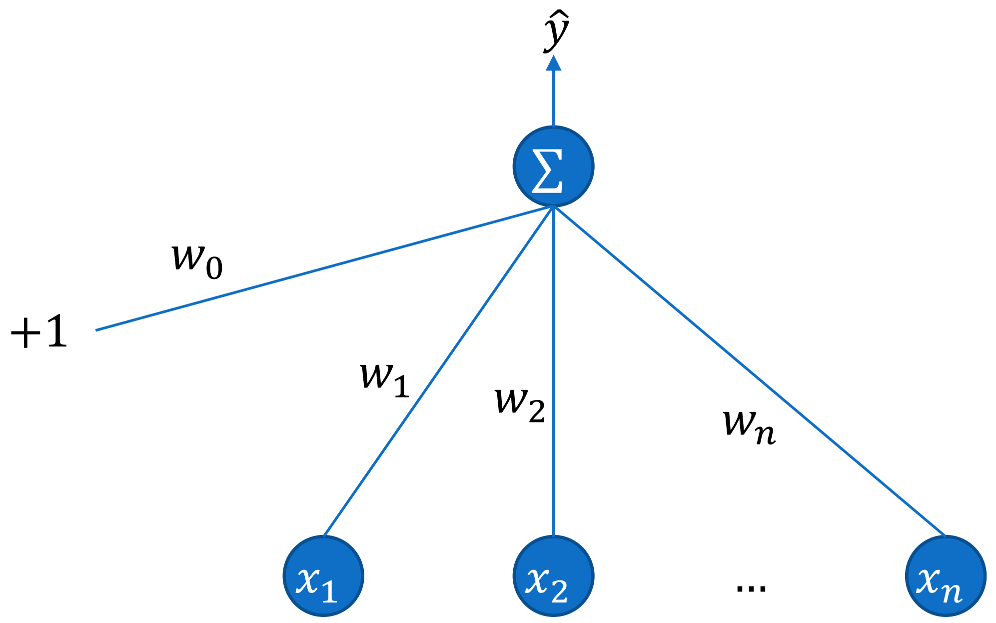
Review

LINEAR REGRESSION

Linear Regression: A Visual Perspective

$$\boldsymbol{h(X) = W^T X = w_0 x_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}$$

Compute Error: $y - \hat{y}$

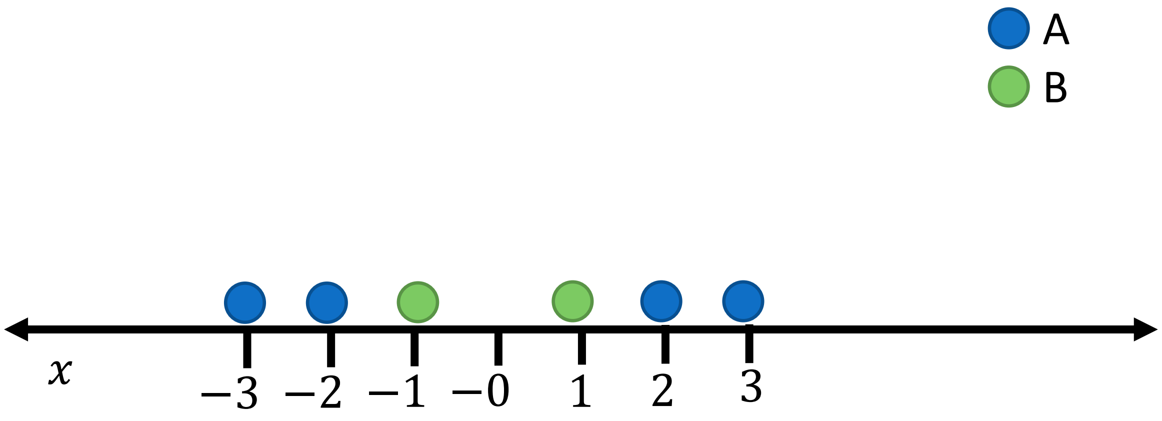


Polynomial Regression

Polynomials: Intuition

- Suppose the following dataset

x	y
-3	A
-2	A
-1	B
1	B
2	A
3	A



We will consider classification label as it is easier to understand…

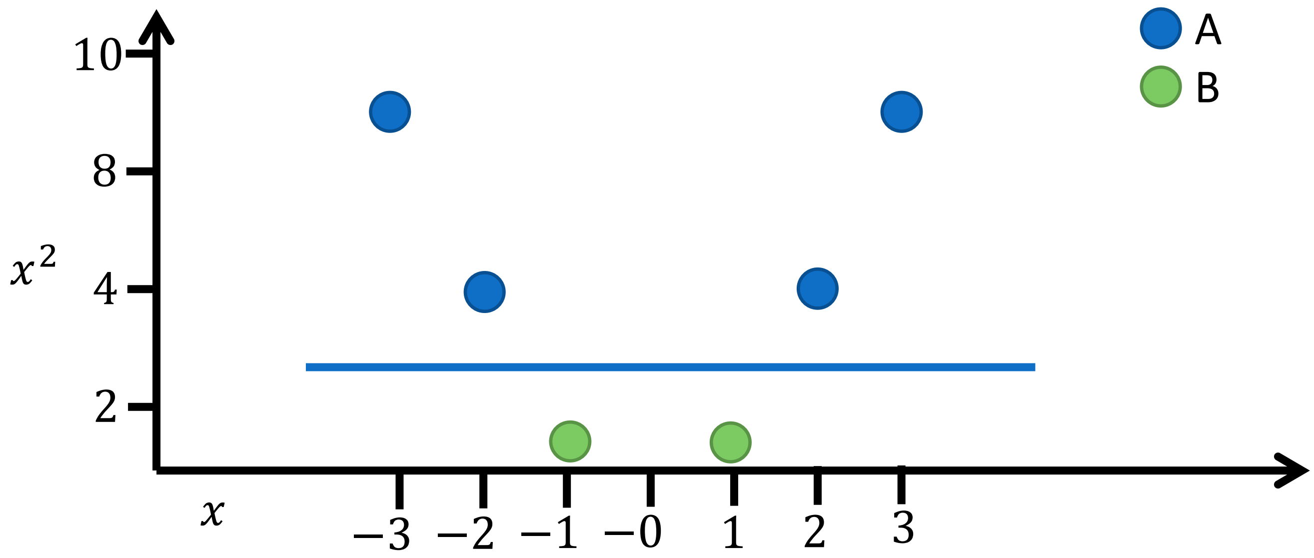
Can a straight line separate out the two classes?

