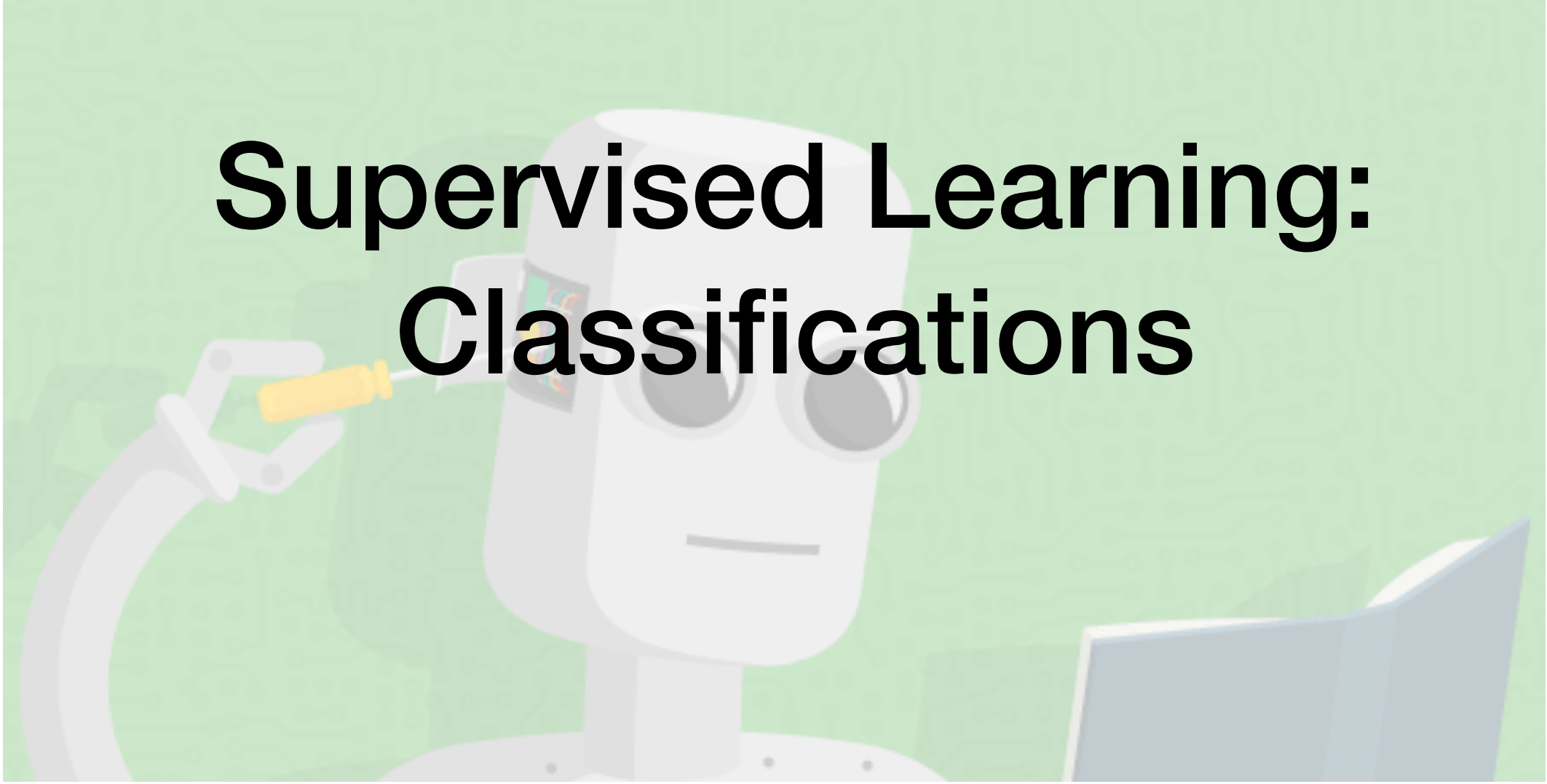
Supervised Learning: Classifications



Dr. Fabien Plisson

Chemoinformatics in Drug Discovery

LANGEBIO, UGA CINVESTAV

October 15-18, 2019 - Irapuato, Mexico

Program

What is Supervised Learning?

Linear Regressions & Assumptions

Regression metrics

Overfitting & Cross-validations

Other regressions

Program

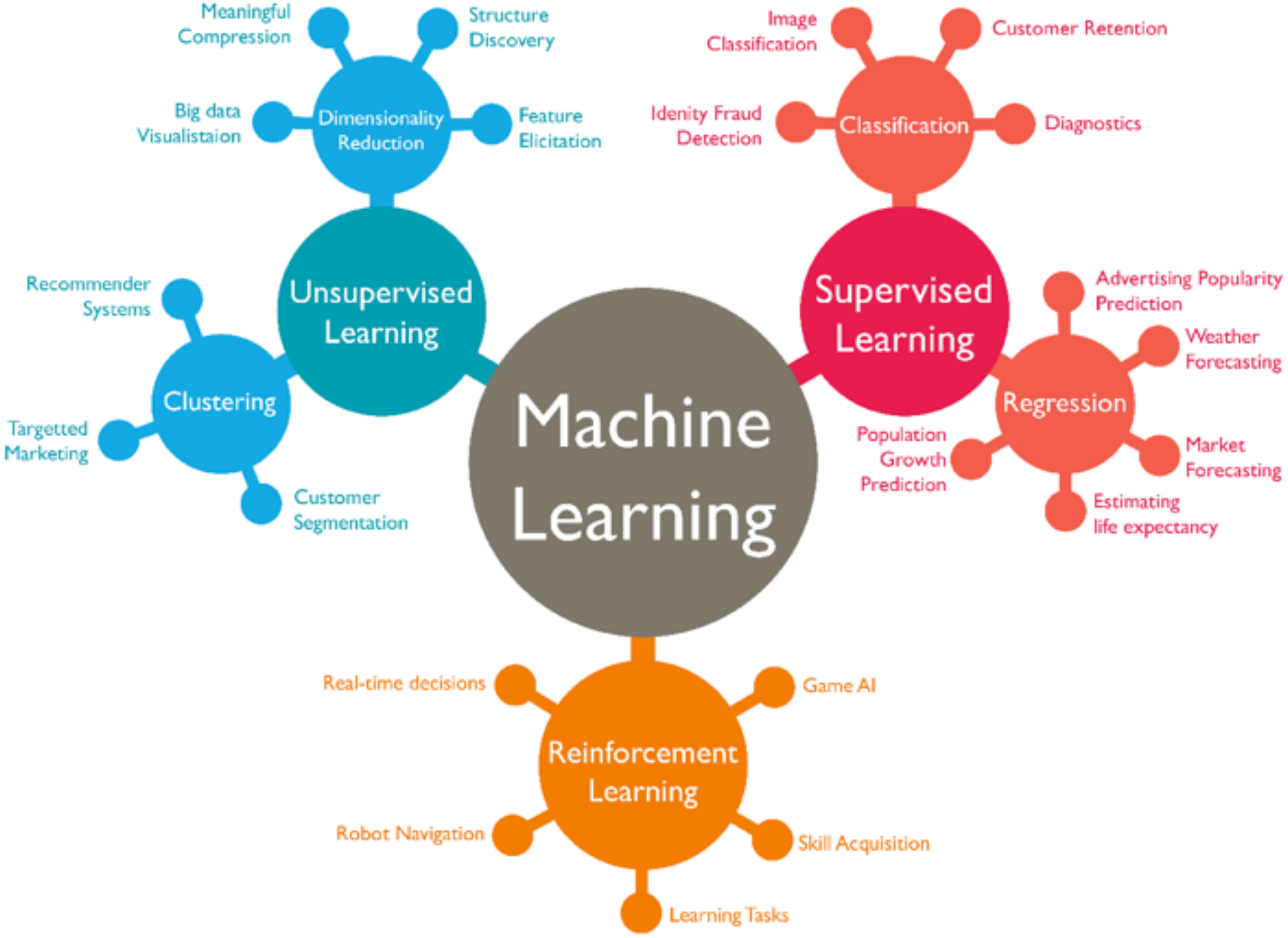
What is Supervised Learning?

