

An Information Age: Math and Technology of Data

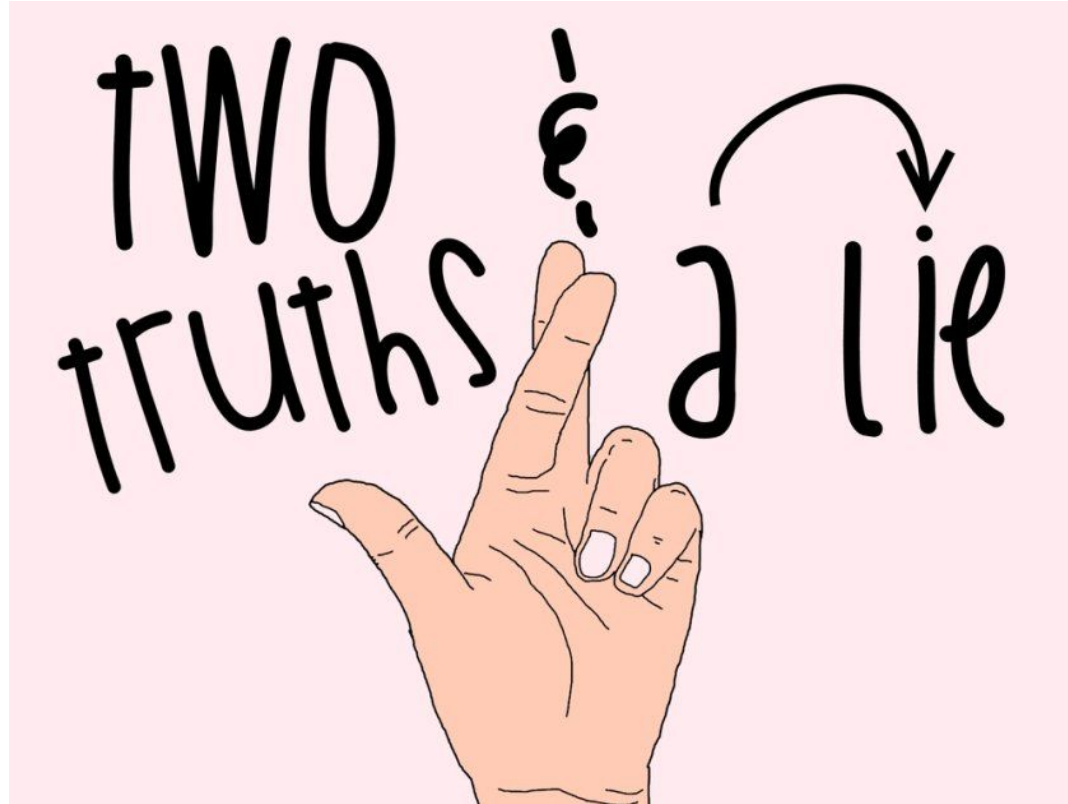
Duke TIP Scholar Weekends

Third Person Introductions

Introduce the person sitting next to you by saying the following.

- Name
- Hometown
- Favorite Food

Two Truths and a Lie



Rules and Expectations

What class rules are important to you?

Rules and Expectations

What are your expectations of the instructors?

Rules and Expectations

What are your expectations of each other?

Warm Up




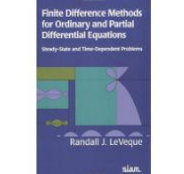







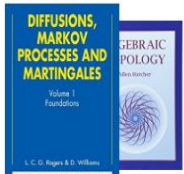
Sort yourselves in into 4 groups based on non-physical attributes.



Amazon

- Amazon tracks everything purchased from their site.
- This is a massive amount of data.
- Machine learning algorithms are used to find patterns in data.
- For example, amazon may recommends items purchased by other customers who have similar shopping habits to you.

Recommended for you, Sarah

 <p>Prime Video - Unlimited Streaming for Prime Members 28 ITEMS</p>	 <p>Golf Equipment 16 ITEMS</p>	 <p>Small Animal Supplies 6 ITEMS</p>	 <p>Reference Books 44 ITEMS</p>
 <p>Video Games 16 ITEMS</p>	 <p>Education & Teaching Books 54 ITEMS</p>	 <p>Engineering Books 44 ITEMS</p>	 <p>Dog Supplies 48 ITEMS</p>
 <p>Science & Math Books 100 ITEMS</p>	 <p>Computer & Technology Books 79 ITEMS</p>	 <p>Business Books 19 ITEMS</p>	 <p>Medical Books 5 ITEMS</p>

Face Detection



Netflix

- How can an algorithm use the data to recommend movies?
- By leveraging patterns in data.
- How would you do it?