It is not possible to "fit" the data with a single line!

Polynomials: Intuition

- We can always take higher powers of the features to make the data linearly separable.

x	x^2	y
-3	9	A
-2	4	A
-1	1	B
1	1	B
2	4	A
3	9	A



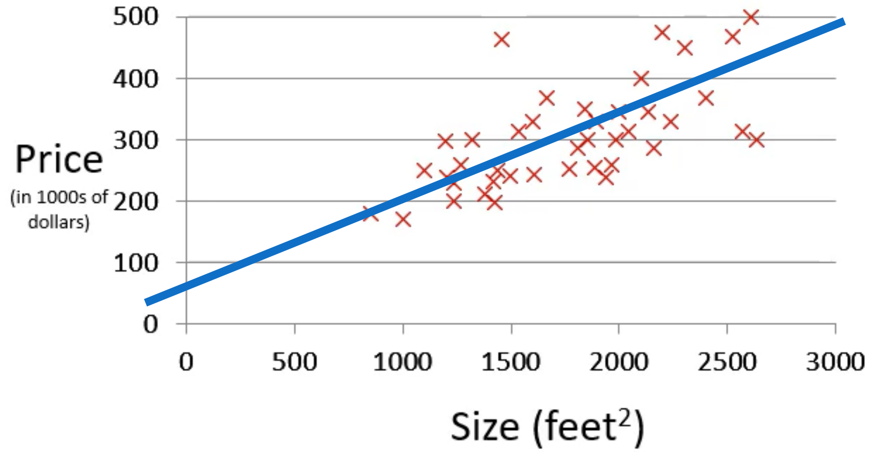
Now we can fit a straight line to separate these two classes.

If it is still not possible to linearly separate the data, maybe adding a 3rd dimension (x_1^3) would do the trick!

Note: This is also called "Polynomial Kernel" and the power that we choose is called degree of polynomial. More on "Kernels" later.

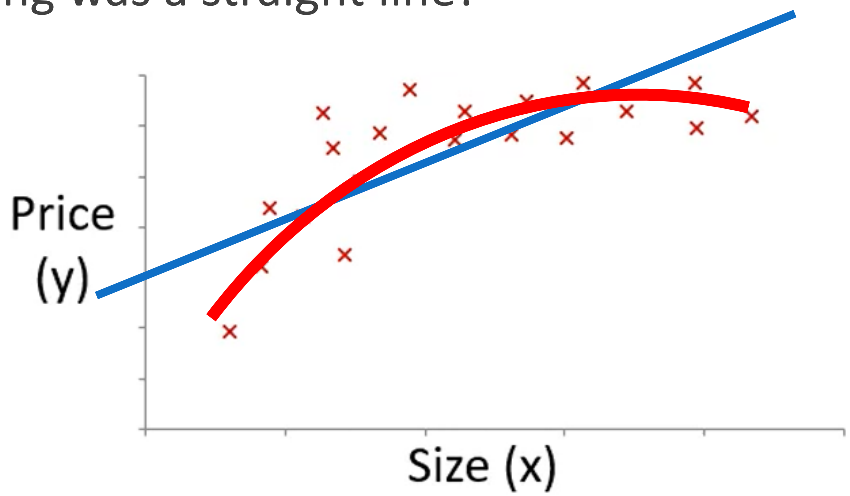
Polynomial Regression

- The relationship between input features and the output label is **Linear**!



- The line we were fitting was a straight line!

A straight line is not a good fit in this case

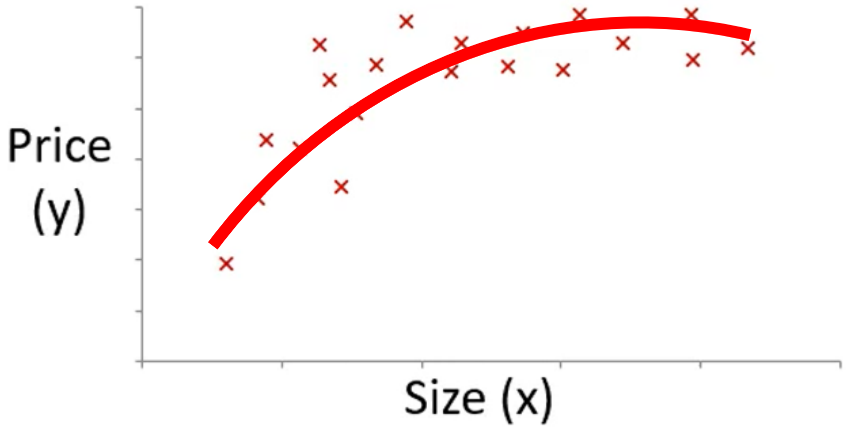


A line with some curve would be a better fit…

Polynomial Regression

- Visualizing Polynomial Degree
  - https://www.desmos.com/calculator

Polynomial Regression



$$w_0 + w_1 x + w_2 x^2$$

$$w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

One feature is now converted to three features!