Linear Regression & Assumptions

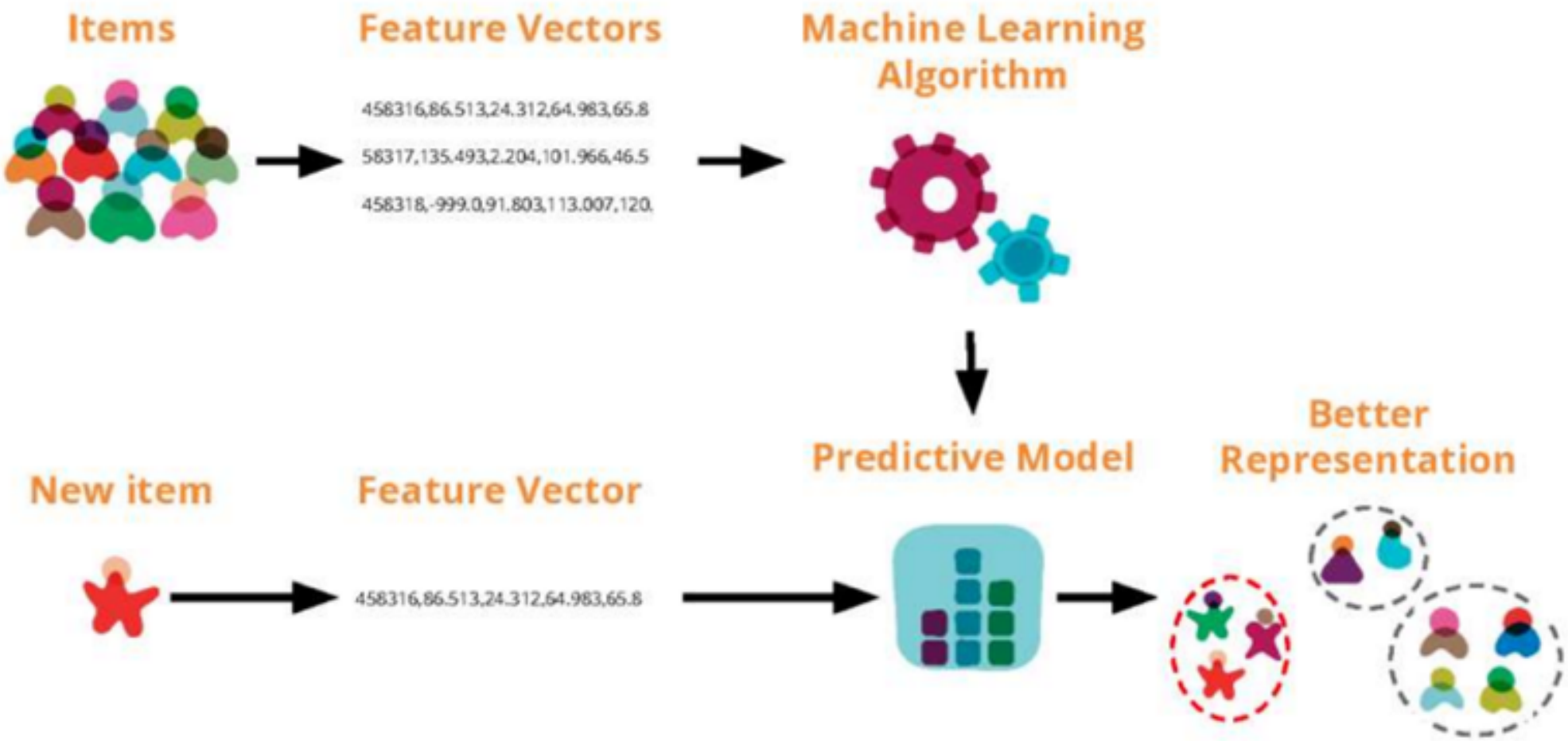
Regression metrics

Overfitting & Cross-validations

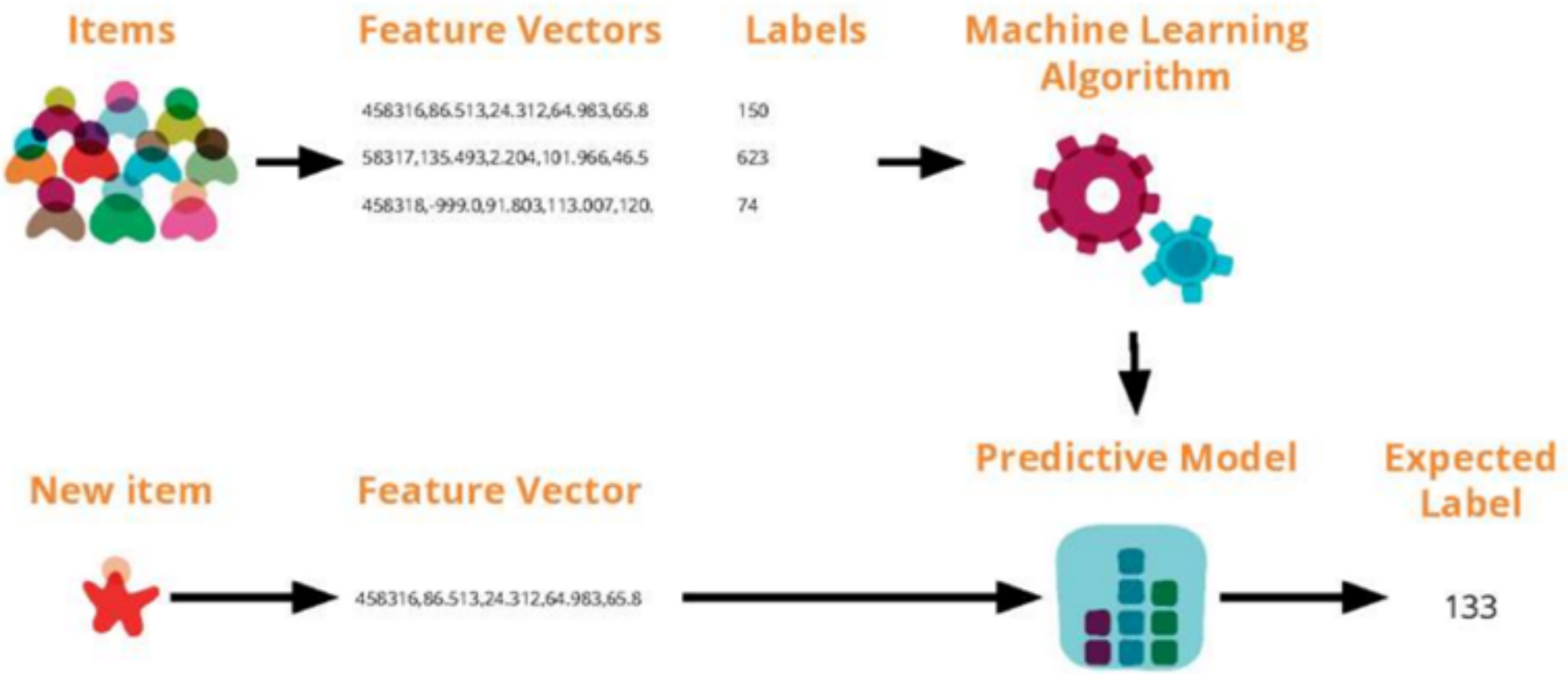
Other regressions


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Unsupervised Learning



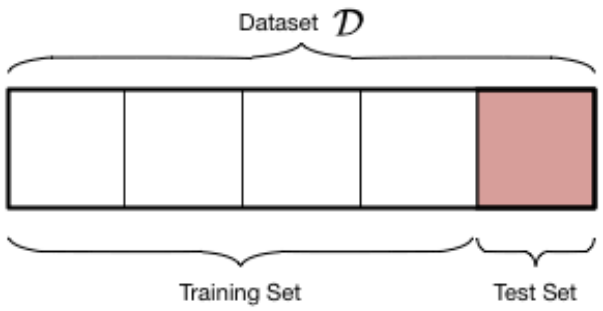
Supervised Learning

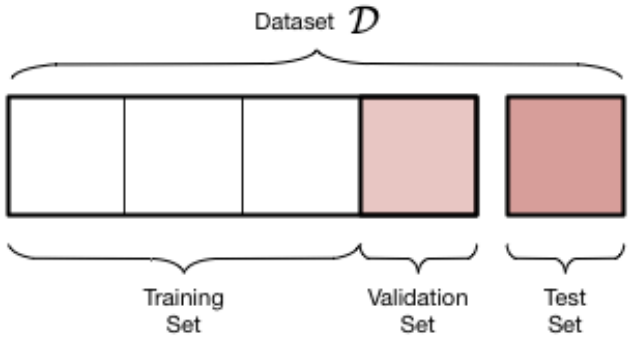


Supervised Learning

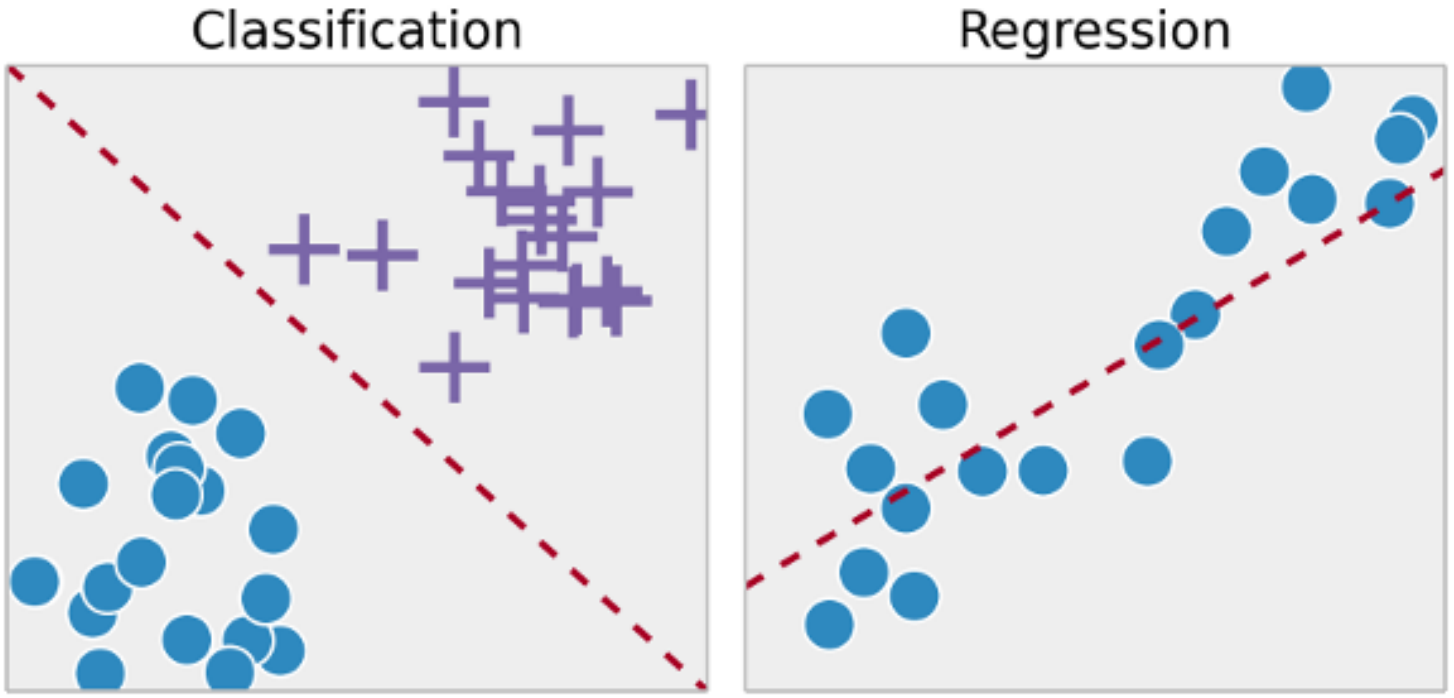
Algorithm that uses common knowledge (**training** dataset) about a topic to predict the responses (to fit a model) for new set of input values (**testing** dataset).

A third dataset, known as **validation** dataset, including input values and responses, part of the training set, is sometimes used to tune hyper-parameters or to reduce overfitting.





Classification vs. Regression



A **classification** model predicts **discrete values** (i.e. classes).

A **regression** model predicts **continuous values** based on past data.

Classification vs. Regression

Assign the response into a **class**.

Predict new input to belong to one class or another.

Discrete / Categorical variable

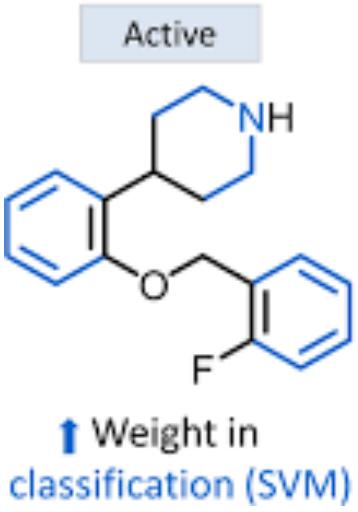
Metric: **Accuracy**

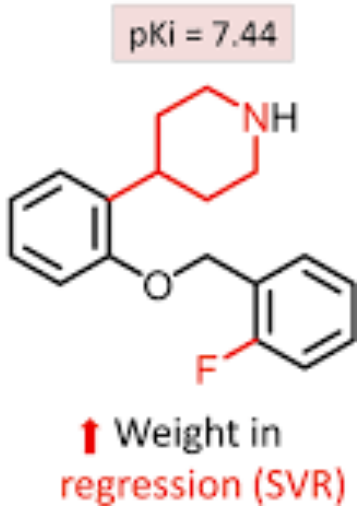
Assign the response into a real **number**.

Predict new input with a new number.

Continuous variable

Metric: **Sum-of-squares error or r^2**





Popular Algorithms

Classification

Decision Trees

Random Forest

Gradient Boosting

Support Vector Machine(s)

K-Nearest Neighbors

Naïve Bayes

Logistic Regression

Regression

Linear Regression

OLS Regression

LOESS (local) Regression

Polynomial Regression

Supervised Dimension Reduction

Genetic Algorithms

Linear Discriminant Analysis (LDA)

Generalized Discriminant Analysis (GDA)

Program

What is Supervised Learning?