The Netflix logo is displayed in a bold, white, sans-serif font with a black drop shadow, set against a solid red rectangular background.

Critic	Star Wars	Raiders of the Lost Arc	Casablanca	Singin' in the Rain
Ethan	****	****	*	**
Christopher	*****	****	**	*
Jimmy	**	**	****	***
Emmi	**	*	***	****
Sophia	*****	?	?	**

Preference Space

Star Wars



Raiders

Recommendations

- Sophia has not seen Raiders of the Lost Arc, but gave Star Wars 5 stars. Since people who like Star Wars also tend to like Raiders of the Lost Arc, we should recommend that Sophia watch Raiders of the Lost Arc.
- Sophia has not seen Casablanca, but gave Singin' in the Rain 2 stars. Since people who did not care for Singin' in the Rain also did not care for Casablanca, we should not recommend that Sophia watch Singin' in the Rain.

Features

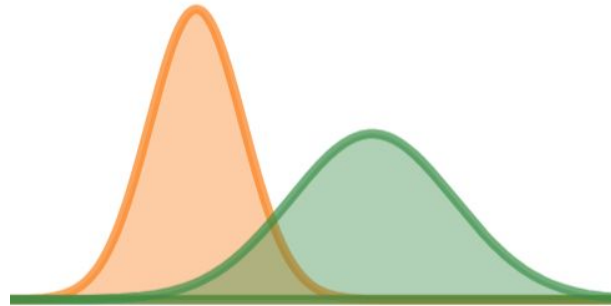
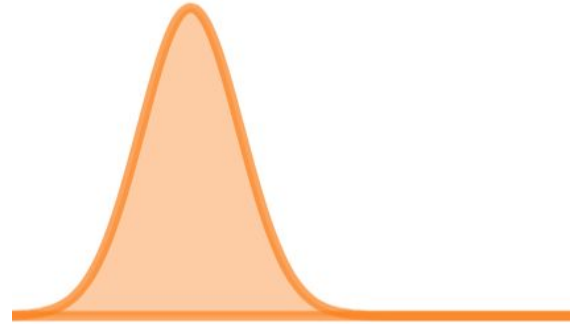
In order to compare critics, we must consider features. A **feature** is a data attribute used to make such a comparison. Features can be either quantitative (size, weight, etc.) or qualitative (color, shape, etc.).

Apples and Oranges

- Suppose we wanted to sort apples and oranges.
- We could use mass as a feature to tell apples and oranges apart.
- Is this a good feature?
- What about in conjunction with another feature such as color?



Apples and Oranges



Features

Think of an example of something you would want to sort.
What features would you use to sort these objects? Why did you choose those features?

Movie Features

Feature	Star Wars	Raiders of the Lost Arc	Casablanca	Singin' in the Rain

Movie Features

Feature	Star Wars	Raiders of the Lost Arc	Casablanca	Singin' in the Rain
Action (1-5)	5	4	2	1
Romance (1-5)	1	2	4	3
Length (min)	121	115	102	103
Harrison Ford	Y	Y	N	N
Year	1977	1981	1942	1952

Feature Space

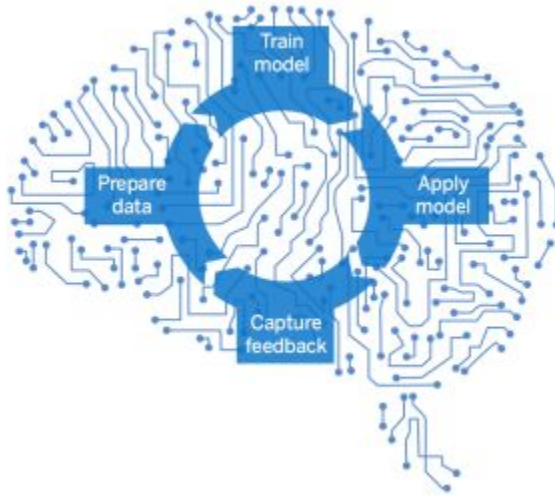


Machine Learning

Machine learning is a subfield of computer science. The objective is to teach a computer to solve problems without explicitly programming it to do so.



Problem Solving



**Problem
statement**



**Extract
features**



Implement



Evaluate

Supervised vs. Unsupervised

- **Supervised**

- Sample inputs with desired outputs are provided to the learner.
- The learner uses this information to determine the output for new inputs according to some algorithm.

- **Unsupervised**

- No labels are given to the learner.
- The learner must figure out how to structure inputs in some way.

Supervised or Unsupervised?

- **Classification**

- Data that include attributes of some object and a categorical label are provided.
- The goal is to use the attributes of an object to place it into the correct category.