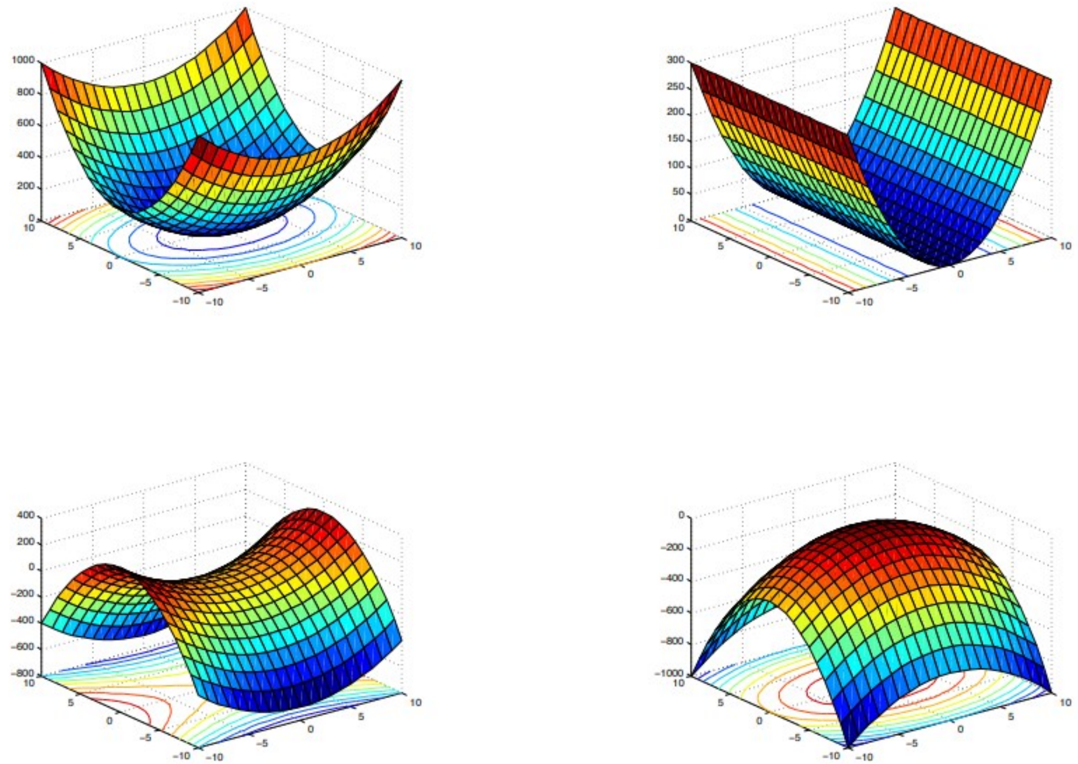
$$h(x) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$$

$$h(x) = w_0 + w_1 (size) + w_2 (size)^2 + w_3 (size)^3$$

$$x_1 = (size)$$
$$x_2 = (size)^2$$
$$x_3 = (size)^3$$

Note: Feature scaling becomes much more important now (as you are taking powers of the original value, making them much much bigger)

Error Surfaces are Still Planes



Credit: http://mezeylab.cb.bscb.cornell.edu/labmembers/documents/supplement%205%20-%20multiple%20regression.pdf

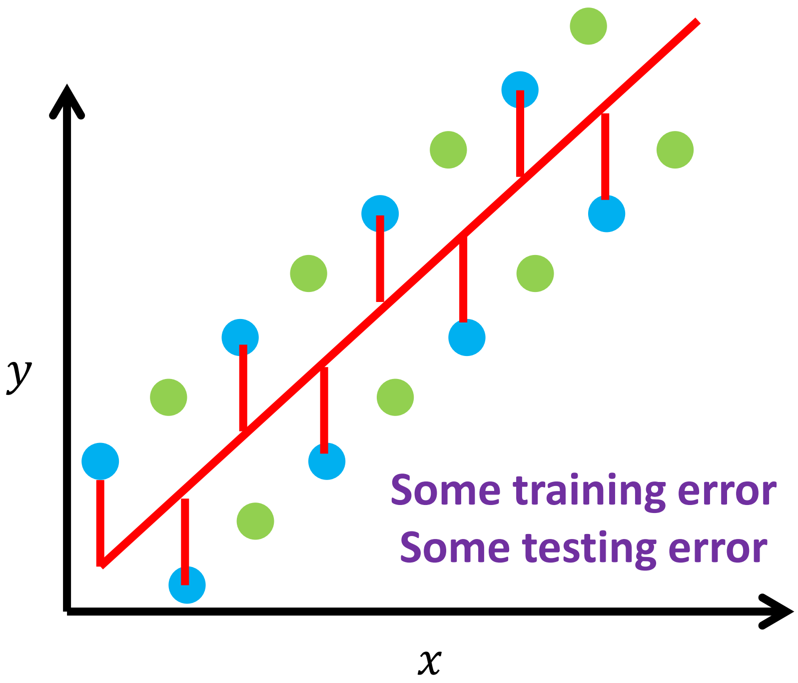
Bias and Variance Tradeoff

The Fitting Problem

- Is it a good idea to always look for 0 training error?