Linear Regression & Assumptions

Regression metrics

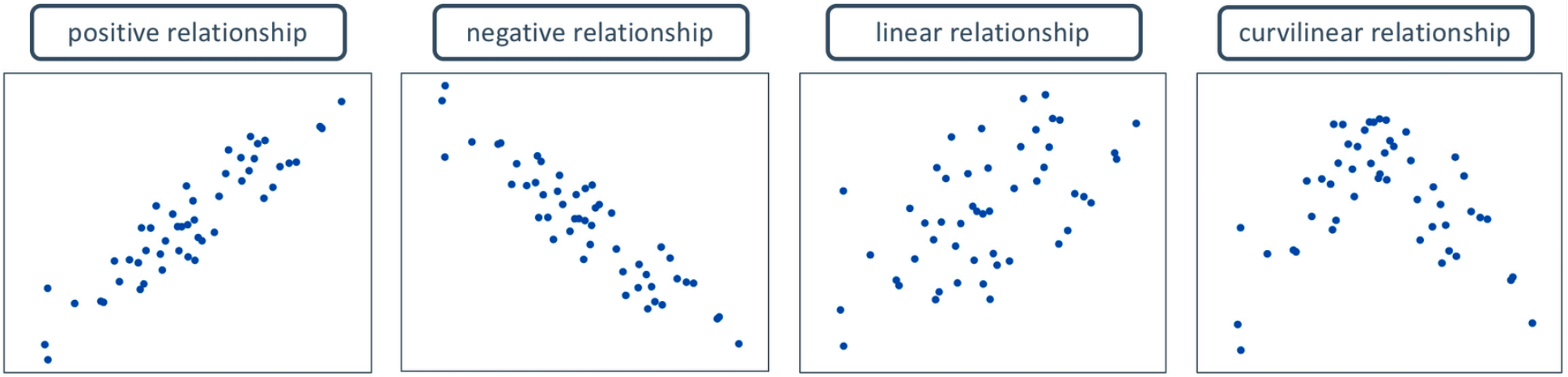
Overfitting & Cross-validations

Other regressions

Correlation between variables

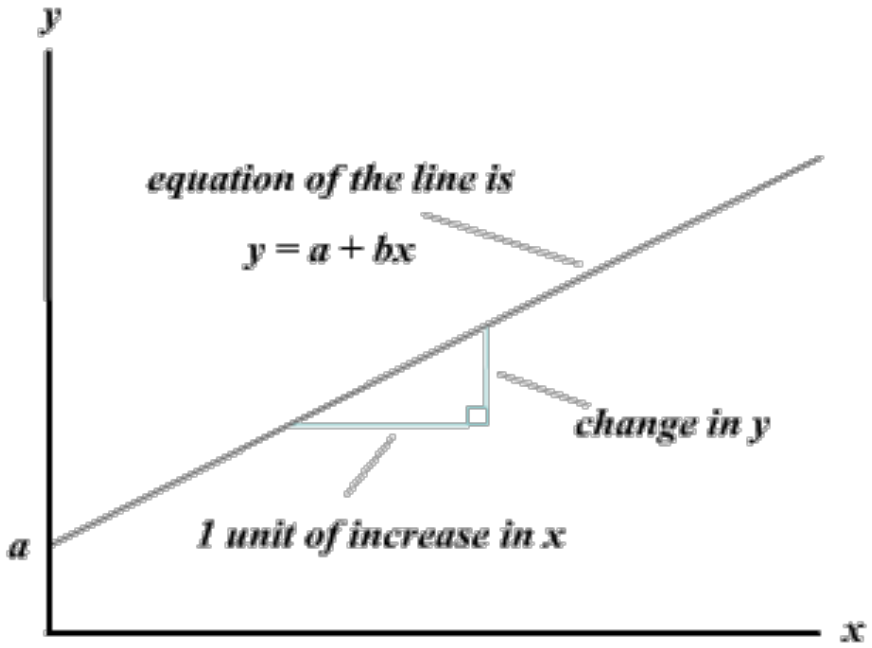
In regression and in statistical modelling, we want to understand the relationship between **an output variable** (a response) and one or more input variables (**univariate or multivariate**).

Simple linear regression is used to model the relationship between two continuous variables.



Simple Linear Regression

When the relationship has a linear (or straight-line) pattern, the correlation provides a numerical measure of the strength and direction of the relationship. We can analyse the data further **by finding an equation for the straight line** that best describes that pattern. That equation predicts the value of the response variable from the value of the explanatory variable. The straight line that best describes the linear pattern is called **the regression line**.

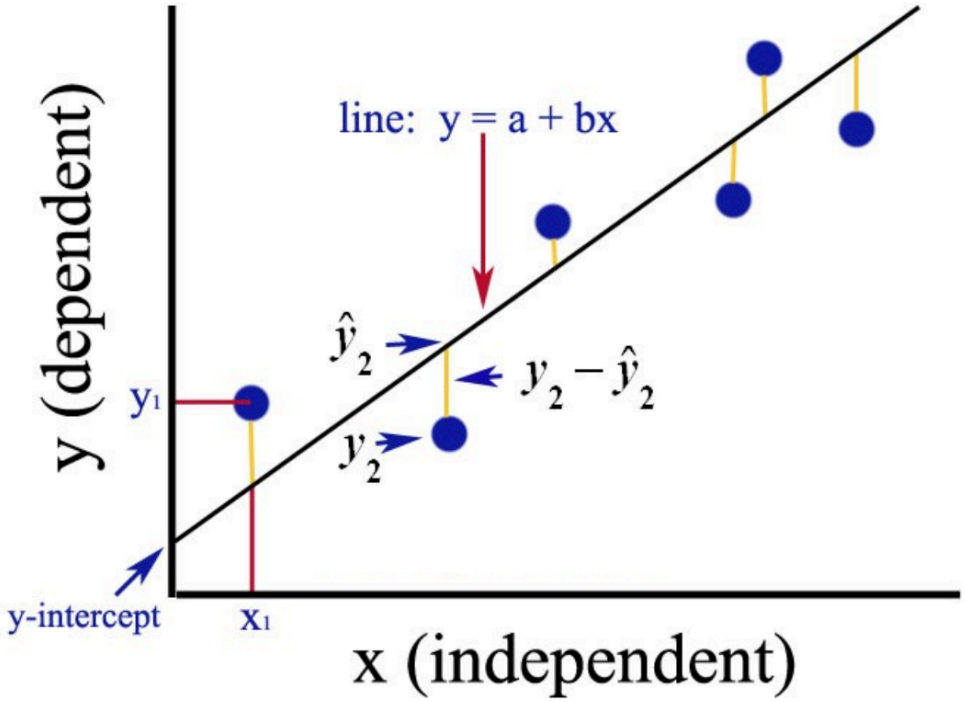


a = y-intercept
b = slope coefficient

Ordinary Least Squares (OLS) Linear Regression

Least squares is one of the methods to find the best fit line for a dataset using linear regression.

The most common application is to create a straight line that **minimises the sum of squares of the errors** generated from the differences in the observed value and the value anticipated from the model.



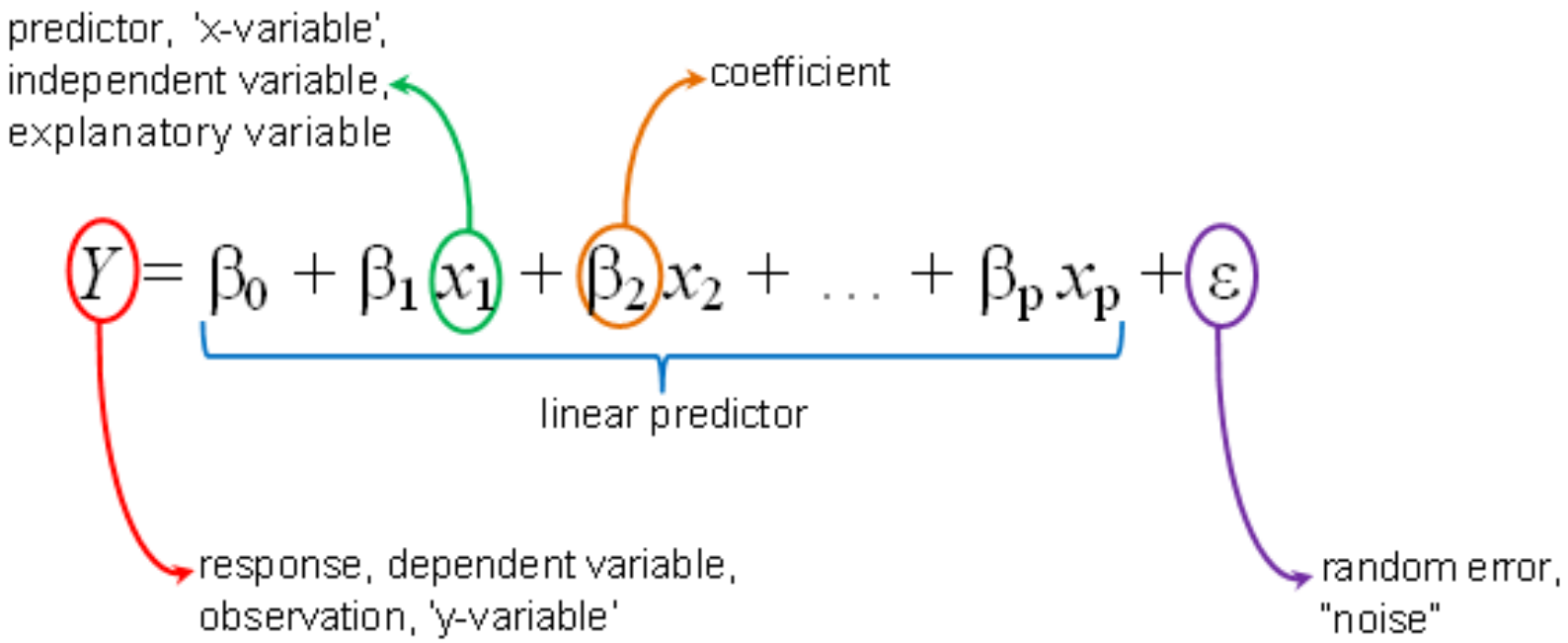
a = y-intercept
b = slope coefficient

y (hat) = observed value
y = predicted value

Minimize: $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Least Squares Method

General Equation Form



What are the assumptions of a linear regression model?

The relationship between the two variables is **linear**.

All variables are **multivariate normal**. The residuals of the linear regression should be normally distributed around a mean of 0.

There isn't much **multicollinearity** among the dependent variables.

There is **homoscedasticity**. That is, the size of the error does not vary by the sizes of the independent variables. The error does not increase substantially if your variables get larger or smaller.

There is no **outliers**.

Assumptions - Linearity

(a) Perfect positive linear relation, $r = 1$

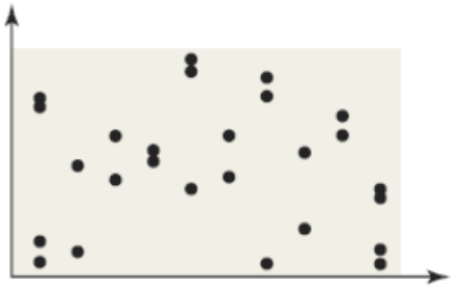
(b) Strong positive linear relation, $r \approx 0.9$

(c) Moderate positive linear relation, $r \approx 0.4$

(d) Perfect negative linear relation, $r = -1$

(e) Strong negative linear relation, $r \approx -0.9$

(f) Moderate negative linear relation, $r \approx -0.4$

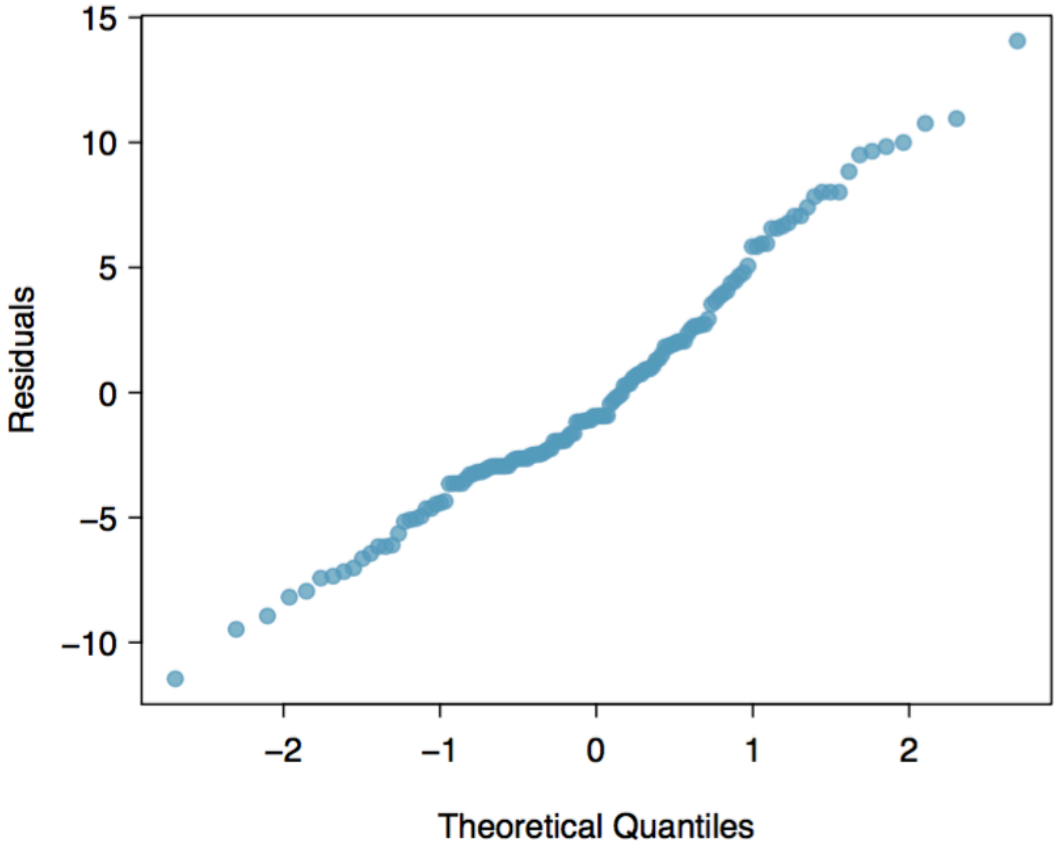


(g) No linear relation, r close to 0.