- **Regression**

- Data that include attributes of some object and a continuous output are provided.
- The goal is to use the attributes of an object to determine the output value.

- **Clustering**

- Data that include attributes of some object are provided.
- The goal is to group the objects together that are similar.

Pitch Prediction

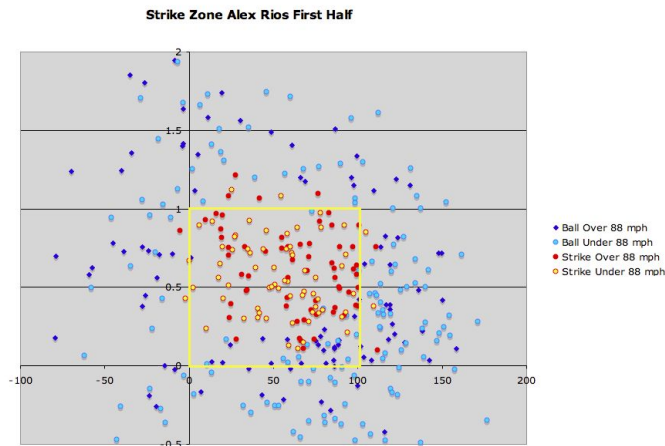
Data on every pitch in the MLB is collected.

- Speed
- Break
- Location
- Count
- Outs
- Score
- Runners
- Pitcher
- Batter

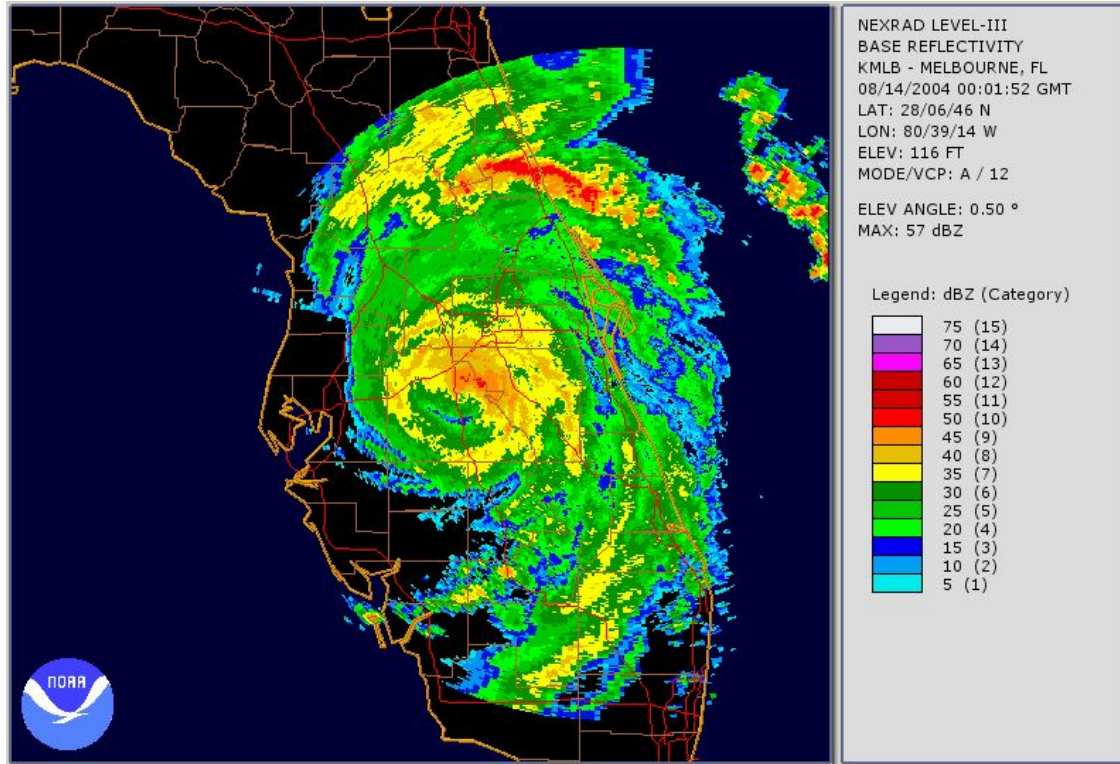
Use this information teaching a computer to classify the pitch.

