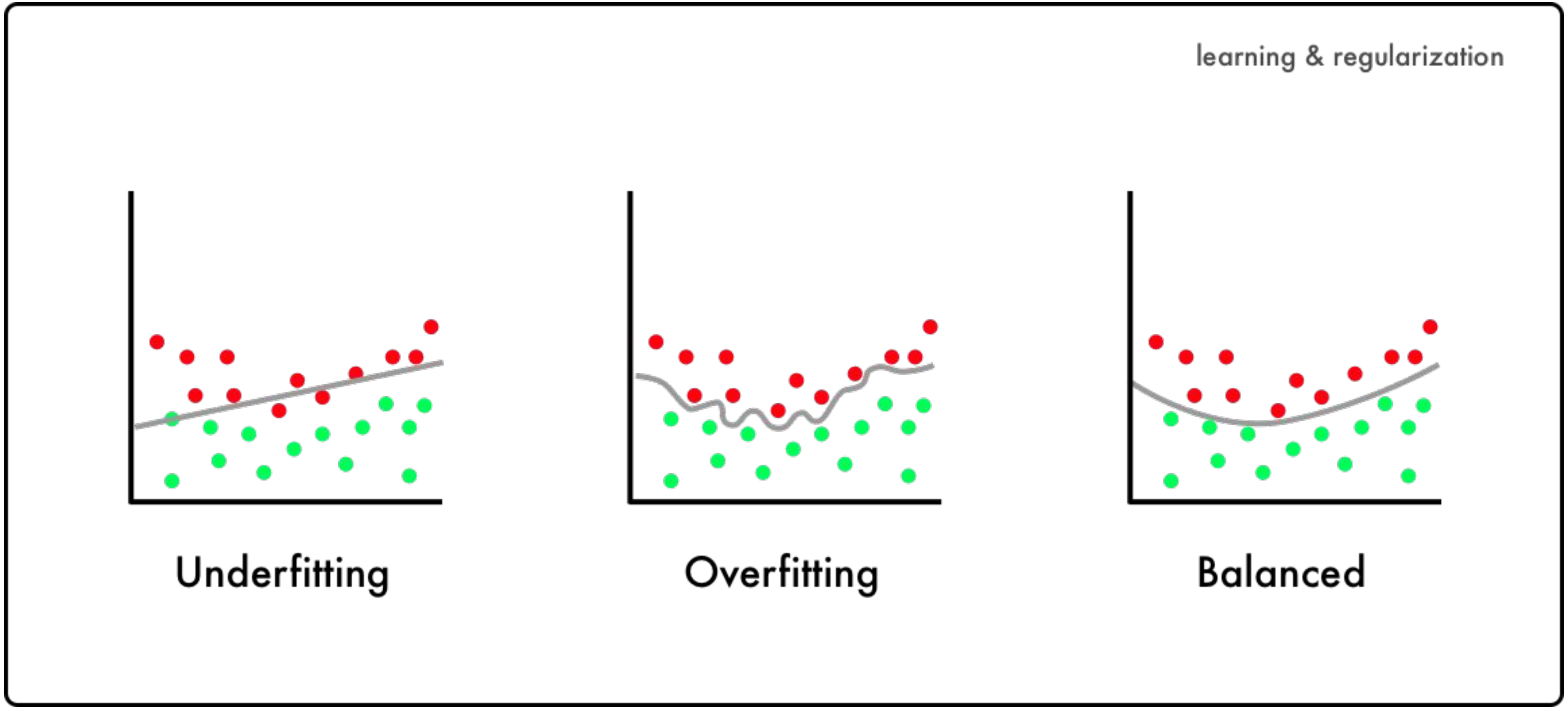
Slide Credit: Aurelien, Hands-on Machine Learning

Gradient Descent

- Check this blog post for interactive explanation of Linear Regression
  - https://machinelearningcompass.com/machine_learning_math/gradient_descent_for_linear_regression/

Slide Credit: Aurelien, Hands-on Machine Learning

Overfitting & Underfitting

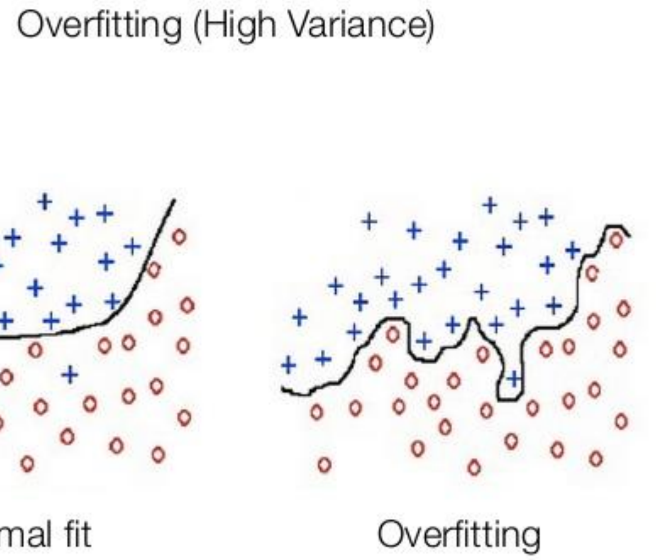


Overfitting

- **Overfitting:** Good performance on the training data, poor generalization to other (test) data.
- Overfitting refers to a model that models the training data to well.
- Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data.
- This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model.