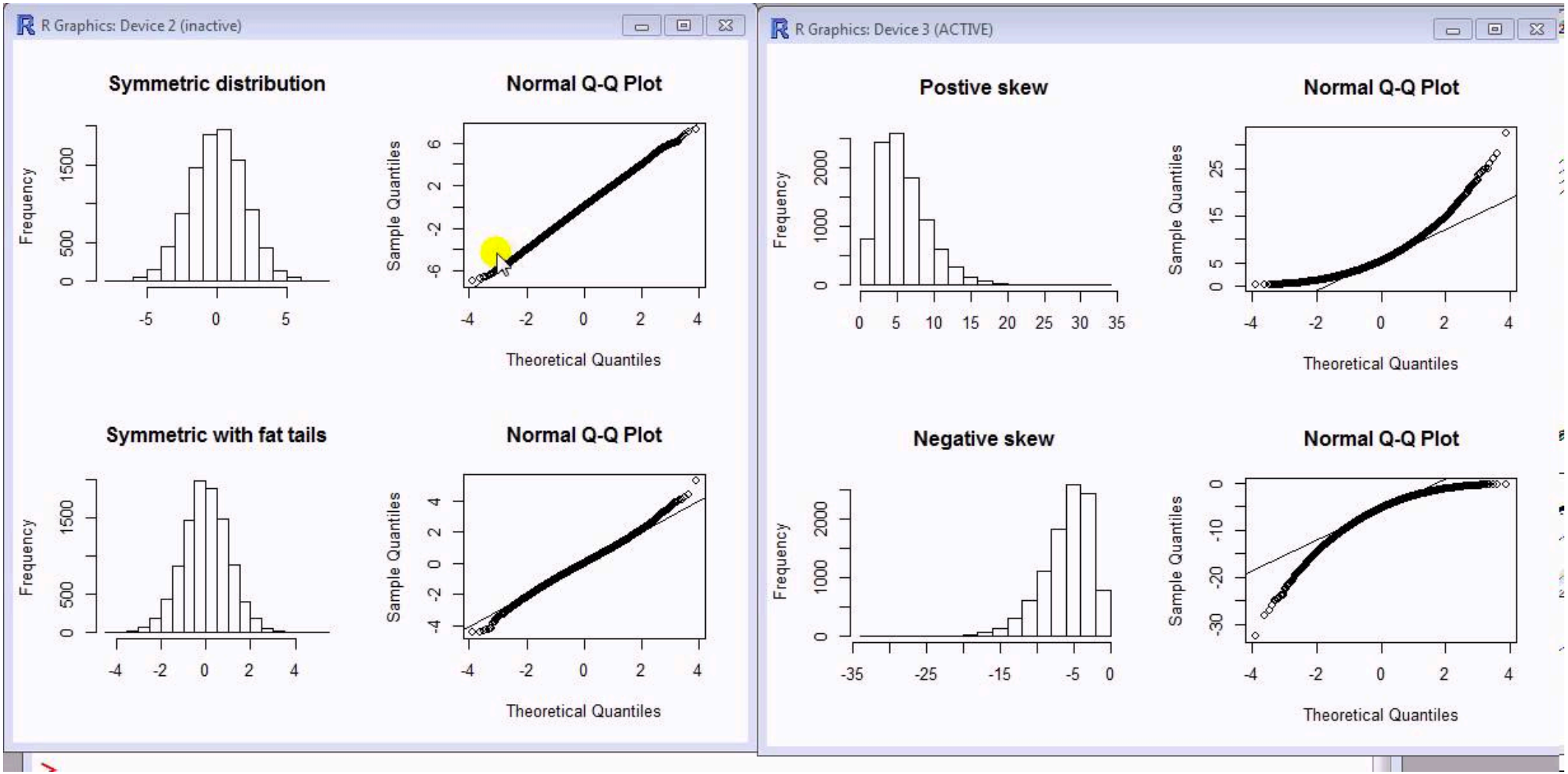
(h) No linear relation, r close to 0.

Assumptions - Normal Distribution with Q-Q plot

Normal probability plot. A normal probability plot of the residuals is shown in Figure 8.8. While the plot exhibits some minor irregularities, there are no outliers that might be cause for concern. In a normal probability plot for residuals, we tend to be most worried about residuals that appear to be outliers, since these indicate long tails in the distribution of residuals.



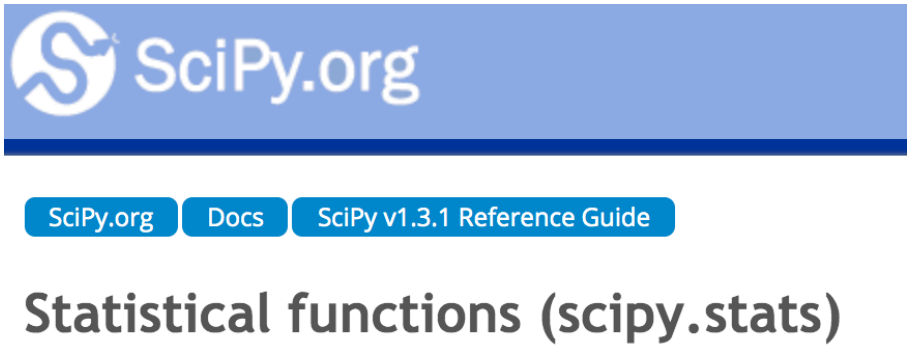
Assumptions - Normal Distribution with Q-Q plot



Assumptions - Normal Distribution with Tests

In statistics, normality tests are used to determine if a data set is well-modelled by a normal distribution and to compute how likely it is for a random variable underlying the data set to be normally distributed.

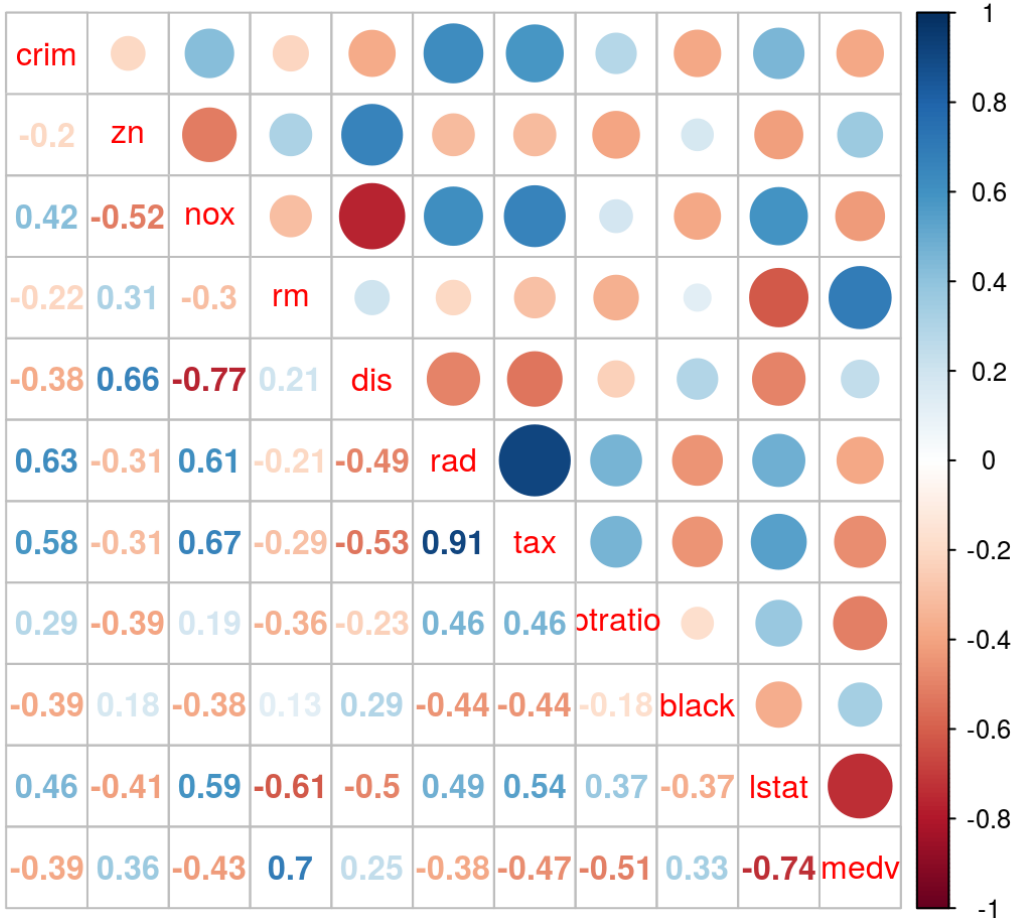
Shapiro-Wilk test (common choice)
Kolmogorov-Smirov test
D'Agostino's K-squared test
Pearson's Chi-squared
…



Assumptions - Multicollinearity

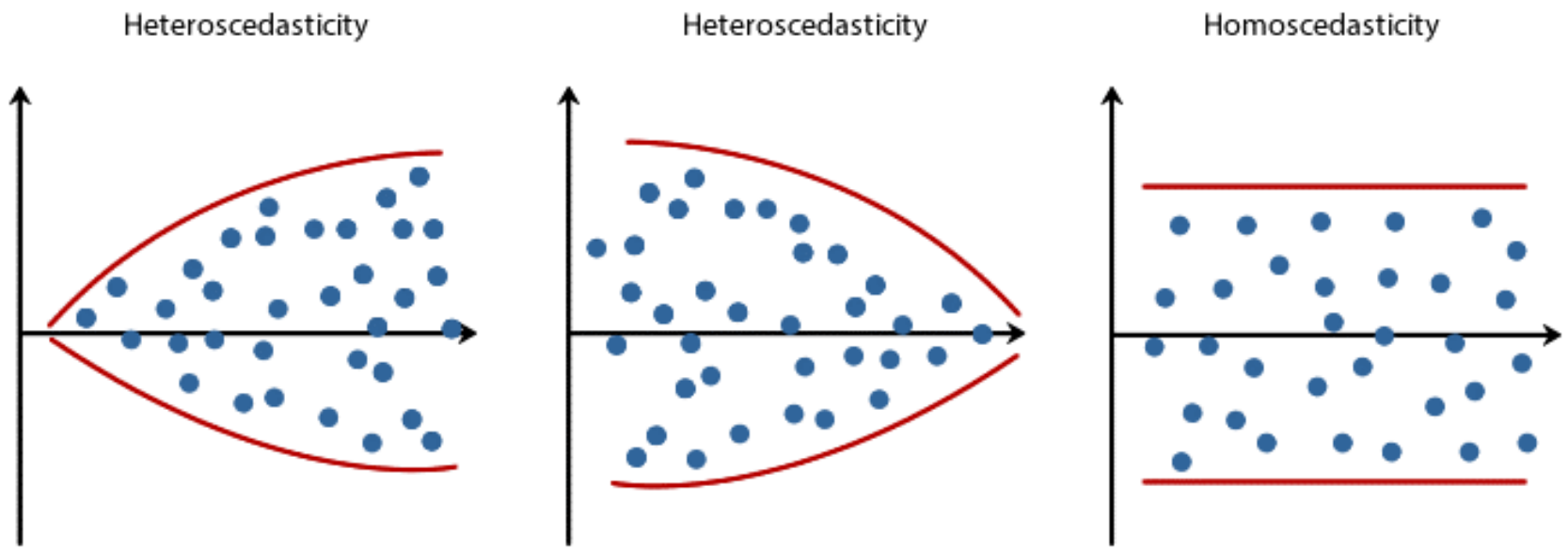
Multicollinearity occurs when the independent variables are too highly correlated with each other.