- Fastball
- Curveball
- Knuckleball
- Change Up
- Slider

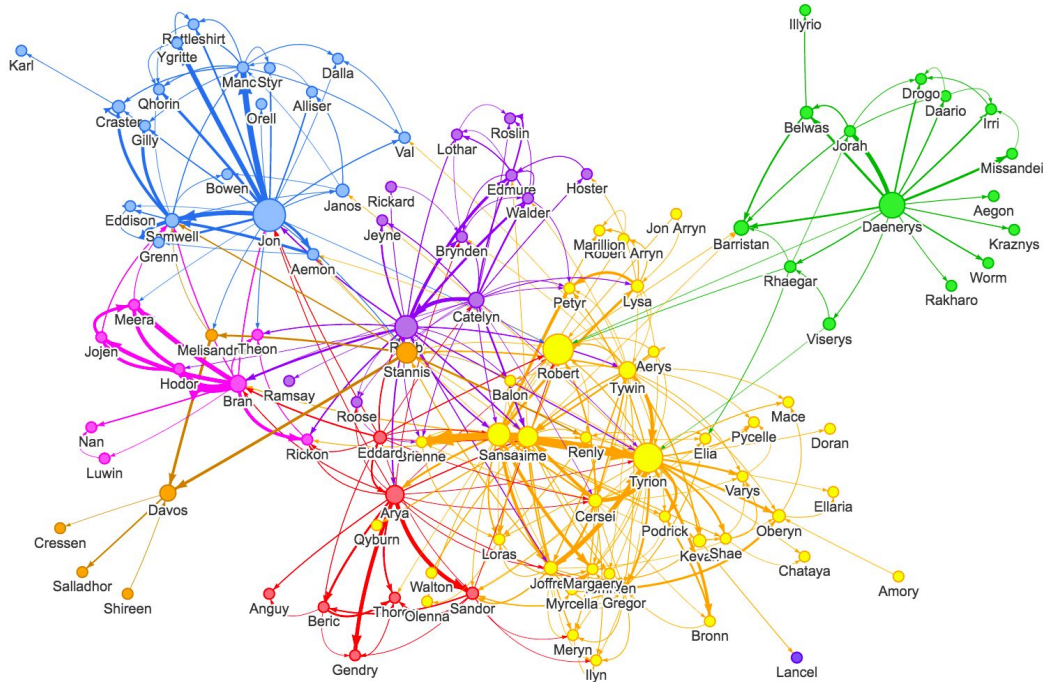
We can then predict what type of pitch will be thrown next in real time.



Predicting Hurricanes



Social Network Clustering



Testing Supervised Learning Algorithms

Suppose we have n data points. We need to use this data to develop and test our algorithm. This is done by partitioning the data.

- Approximately 80% of the data should be used to train our algorithm. This is called **training data**.
- Approximately 20% of the data should be used to validate or test our algorithm. This is called **validation data**.

Popular Machine Learning Algorithms

- *K*-Nearest Neighbors
- Classification Trees
- *K*-Means
- Linear Regression
- Logistic Regression
- Gaussian Naive Bayes
- Support Vector Machines



Popular Machine Learning Algorithms

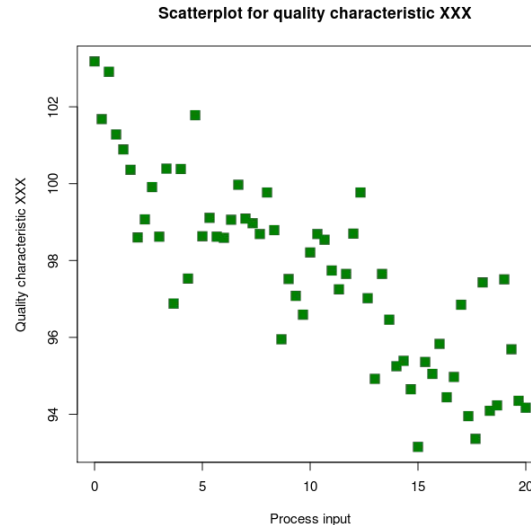
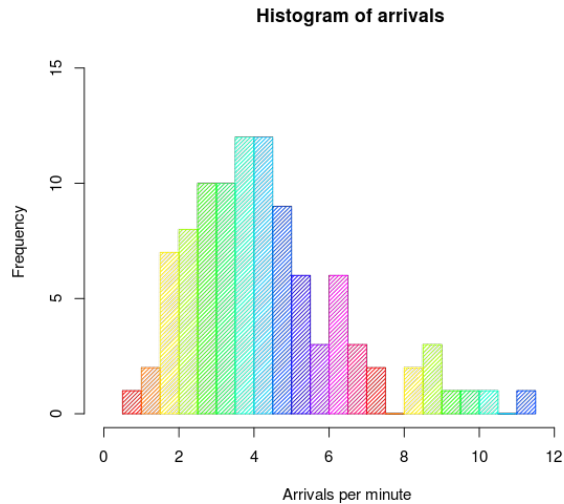
- *K*-Nearest Neighbors
- Classification Trees
- *K*-Means
- Linear Regression
- Logistic Regression
- Gaussian Naive Bayes
- Support Vector Machines



Descriptive Statistics

Descriptive Statistics

Before we begin to perform any analysis on our data, we should describe it with quantities and graphs.