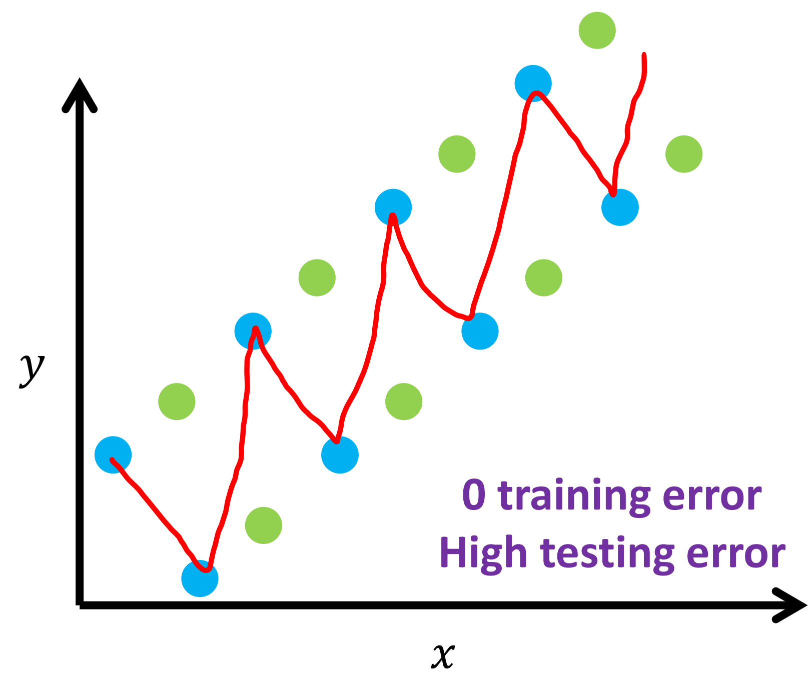
Training Data
Testing Data

What is MSE on this testing data?



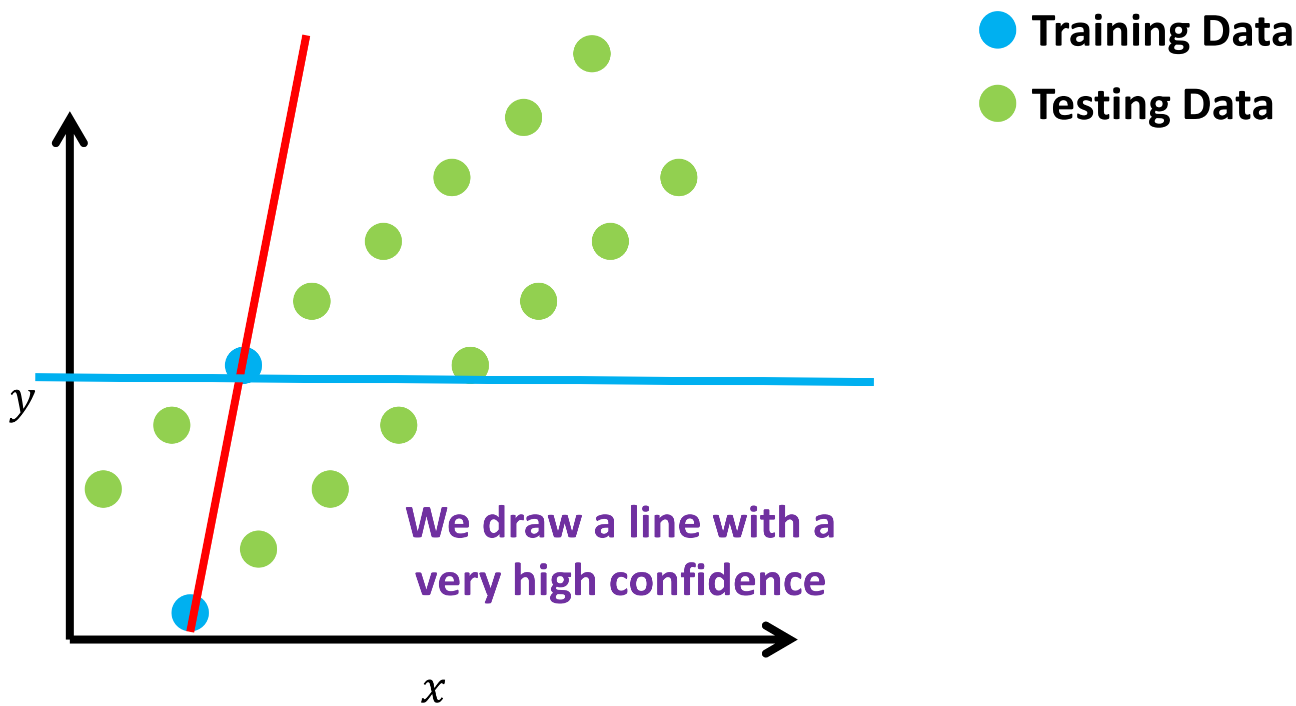
What is MSE on this training data?

What is MSE on this testing data?



What is MSE on this training data?

The Fitting Problem



This is a high variance model as the performance of the model "varies" a lot across train and test datasets.

$$h(x) = w_0 + w_1 x$$

High variance

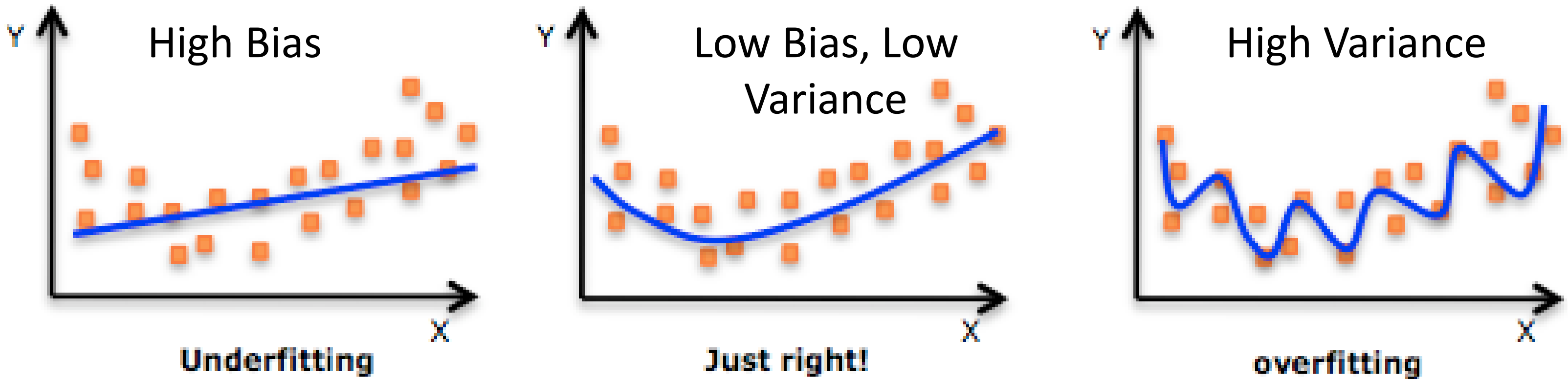
The slope of the line is high that means the value of w_1 is really high (i.e., the model is giving high "weight to feature x.