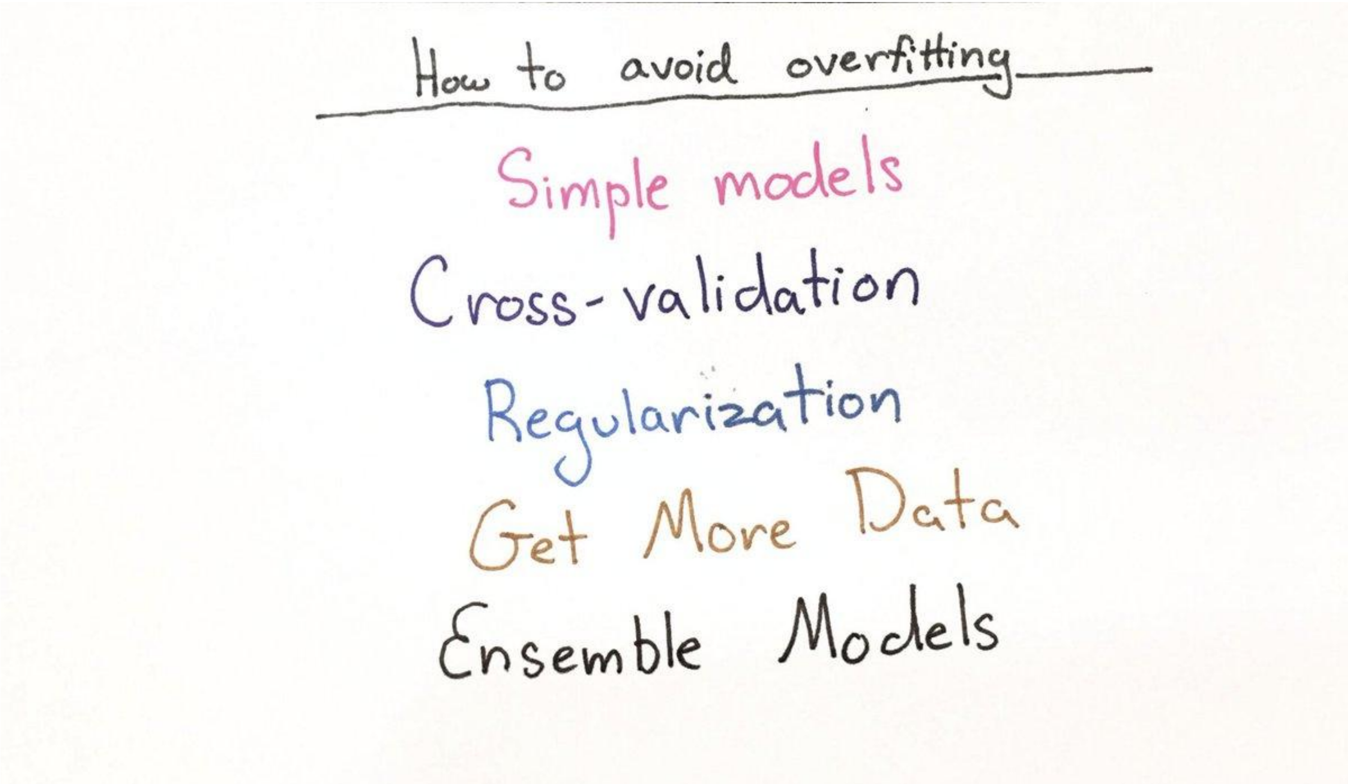
Slide Credit: https://machinelearning101.readthedocs.io/

Overfitting



http://blog.csdn.net/love_tea_cat/article/details/25972921

How to avoid overfitting



Slide Credit: Machine Learning Flashcards

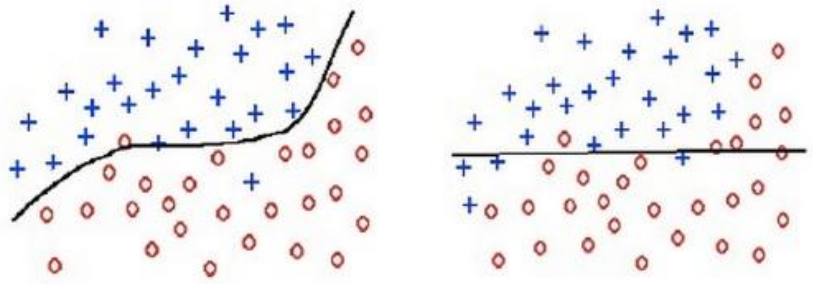
Underfitting

- **Underfitting:** Poor performance on the training data and poor generalization to other (test, val) data.
- Underfitting refers to a modal that can neither model the training data nor generalize to new data.
- Underfitting is often not discussed as it is easy to detect given a good performance metric.