Measures of Center

- The **mean**, μ , is the average of the data values.
- The **median**, Q_2 , is the midpoint of the data values.
- The **mode** is the most frequently occurring data value.

Percentiles

A **percentile** indicates the value below which a given percentage of observations in a group of observations. For example, if you were to score in the 95th percentile, you scored higher than 95% of the population.

Percentiles

- Some percentiles are quite common, such as the 25th percentile and the 75th percentile.
- We call the 25th percentile the **lower (first) quartile** (Q_1) and the 75th percentile the **upper (third) quartile** (Q_3).
- Note that the median is actually the 50th percentile.

Five Number Summary

- Minimum
- First Quartile
- Median
- Third Quartile
- Maximum

Measures of Spread

- The **variance**, σ^2 , is a measure of how spread data is from the average value.
- The **standard deviation**, σ , is simply the square root of variance.
- The **range** is the difference between the maximum and minimum values of the data set.
- The **interquartile range (IQR)** is the difference between Q_3 and Q_1 .

Variance and Standard Deviation

$$\sigma^2 = \frac{\sum_{i=1}^N (X - \mu)^2}{N}$$

Find the variance and standard deviation of the following data set.

2, 3, 5, 7, 11

Variance and Standard Deviation

$$\mu = (2 + 3 + 5 + 7 + 11)/5 = 5.6$$

$$\sigma^2 = ((2 - 5.6)^2 + (3 - 5.6)^2 + (5 - 5.6)^2 + (7 - 5.6)^2 + (11 - 5.6)^2)/5 = 12.8$$

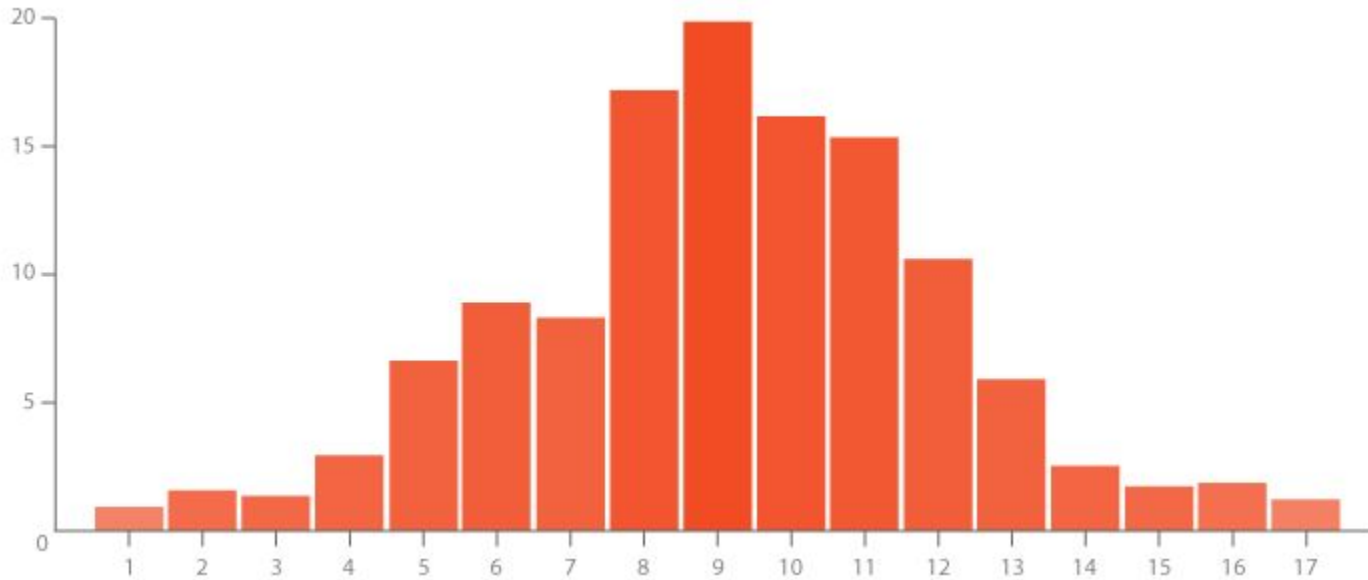
$$\sigma = (12.8)^{1/2} \approx 3.58$$

Visualizing Data

We can visualize data using the following plots.