High bias

The slope of the line is 0 that means the value of w_0 is really high (i.e., high bias). The model is not giving any "weight" to feature x.

The Fitting Problem

Bias and Variance



A Real Example...

Don't learn training data specific features...

Credit: http://www.petanimalexpo.co.nz/

Bias and Variance

- **Bias:** The difference between the average prediction of our model and the correct value which we are trying to predict.
  - If the average predicted values are far off from the actual values, then the bias is high.
  - Model with high bias pays **little attention to the training data** and oversimplifies the model
  - When a model has a high bias, then it implies that the model is too simple and does not capture the complexity of the data, thus **underfitting the data.**
  - It leads to a **high error on both training and test data.**

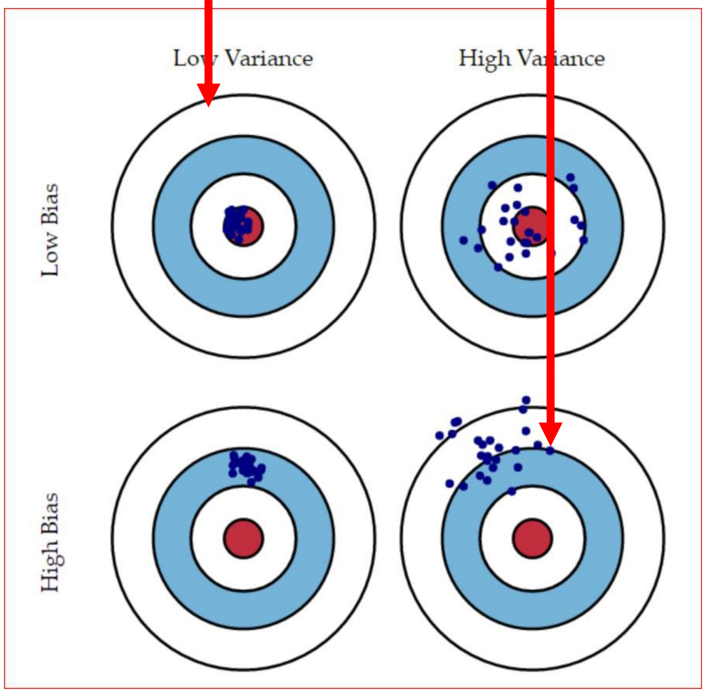
- **Variance:** The variability of model prediction for a across datasets (data points), i.e., how scattered are the predicted values from the actual values.
  - Model with high variance pays a lot of attention to the training data and does not generalize on the data which it has not seen before.
  - As a result, such **model performs very well on training data, but has high error rates on test data.**
  - High variance causes **overfitting** that implies that the algorithm models random noise present in the data.

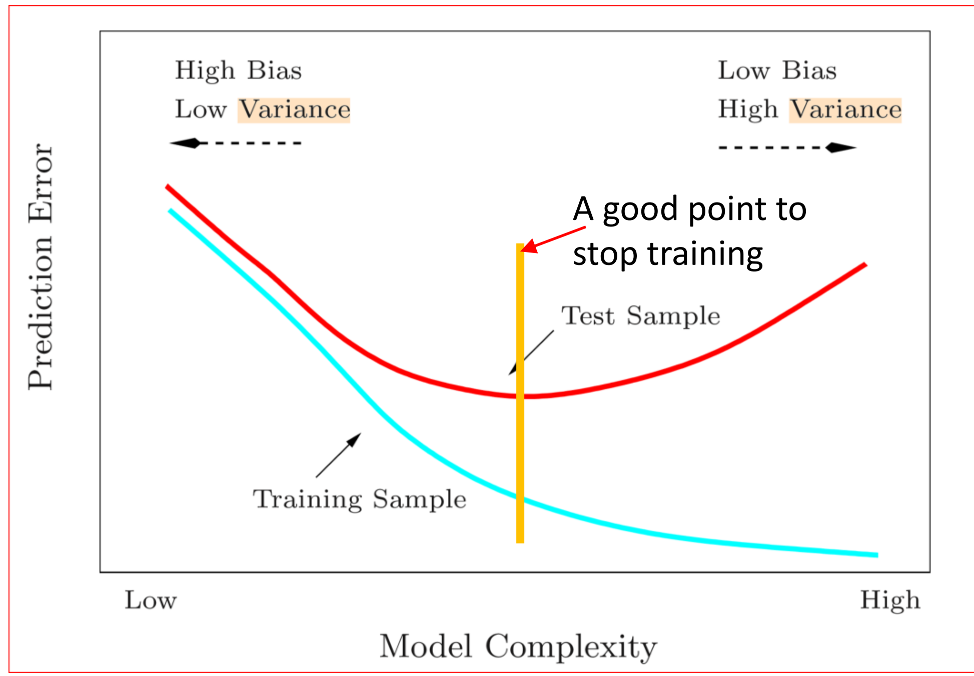
Credit: Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani and Jerome Friedman
https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Bias and Variance