Solution: remove highly correlated variables - bring limited information to the prediction.



Assumptions - Homoscedasticity

Homoscedasticity refers to the assumption that the dependent variable Y exhibits similar amounts of variance across the range of values for an independent variable x_i.

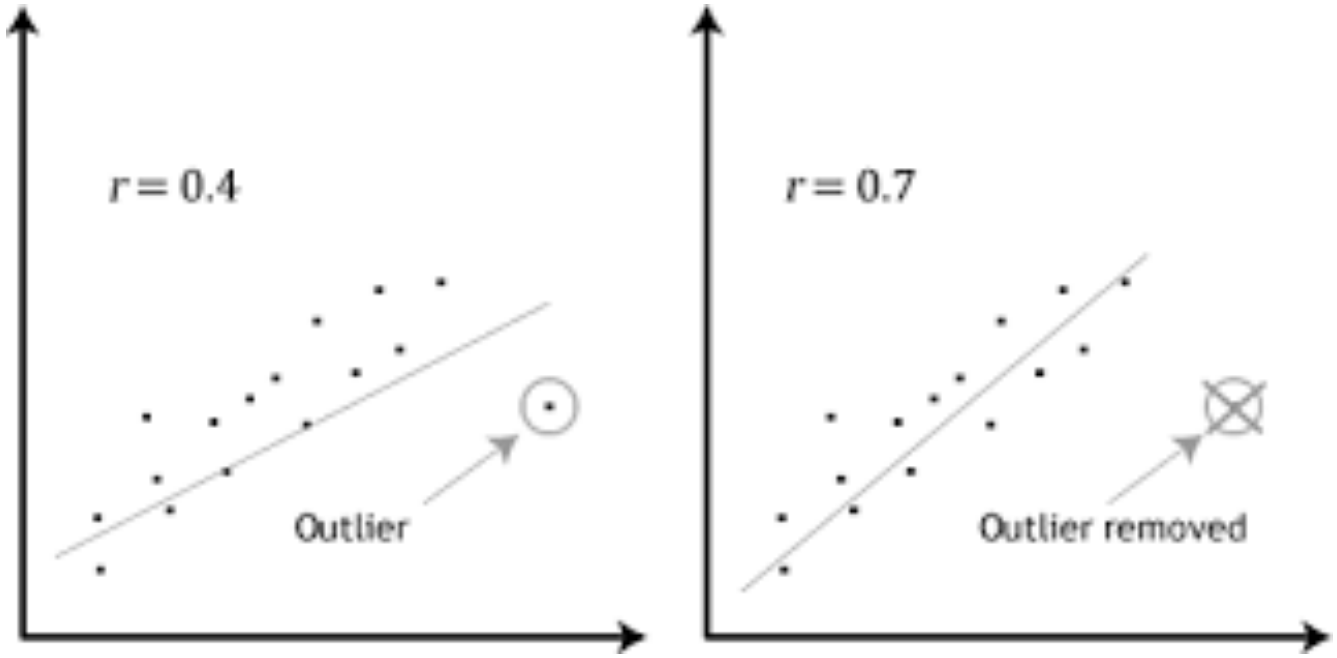


Solution: applying transformation to Y

Assumptions - Outlier detection

Outliers are atypical observations that can affect the prediction.

Solution: remove outliers



Program

What is Supervised Learning?

Linear Regression & Assumptions

Regression metrics

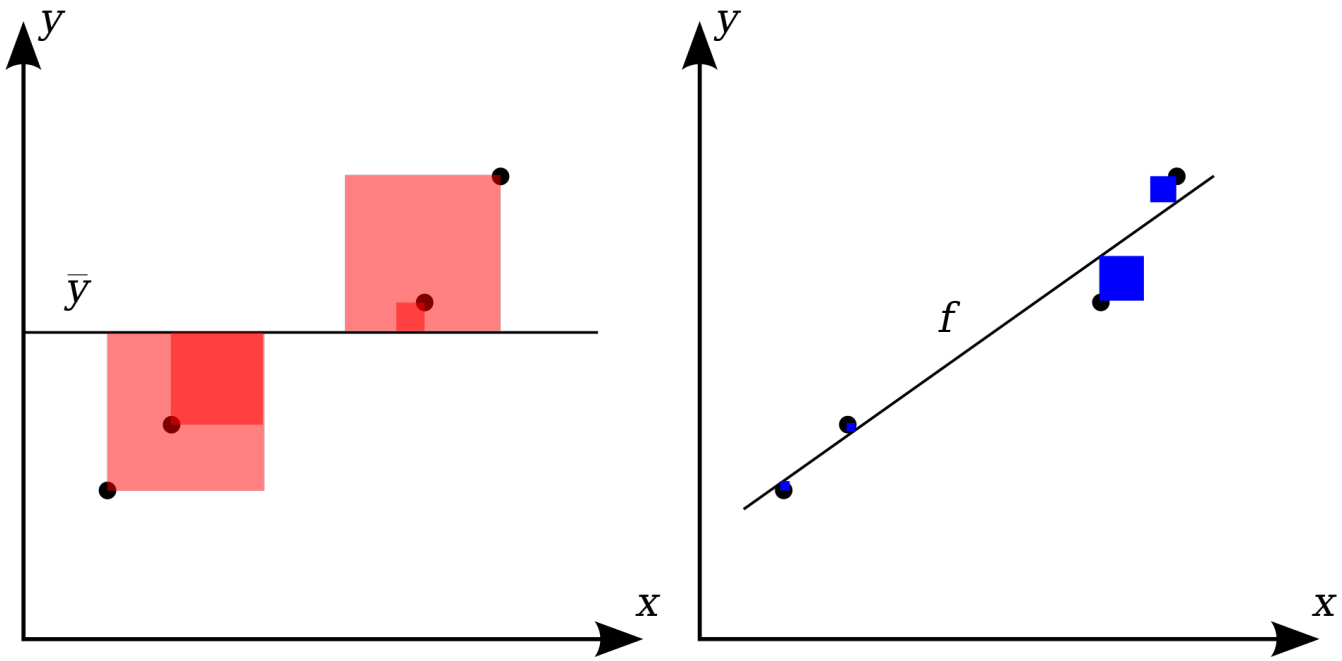
Overfitting & Cross-validations

Other regressions

Coefficient of determination R^2 "R squared"

R^2 is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2, \qquad SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$



$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

Adjusted R^2

The **adjusted R-squared** is a modified version of **R-squared** that has been **adjusted** for the number of predictors in the model.

The **adjusted R-squared** increases only if the new term improves the model more than would be expected by chance.

n = number of observations
q = number of predictors

$$R^2_{adj} = 1 - \left(\frac{\left(1 - R^2\right)(n-1)}{n-q} \right)$$

Root Mean Squared Error (RMSE)

Root-mean-square error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values observed.

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

Akaike's Information Criterion (AIC)

Bayesian Information Criterion (BIC)

The inclusion of **additional variables** can increase error and lead to **overfitting**. Both **AIC** and **BIC** attempt to resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC. The lower AIC/BIC is, better is the model.

$$\mathrm{AIC}_i = -2\log L_i + 2p_i$$

$$\mathrm{BIC}_i = -2\log L_i + p_i \log n$$

Scikit-learn: Regression Metrics

See the Regression metrics section of the user guide for further details.

metrics.explained_variance_score (y_true, y_pred)	Explained variance regression score function
metrics.max_error (y_true, y_pred)	max_error metric calculates the maximum residual error.
metrics.mean_absolute_error (y_true, y_pred)	Mean absolute error regression loss
metrics.mean_squared_error (y_true, y_pred[, …])	Mean squared error regression loss
metrics.mean_squared_log_error (y_true, y_pred)	Mean squared logarithmic error regression loss
metrics.median_absolute_error (y_true, y_pred)	Median absolute error regression loss
metrics.r2_score (y_true, y_pred[, …])	R^2 (coefficient of determination) regression score function.

https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

Program

What is Supervised Learning?

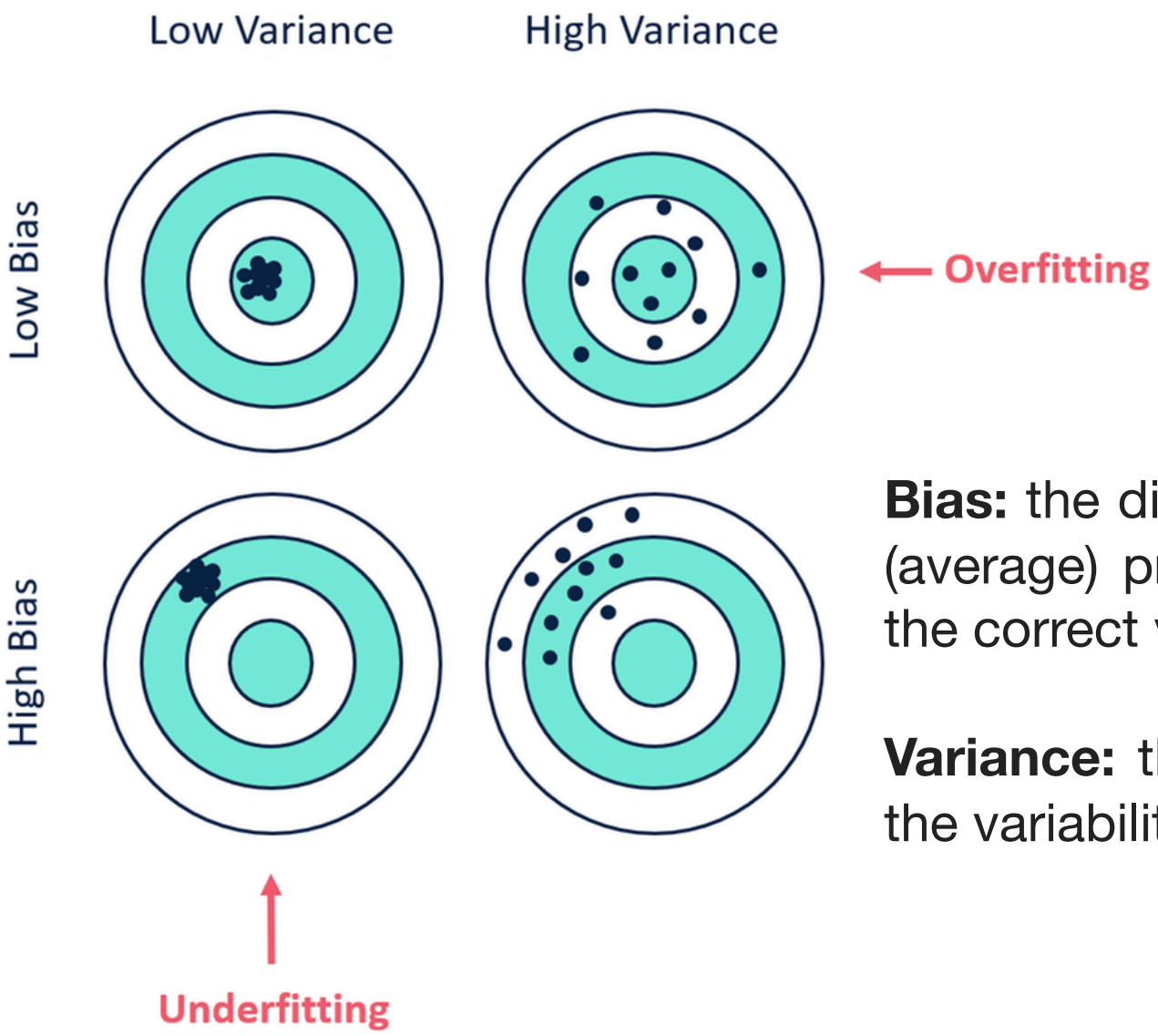
Linear Regression & Assumptions

Regression metrics

Overfitting & Cross-validations

Other regressions

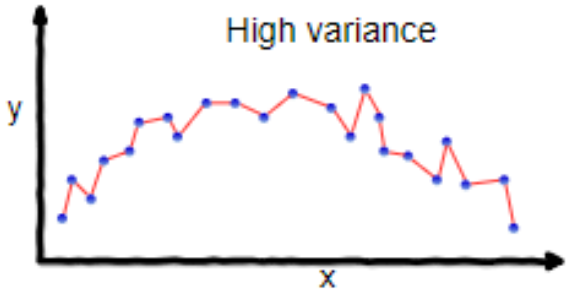
Bias vs. Variance



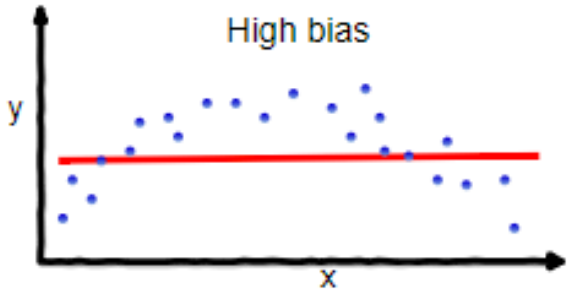
Bias: the difference between expected (average) prediction of the model and the correct value.