Slide Credit: https://machinelearning101.readthedocs.io/

Underfitting

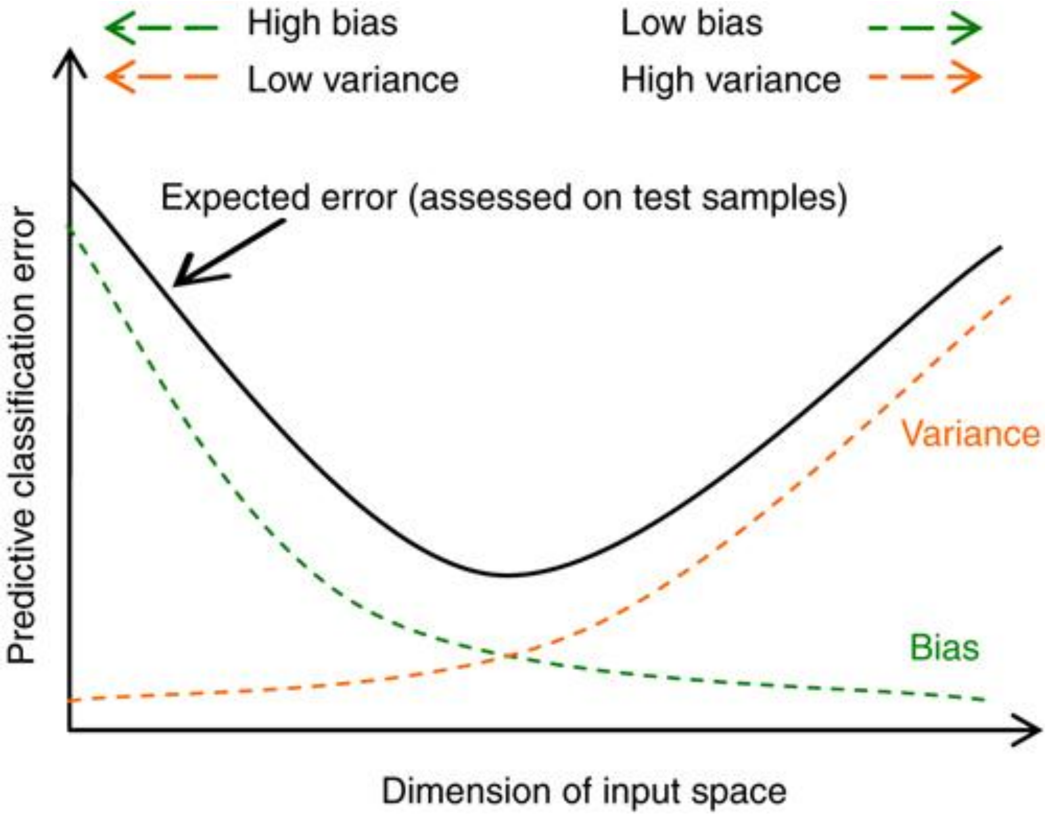
Underfitting (High Bias)