Ideal Situation

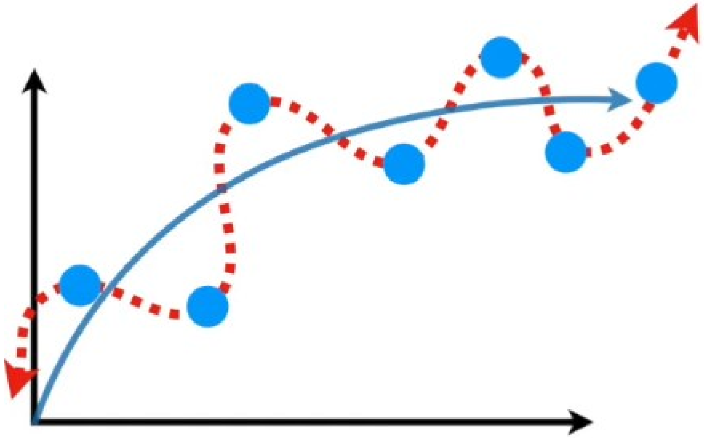
Worst Situation



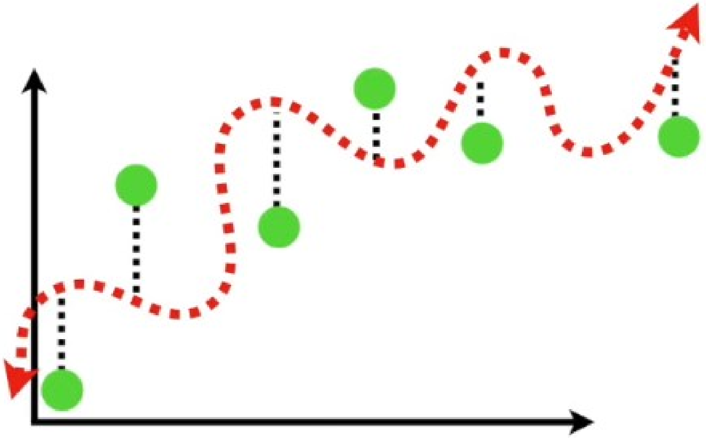


Credit: Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani and Jerome Friedman
https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229
https://medium.com/datadriveninvestor/bias-and-variance-in-machine-learning-51fdd38d1f86

Low Bias and High Variance



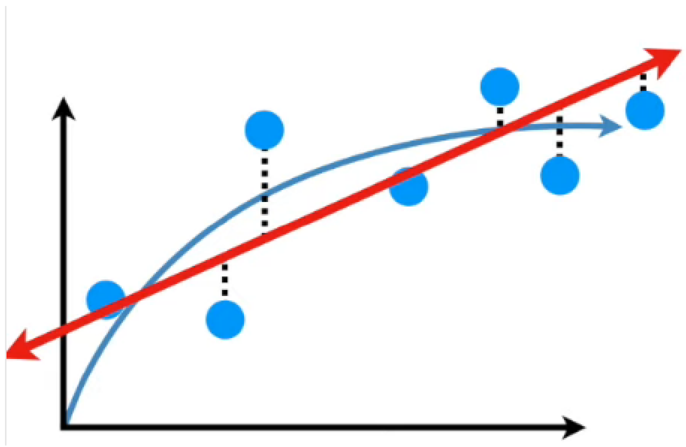
Training Dataset



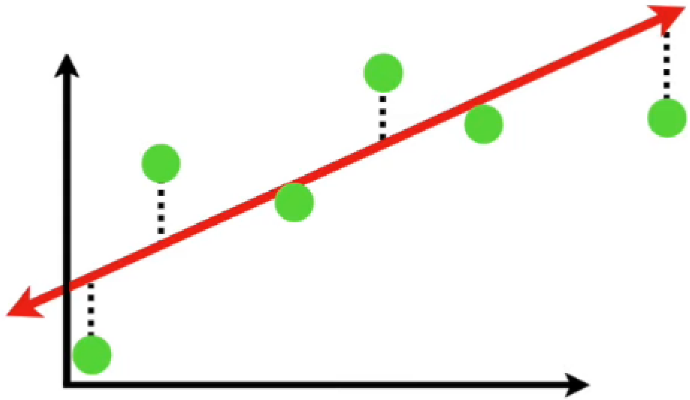
Testing Dataset

Because this model fits well to the training set (blue dots) but not so well on the testing set (green dots), we say that the model is "Overfitting"

High Bias and Low Variance



Training Dataset

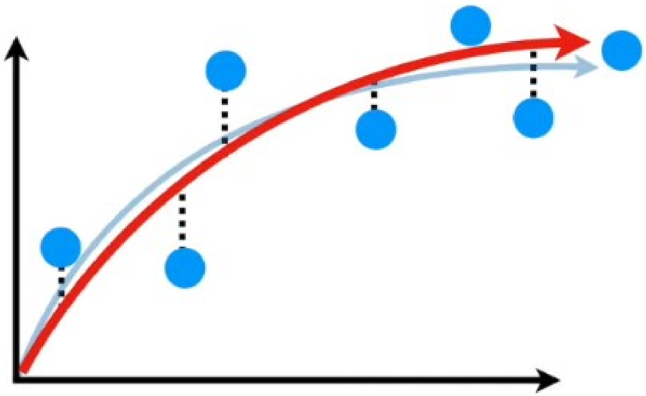


Testing Dataset

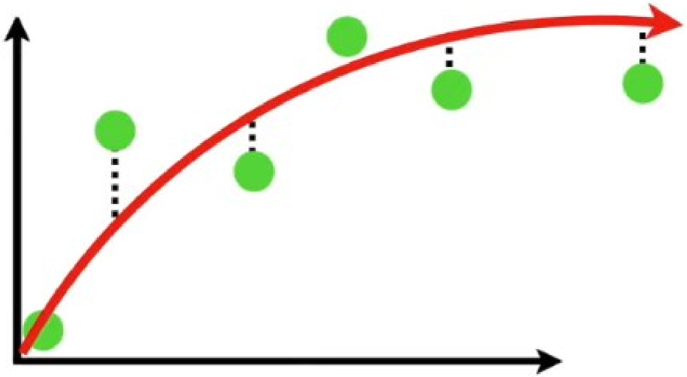
Because this model does not fit to the training set (blue dots) but fits well on the testing set (green dots), we say that the model is "Underfiting"

An Ideal Algorithm should...