Variance: the amount which indicates the variability of any model prediction.

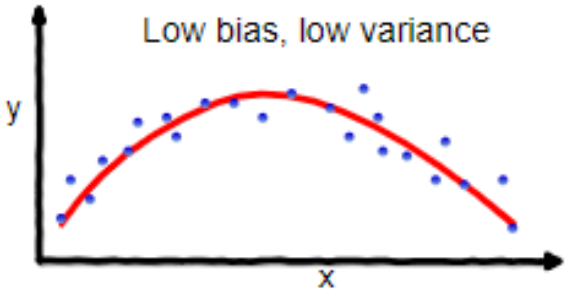
Underfitting & Overfitting



overfitting

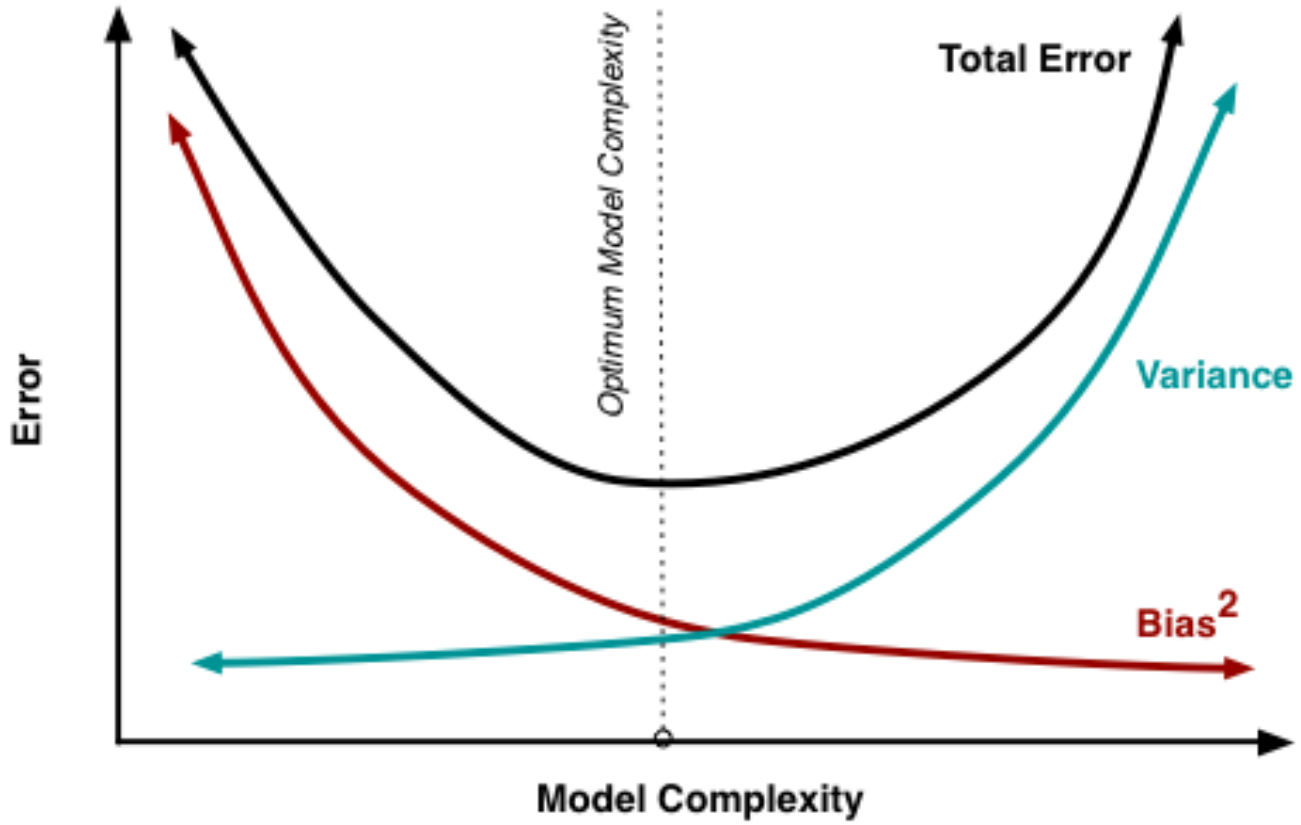


underfitting

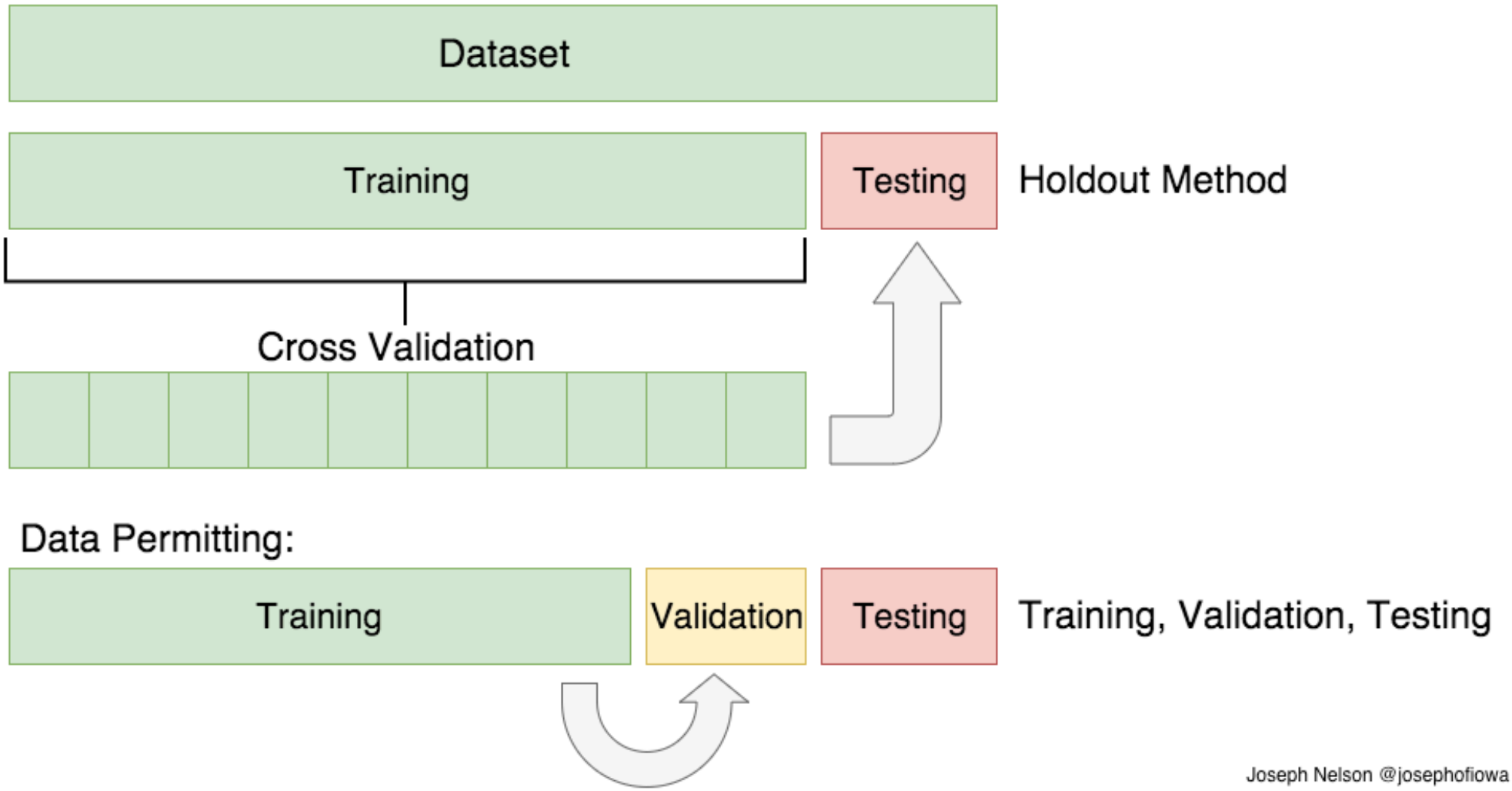


Good balance

Bias-Variance Tradeoff

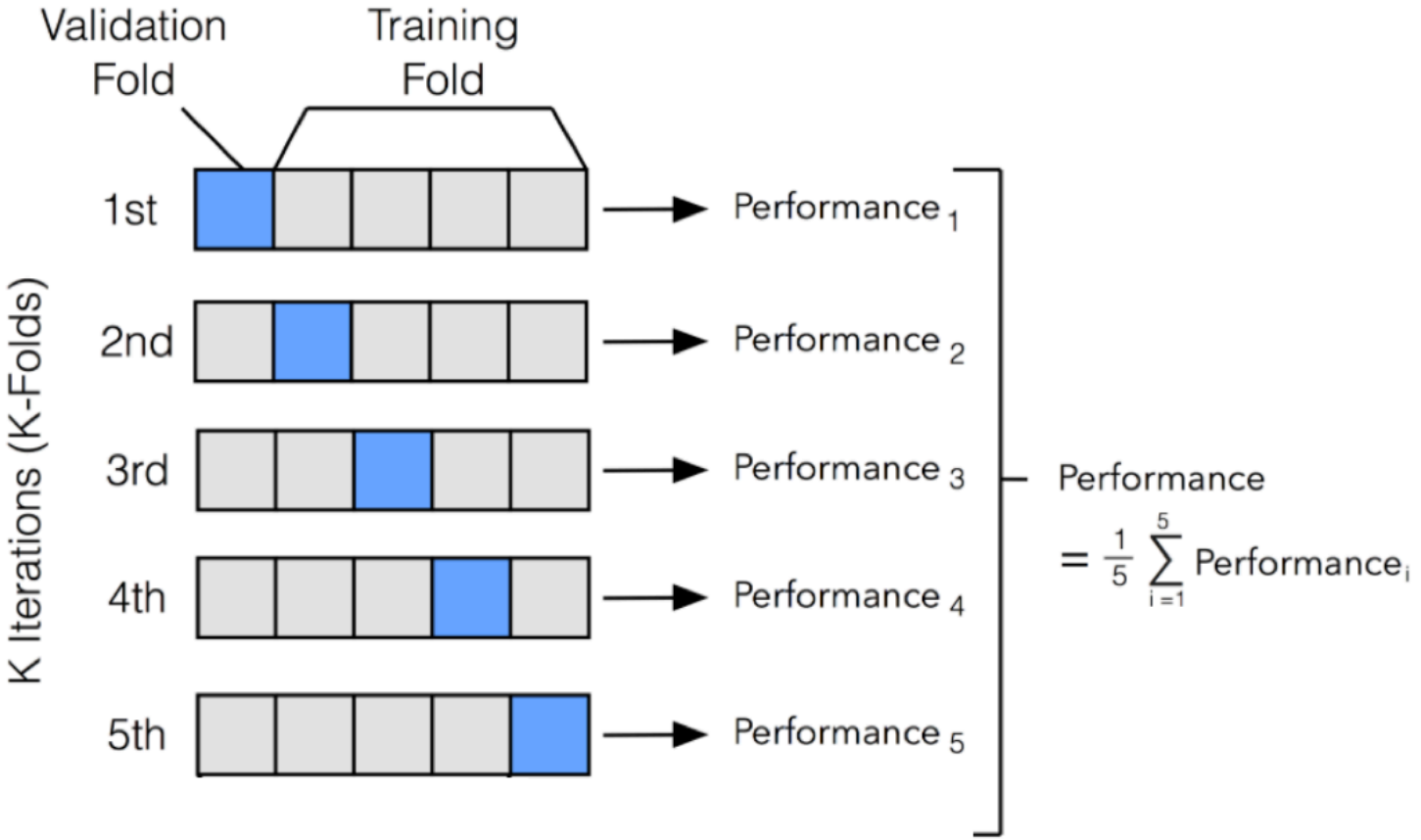


Solution: Cross-validation (CV) methods

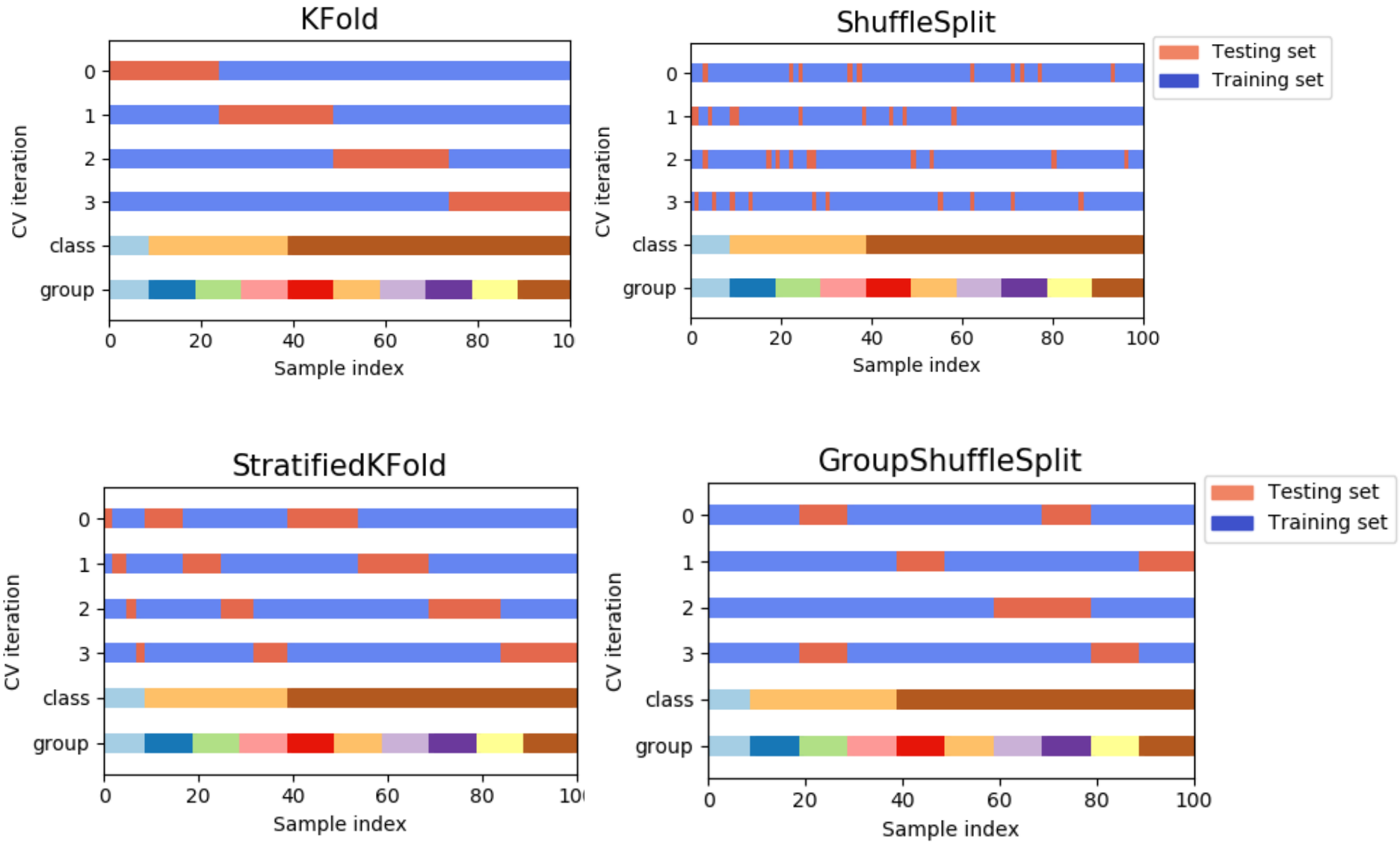


Example: K-fold CV

1. Divide Model set into K folds.
2. Use K-1 folds for training and 1 for testing.
3. Repeat steps 1 & 2, K times, rotating the test set.
4. Determine an expected performance metric (RMSE, R^2…) based on the results across the iterations.



Scikit-Learn: Cross-validation methods



https://scikit-learn.org/stable/modules/cross_validation.html

Program

What is Supervised Learning?

Linear Regression & Assumptions

Regression metrics

Overfitting & Cross-validations

Other regressions

Reducing Overfitting with regularised linear regressions