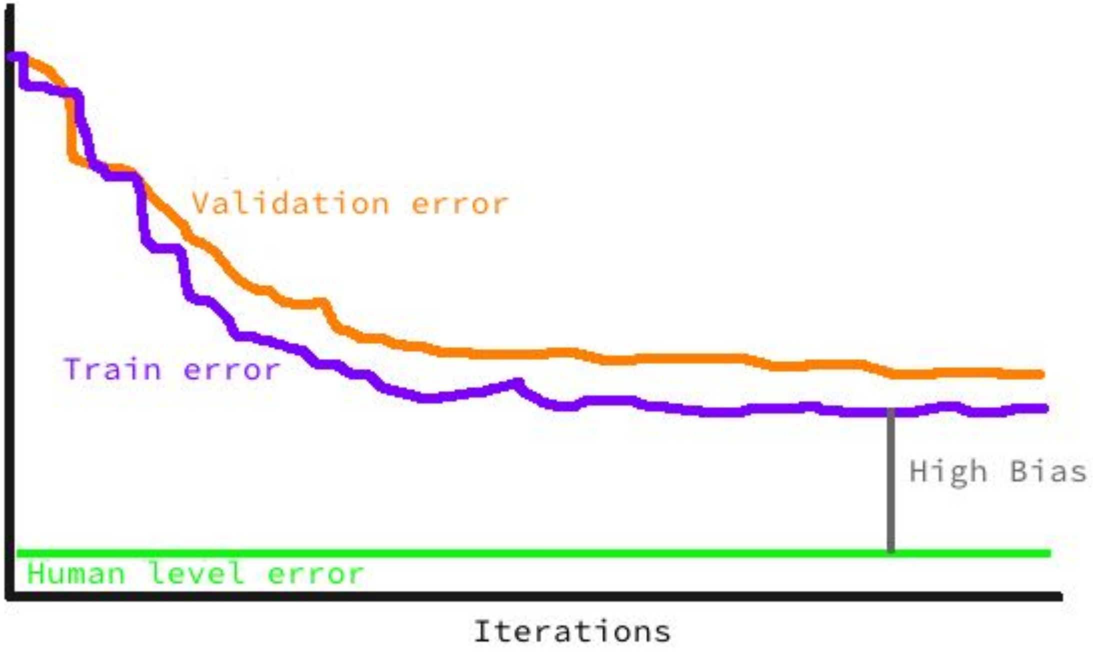
Normal fit Underfitting

http://blog.csdn.net/love_tea_cat/article/details/25972921

Understanding the Bias Variance Tradeoff

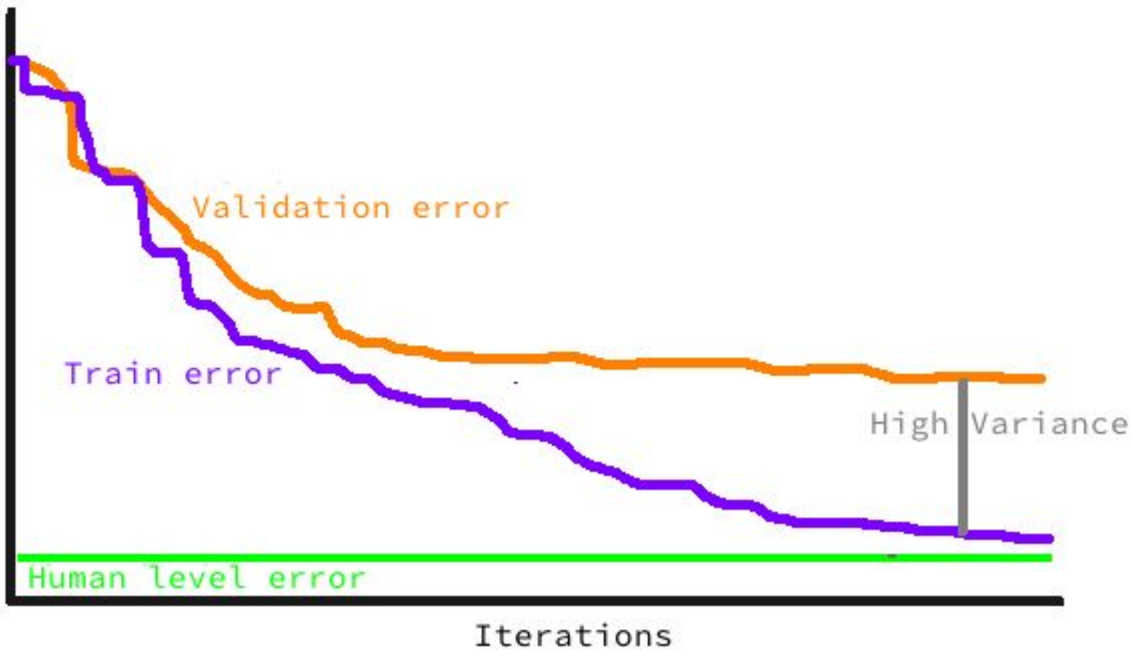


High Bias



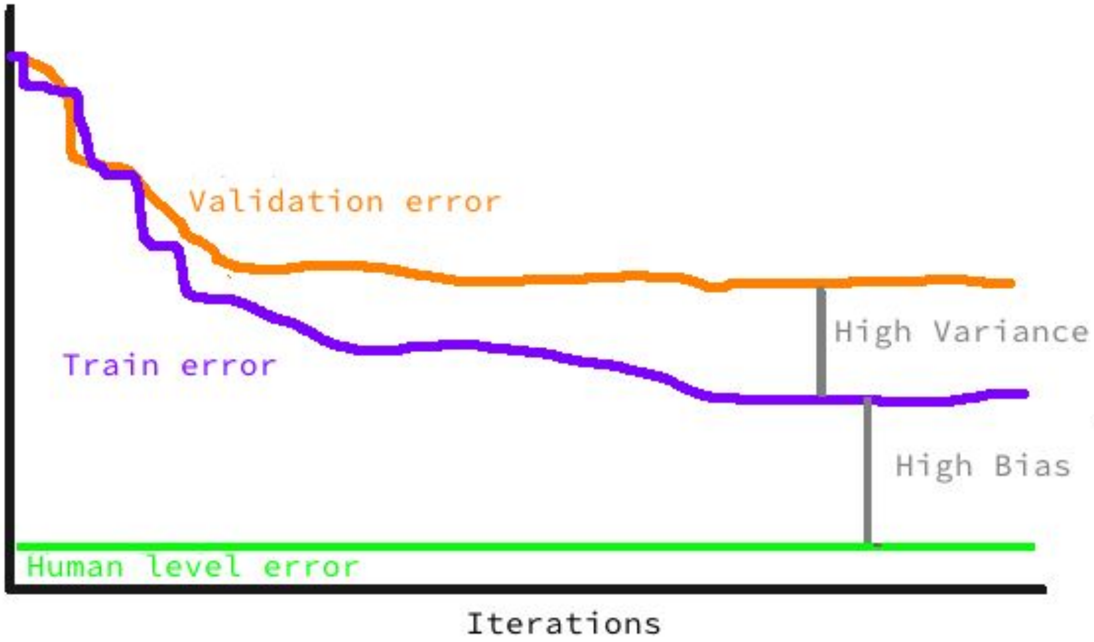
- Bigger model
- Train longer
- New architecture
- Example model:
    - Linear Regression