- **Have lower bias** so it can accurately model true relationship
- **Have low variability** so it can predict consistently across different datasets (splits)!



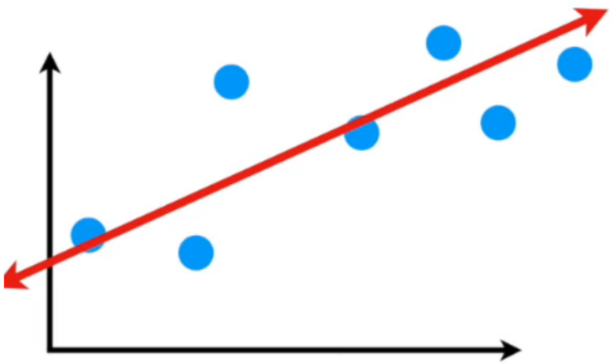
Training Dataset



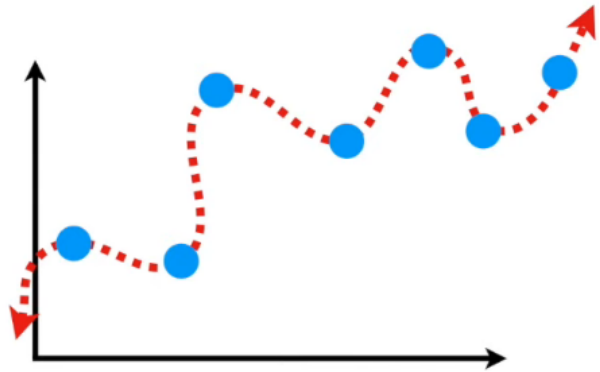
Testing Dataset

Bias-Variance Tradeoff

- This is achieved by finding sweet spot between simple model (left) and complex model (right)



Simple Model



Complex Model

How to find the sweet spot between Bias and Variance?

Finding Sweet Spot between Simple and Complex Model

- **Bagging**
- **Feature Reduction**
  - Feature Selection (Statistical, Automated, and Manual)
  - Feature Extraction
- **Regularization**
  - Reduce magnitude/values of parameters w_j
  - Works well when we have a lot of features, each of which contributes a bit to prediction
- **Boosting**

These solutions are used to remove overfitting

Bias and Variance: Summary

- **High Bias:**
  - High Training Error
  - Validation Error or Testing Error is Close to Training Error
- **High Variance:**
  - Low Training Error
  - High Validation Error or High Testing Error
- **Fixing High Bias (possibly): It's due to simple model.**
  - Add more input features
  - Add more complexity by introducing polynomial features
  - Decrease regularization term

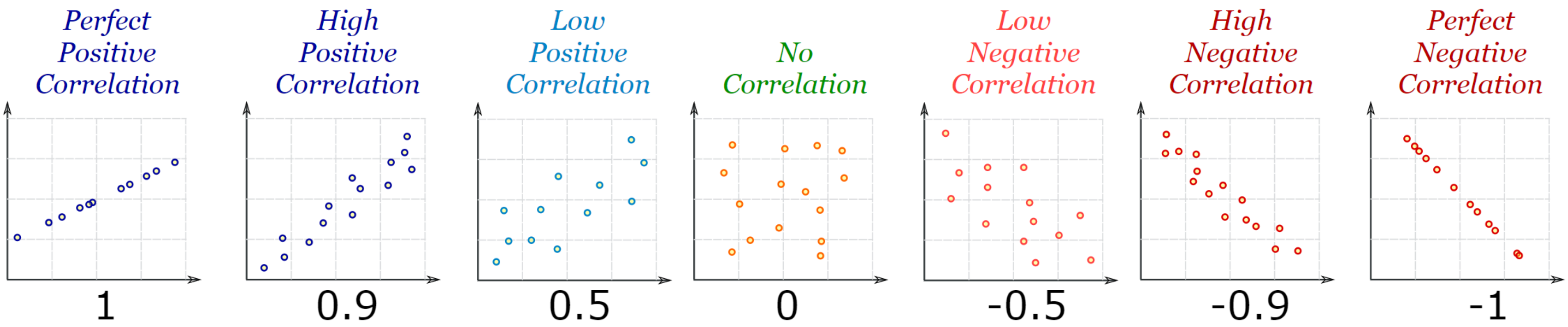
More on regularization later…

- **Fixing High Variance (possibly): It's due complex model.**
  - Getting more training data
  - Reduce input features
  - Increase regularization term