Regularisation favours simpler models to more complex models to prevent your model from overfitting to the data. Regularised regressions apply penalties (L1, L2 or both) to the coefficients of linear models **shrinking** some to zero.

Least Absolute Shrinkage Selector Operator (LASSO)

$$L = \sum(\hat{Y}i - Yi)^2 + \lambda \sum |\beta|$$

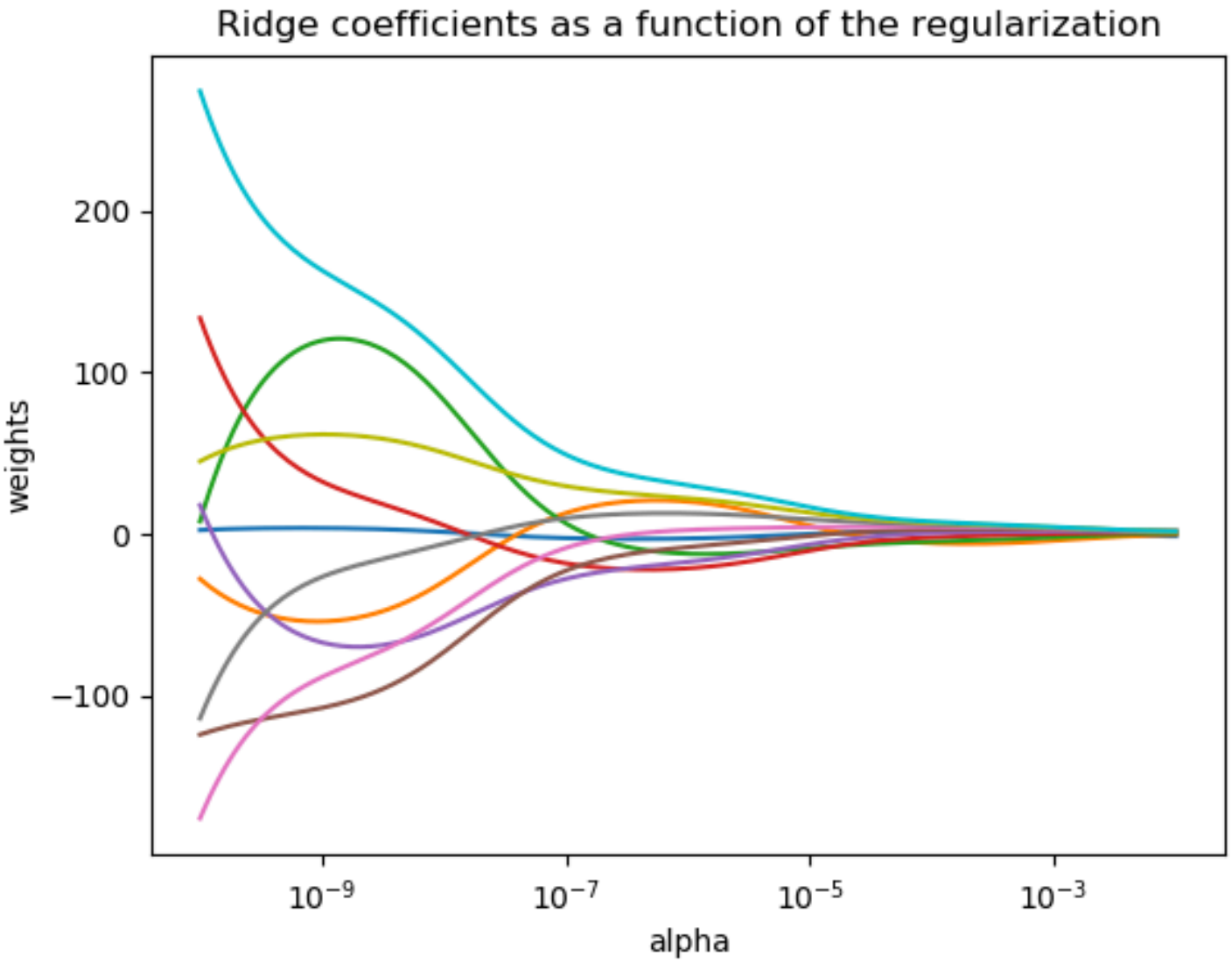
Ridge

$$L = \sum(\hat{Y}i - Yi)^2 + \lambda \sum \beta^2$$

Elastic Net

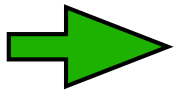
$$L = \sum(\hat{Y}i - Yi)^2 + \lambda \sum \beta^2 + \lambda \sum |\beta|$$

The complexity parameter $\alpha \geq 0$ controls the amount of shrinkage: the larger the value of α, the greater the amount of shrinkage and thus the coefficients become more robust to collinearity.

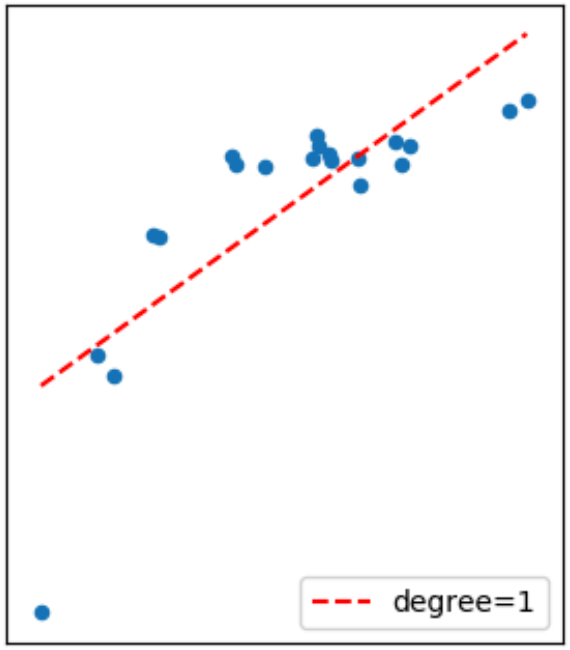


Reducing Underfitting with Polynomial regressions

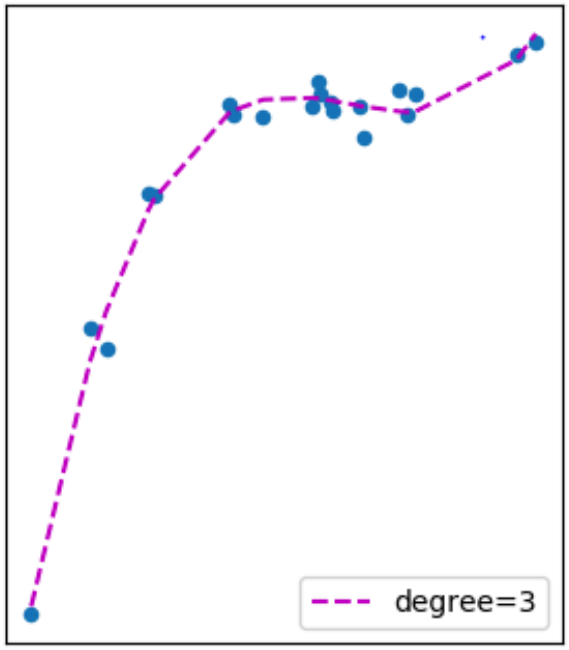
$$Y = \theta_0 + \theta_1 x$$



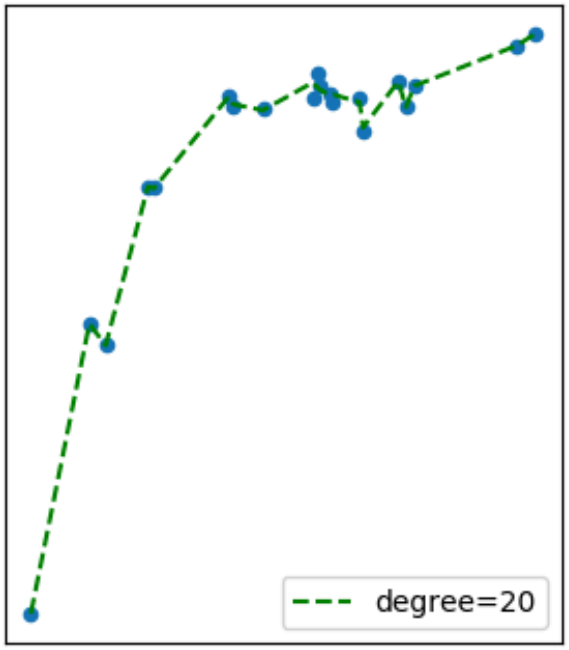
$$Y = \theta_0 + \theta_1 x + \theta_2 x^2$$