High Variance

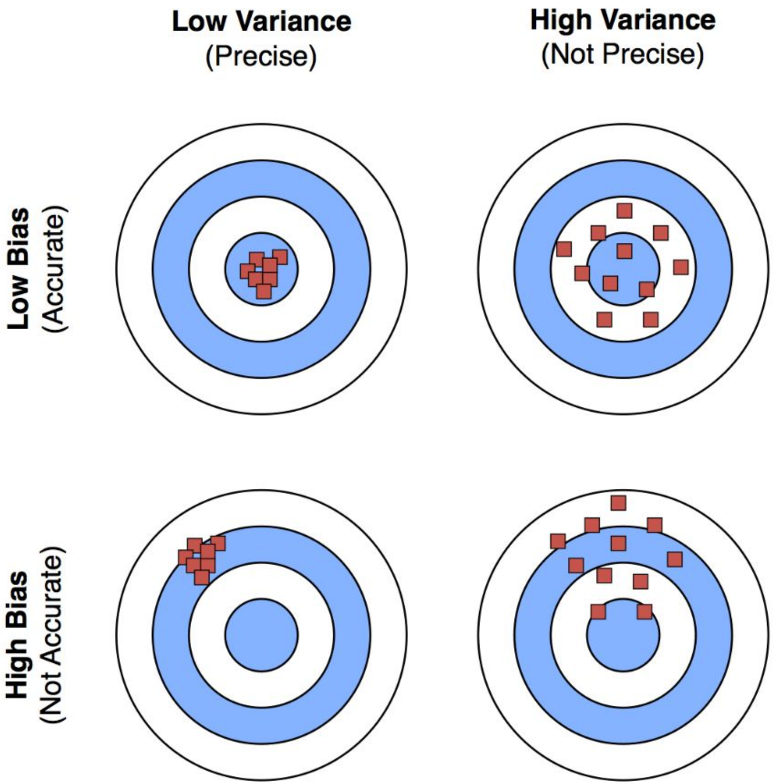


- Regularization
- Bagging Methods
- Dimensionality Reduction
- Feature Selection
  - Example model:
    - KNN, Decision Tree