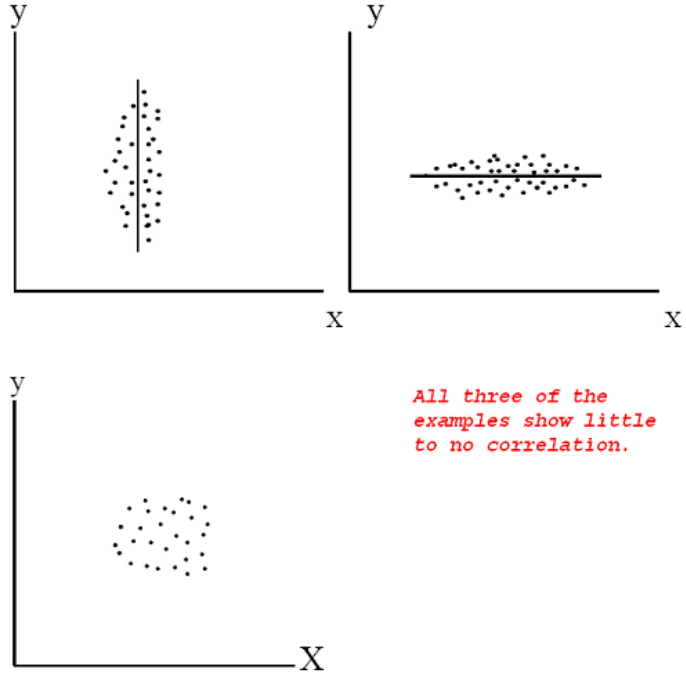
Credit: Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani and Jerome Friedman
https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Manual Feature Selection

❑ Recall Correlation



Should we have a low or high correlation between input features and the output label?



Manual Feature Selection: Example

- Consider a dataset that has an output label “# shark attacks”

# swimmers	watched _jaws	temp	stock_price	# attacks
…	…	…	…	…
…	…	…	…	…

- **attacks:** Number of shark attacks (output variable)
- **swimmers:** Number of swimmers in water
- **watched_jaws:** Percentage of swimmers who watched iconic Jaws movies
- **temp:** Average temperature of the day
- **stock_price:** The price of your favorite tech stock that day (a totally unrelated variable)

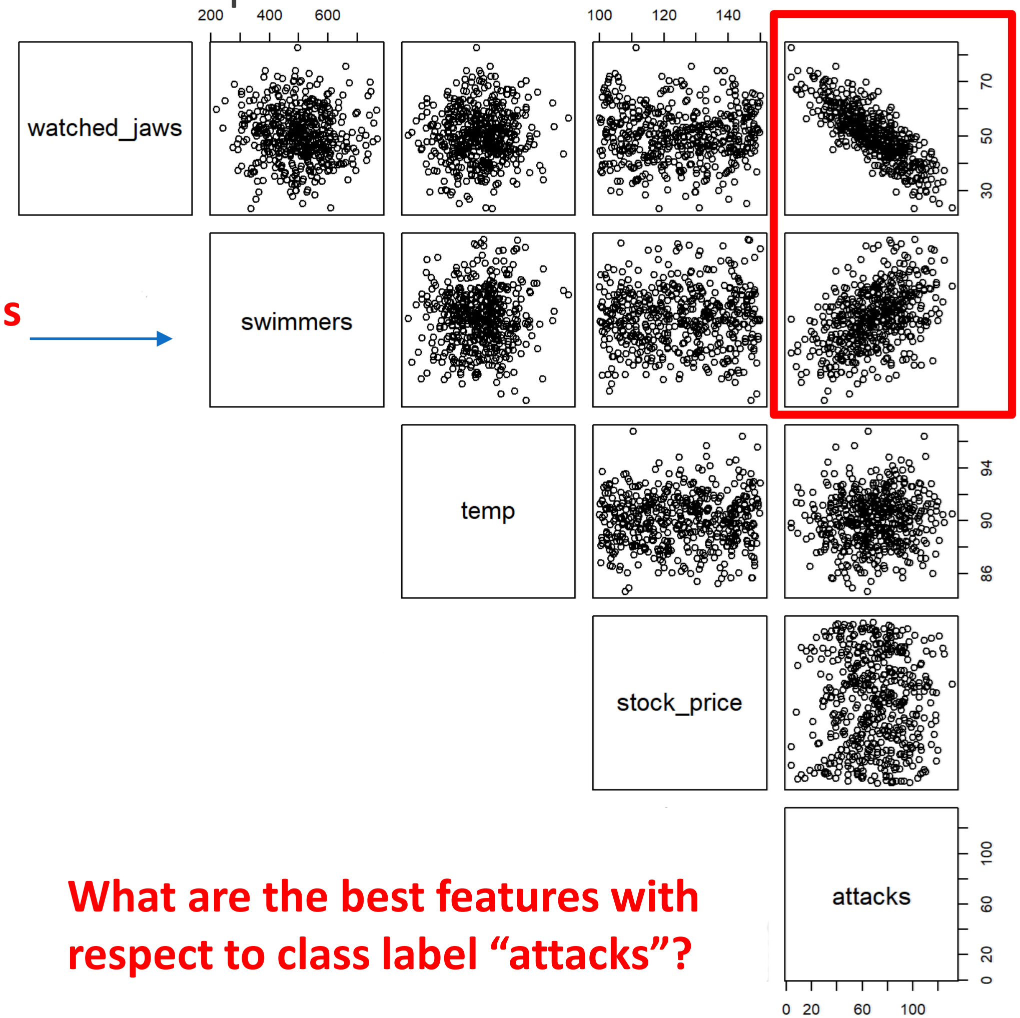
Credit: https://scienceloft.com/technical/understanding-lasso-and-ridge-regression/

Manual Feature Selection: Example

- Plot all features and label as a scatter plot

Why "watched_jaws" is important?

Why "# swimmers" is important?



What are the best features with respect to class label "attacks"?

Automatic Feature Selection

- **Can we somehow "minimize" the contribution of least important features in the output?**
  - **Lasso Regression**

Reference

- Josh Strammer (StatQuest)
  - https://www.youtube.com/channel/UCtYLUTtgS3k1Fg4y5tAhLbw

Book Reading

- Murphy – Chapter 7