- Histogram (Univariate)
- Box and Whisker (Univariate)
- Scatterplot (Bivariate)

Histograms

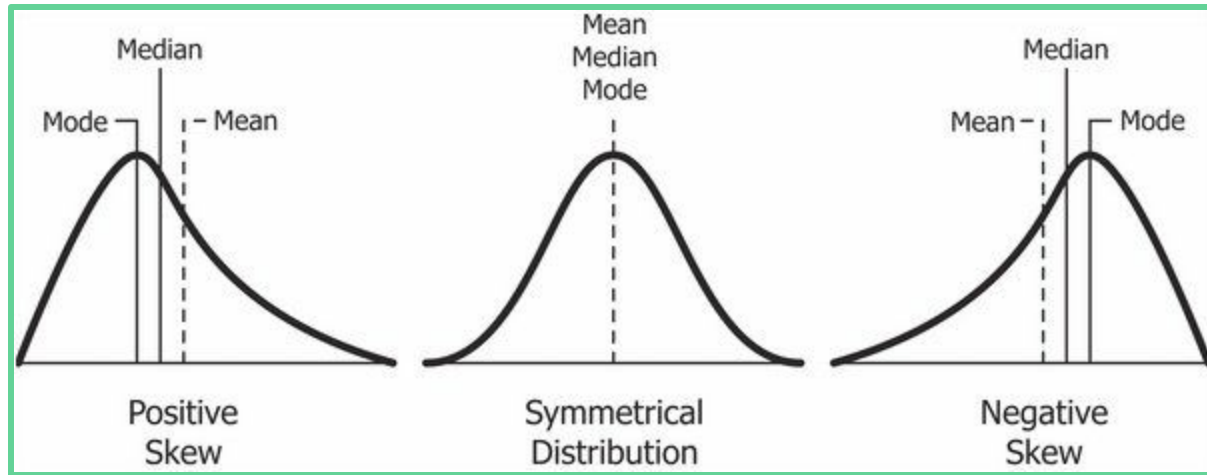


Histograms

What descriptive statistics can a histogram tell us?

Skewness

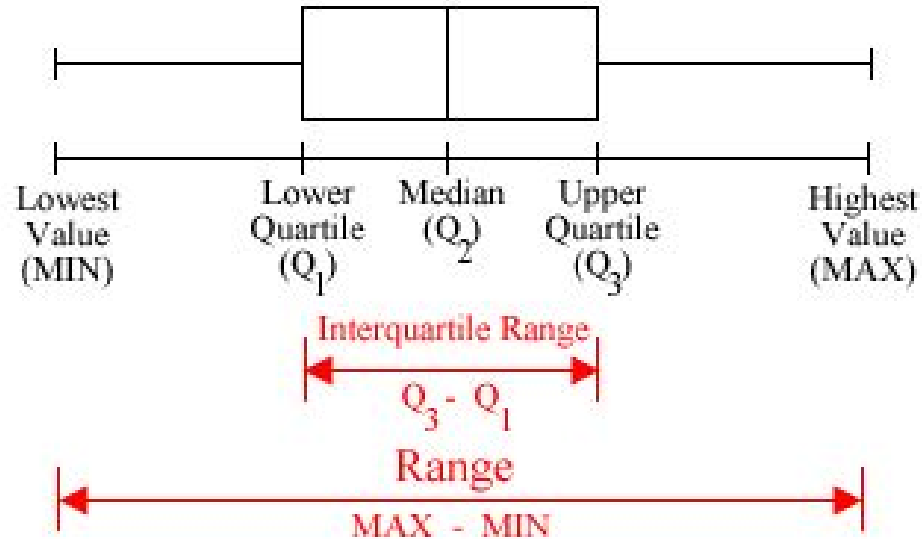
Not all histograms are **symmetric**. Rather, some are skewed to the **right (positive skew)** or **left (negative skew)**.



Skewness

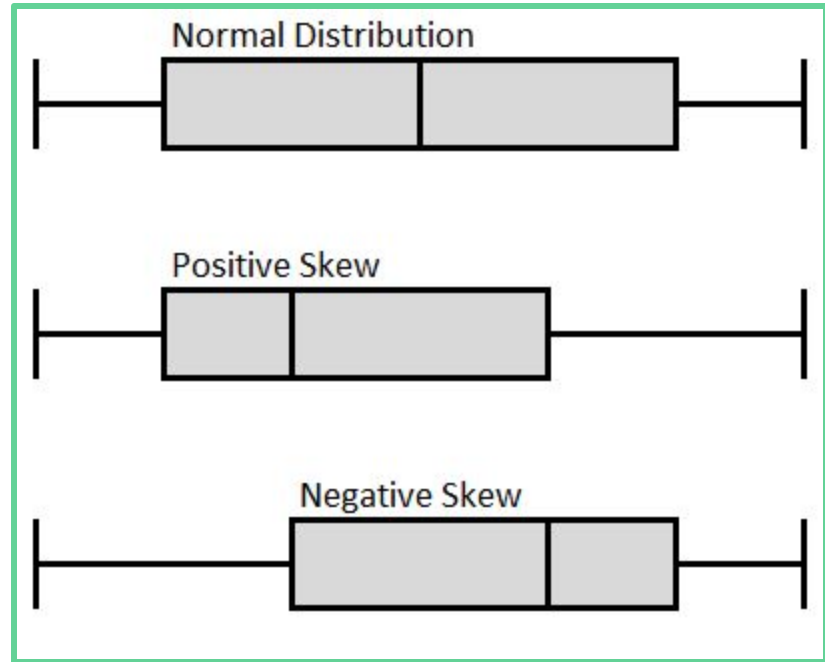
Can you think of data that would be symmetric, skewed to the right, or skewed to the left?