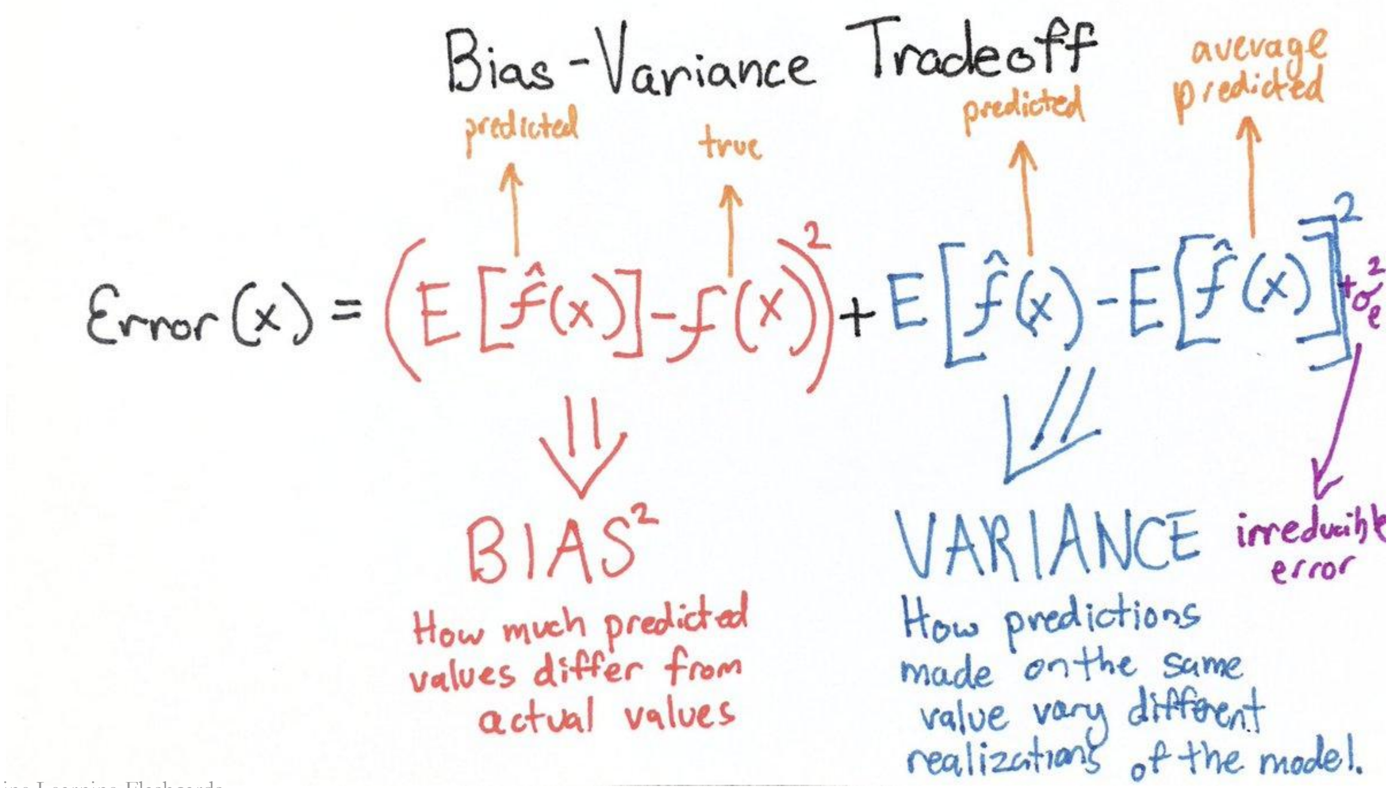
High Variance & High Bias



Graphical Definition



Mathematical Definition



Slide Credit: Machine Learning Flashcards

Regularization

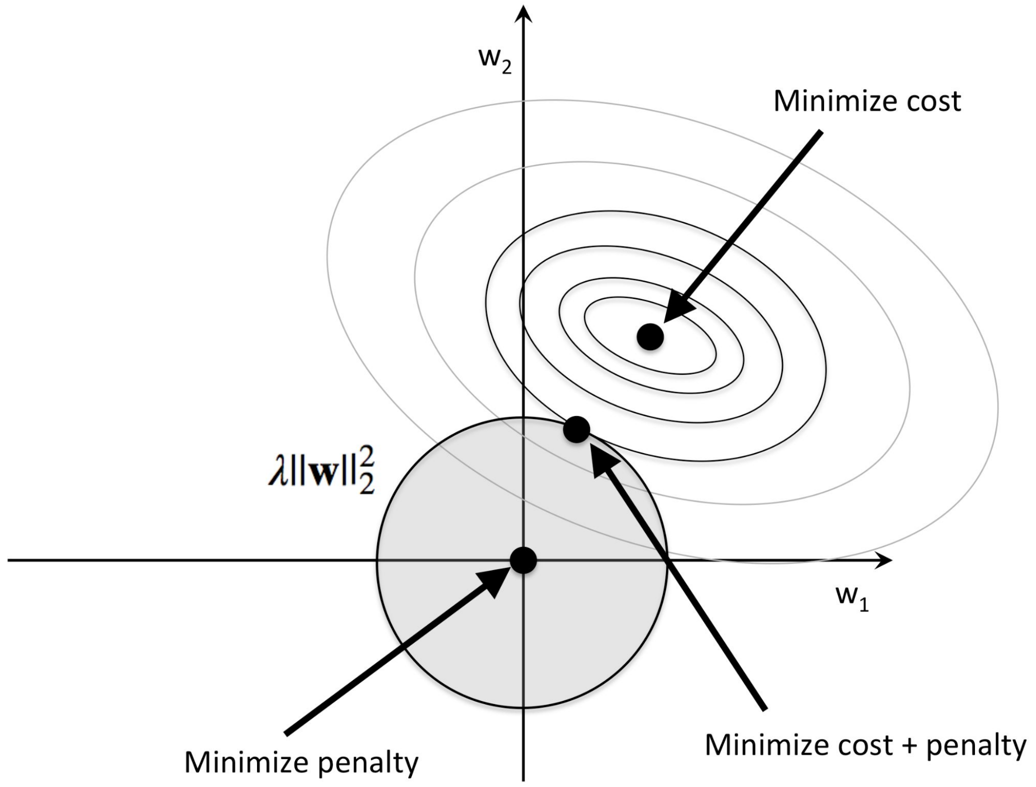
- Regularization is a technique used in an attempt to solve the overfitting problem in statistical mode and can be motivated as a technique to improve the generalizability of a learned model.
- There can be different types such as:
  - Ridge Regularization
  - Lasso Regularization
  - Elastic Net