Underfit
High Bias
Low Variance



Correct Fit
Low Bias
Low Variance



Overfit
Low Bias
High Variance

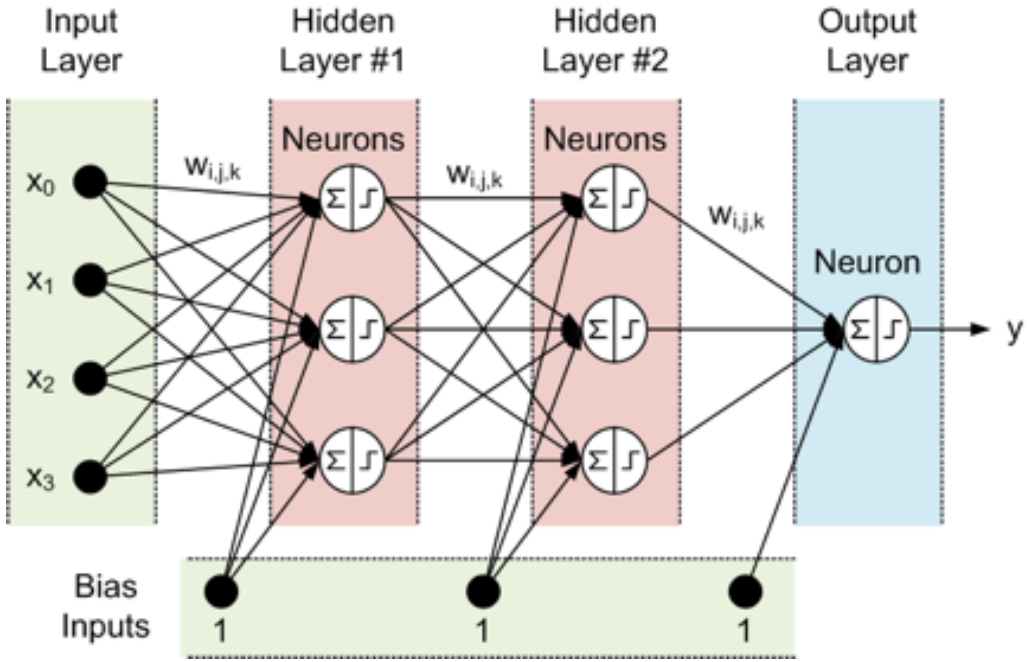
Neural Network Regression

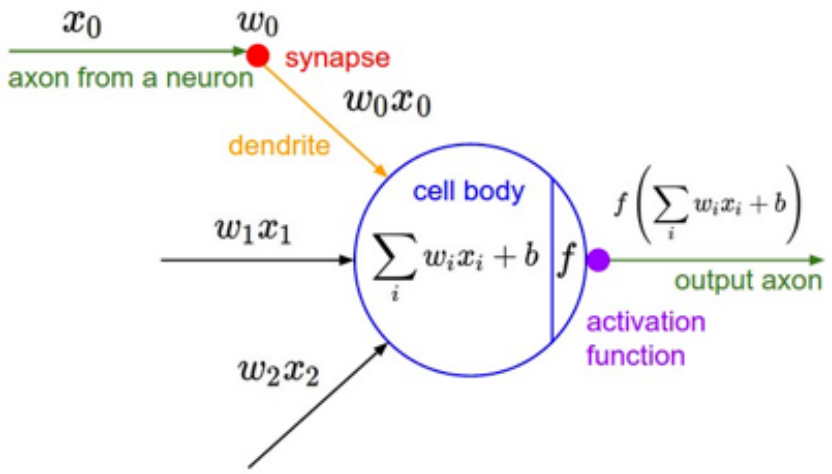
Input Layer contains cells that represent independent variables

Hidden Layer extracts the required features from the input data:

- Calculate the **weighted sum** of each cell in each layer.
- Rectify new values with **activation function**.

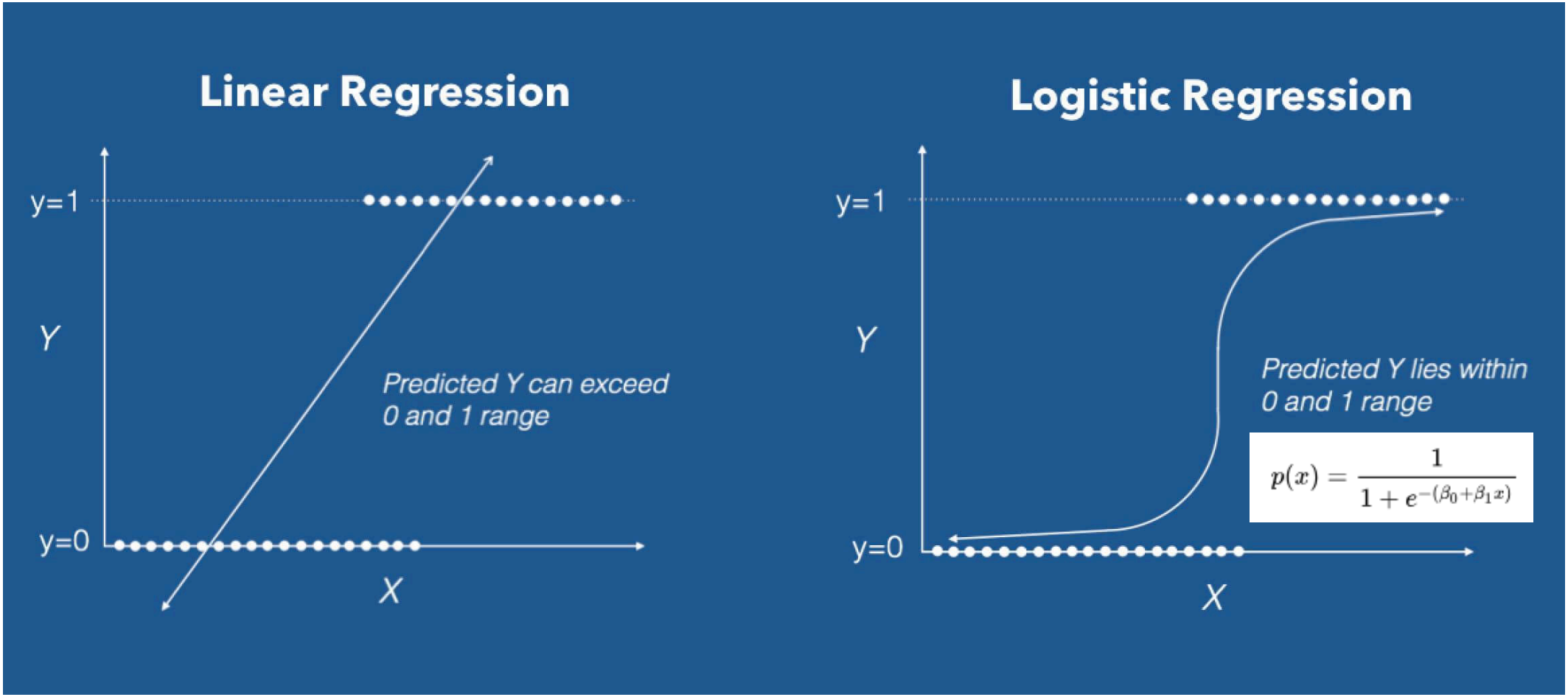
Output Layer is directly related to the type of work (classification or regression)





Logistic Regression

Logistic Regression method is quite similar to Linear Regression. The term “Logistic” is taken from the **Logit function** that is used in this **method of binary classification**.



Spam Detection : Predicting if an email is Spam or not

Credit Card Fraud : Predicting if a given credit card transaction is fraud or not

Health : Predicting if a given mass of tissue is benign or malignant

Marketing : Predicting if a given user will buy an insurance product or not

Banking : Predicting if a customer will default on a loan.

Algorithms for regressions and/or classifications

1.1. Generalized Linear Models

- 1.1.1. Ordinary Least Squares
  - 1.1.1.1. Ordinary Least Squares Complexity
- 1.1.2. Ridge Regression
  - 1.1.2.1. Ridge Complexity
  - 1.1.2.2. Setting the regularization parameter: generalized Cross-Validation
- 1.1.3. Lasso
  - 1.1.3.1. Setting regularization parameter
    - 1.1.3.1.1. Using cross-validation
    - 1.1.3.1.2. Information-criteria based model selection
    - 1.1.3.1.3. Comparison with the regularization parameter of SVM
- 1.1.4. Multi-task Lasso
- 1.1.5. Elastic-Net
- 1.1.6. Multi-task Elastic-Net
- 1.1.7. Least Angle Regression
- 1.1.8. LARS Lasso
  - 1.1.8.1. Mathematical formulation
- 1.1.9. Orthogonal Matching Pursuit (OMP)
- 1.1.10. Bayesian Regression
  - 1.1.10.1. Bayesian Ridge Regression
  - 1.1.10.2. Automatic Relevance Determination - ARD
- 1.1.11. Logistic regression
- 1.1.12. Stochastic Gradient Descent - SGD
- 1.1.13. Perceptron
- 1.1.14. Passive Aggressive Algorithms
- 1.1.15. Robustness regression: outliers and modeling errors
  - 1.1.15.1. Different scenario and useful concepts
  - 1.1.15.2. RANSAC: RANdom SAmple Consensus
    - 1.1.15.2.1. Details of the algorithm
  - 1.1.15.3. Theil-Sen estimator: generalized-median-based estimator
    - 1.1.15.3.1. Theoretical considerations
  - 1.1.15.4. Huber Regression
  - 1.1.15.5. Notes
- 1.1.16. Polynomial regression: extending linear models with basis functions

1.2. Linear and Quadratic Discriminant Analysis

1.3. Kernel ridge regression

1.4. Support Vector Machines

1.5. Stochastic Gradient Descent

1.6. Nearest Neighbors

1.7. Gaussian Processes

1.8. Cross decomposition

1.9. Naive Bayes

1.10. Decision Trees

1.11. Ensemble methods

1.12. Multiclass and multilabel algorithms

1.17. Neural network models (supervised)

https://scikit-learn.org/stable/supervised_learning.html

References

1. An Introduction to Statistical Learning: with Applications in R. David M Diez, Christopher D Barr, Mine Çetinkaya-Rundel. OpenIntro Statistics: Third Edition
2. Applied Predictive Modeling (2013) Kuhn, Max, Johnson, Kjell
3. Clear Explanations Youtubers: Victor Lavrenko, edureka!, Luis Serrano
4. Data Science for Business – Foster Provost & Tom Fawcett (2013) O'REILLY

https://github.com/BarbaraDiazE/CABANA_CHEMOINFORMATICS