Box and Whisker Plots

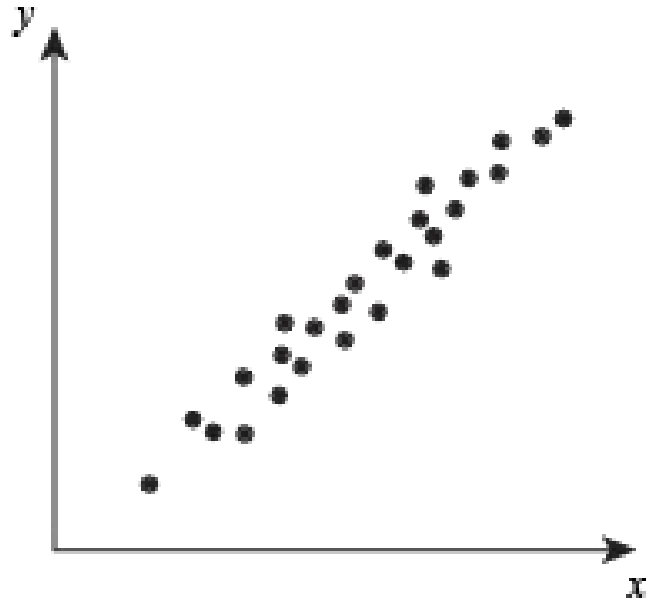


Skewness

Just like histograms, boxplots can be **symmetric** or skewed to the **right (positive skew)** or **left (negative skew)**.



Scatterplots



Scatterplots

Think of two variables that may have some relationship and sketch a scatterplot of what you think the data would look like.

Descriptive Statistics in Spreadsheets



Descriptive Statistics

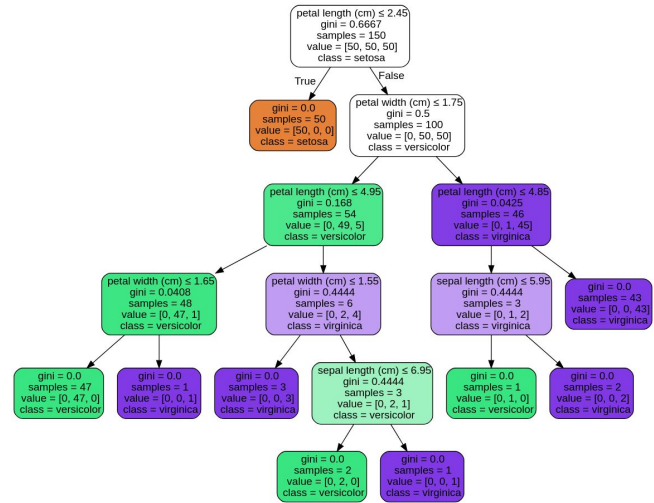
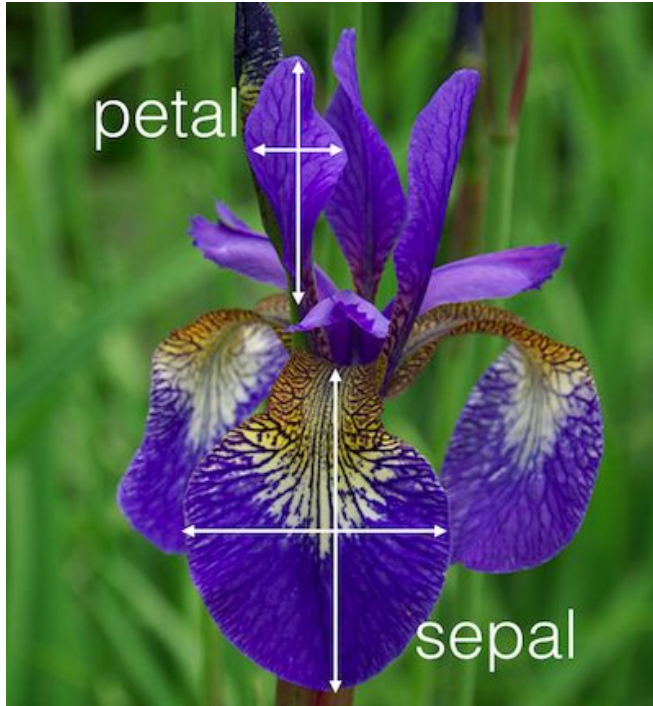
Find a data set online and analyze it by providing the following. Is the data skewed? What can you infer from this data?