Ridge Regression

- Always has a unique solution.
- Tuning parameter alpha.

$$\min_{w \in \mathbb{R}^p, b \in \mathbb{R}} \sum_{i=1}^{n} (w^T \mathbf{x}_i + b - y_i)^2 + \alpha ||w||^2$$

Geometric Interpretation of Ridge Regression

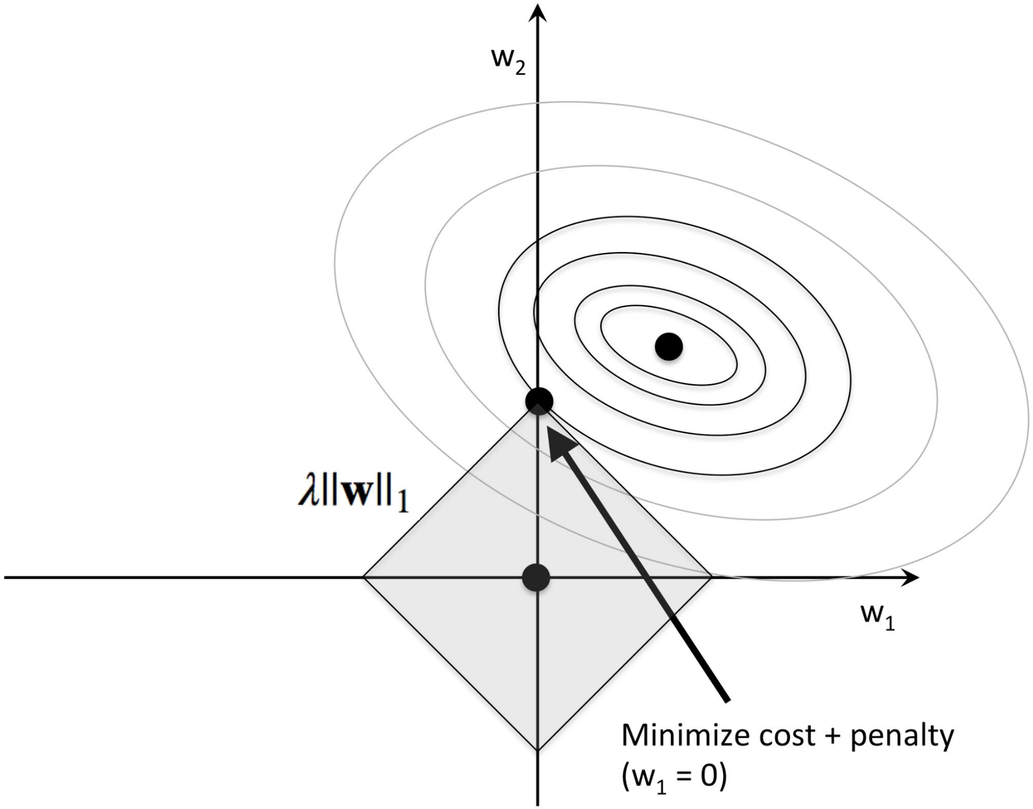


Lasso Regression

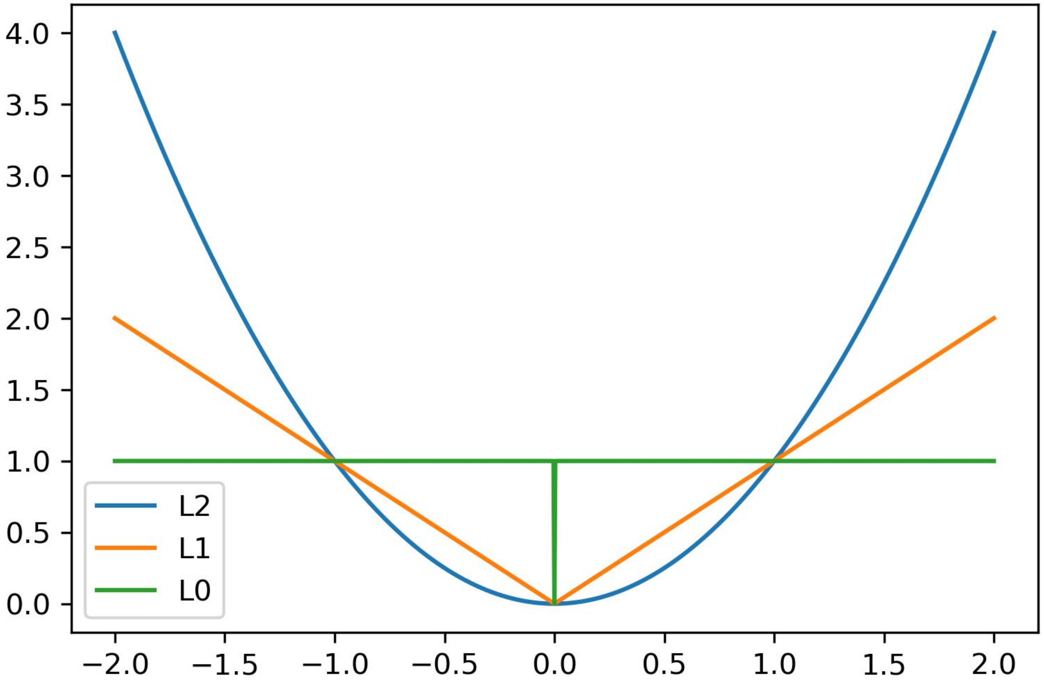
- Shrinks w towards zero like Ridge
- Sets some w exactly to zero - automatic feature selection!

$$\min_{w \in \mathbb{R}^p, b \in \mathbb{R}} \sum_{i=1}^{n} (w^T \mathbf{x}_i + b - y_i)^2 + \alpha ||w||_1$$

Geometric Interpretation of Lasso Regression



Understanding L1 and L2 Penalties



$$\ell_2(w) = \sqrt{\sum_i w_i^2}$$

$$\ell_1(w) = \sum_i |w_i|$$

$$\ell_0(w) = \sum_i 1_{w_i != 0}$$

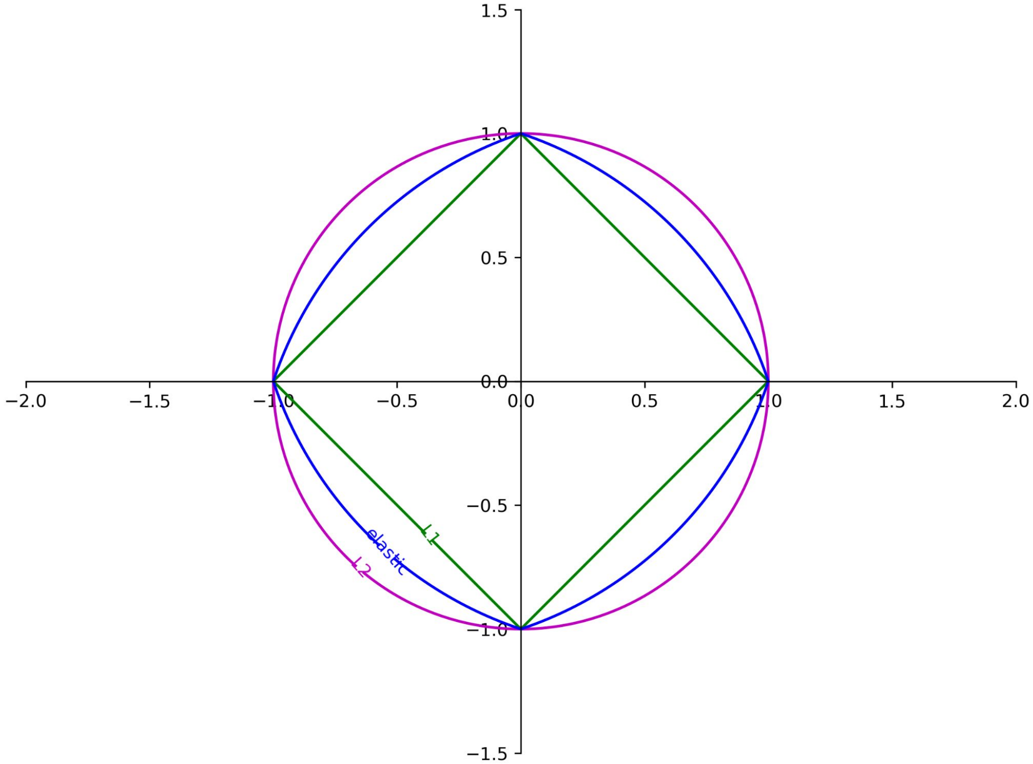
Slide Credit: Amuller, COMS4995

Elastic Net

- Combines benefits of Ridge and Lasso
- 2 parameters to tune

$$\min_{w\in\mathbb{R}^p, b\in\mathbb{R}} \sum_{i=1}^{n} ||w^T\mathbf{x}_i + b - y_i||^2 + \alpha_1 ||w||_1 + \alpha_2 ||w||_2^2$$

Elastic Net



Slide Credit: Amuller, COMS4995

Recap

- Basics of ML
- KNN
- Train / Test / Validation
- Evaluation Metrics
- Regression
- Regularization
- Overfitting & Underfitting

Next Class:

Feature Selection & Details of Evaluation metrics and Cross Validation