- Mean
- Median
- Mode
- Minimum
- First Quartile
- Third Quartile
- Maximum
- Variance
- Standard Deviation
- Range
- IQR

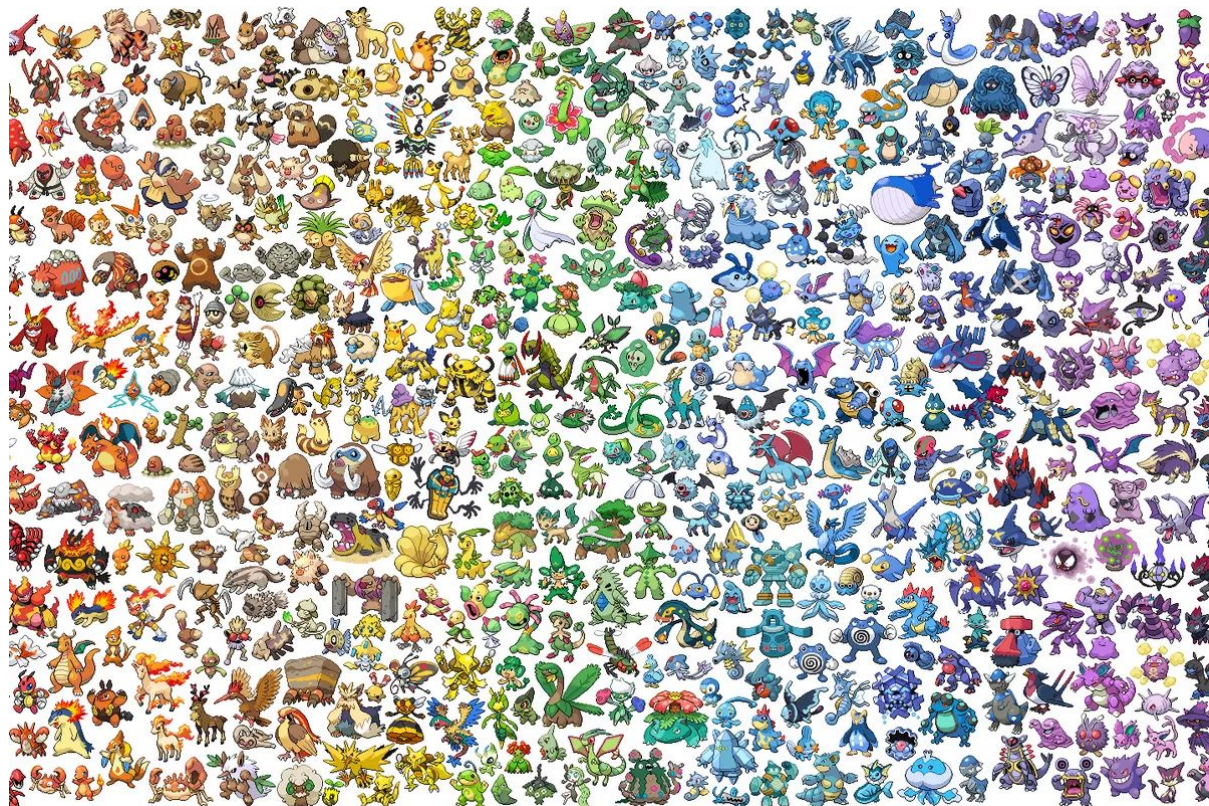
Descriptive Statistics in Orange



Iris Data



Pokemon Data



Descriptive Statistics

Using the dataset you found online create the following using Orange.

- Histogram
- Box Plot
- Scatterplot

Do you have any new insights from these plots?

Measuring Model Performance

Accuracy vs. Precision



**High Accuracy
High Precision**



**Low Accuracy
High Precision**



**High Accuracy
Low Precision**



**Low Accuracy
Low Precision**

Terminology

- **Positive (P)**: Observation is positive.
- **Negative (N)**: Observation is not positive.
- **True Positive (TP)**: Observation is positive, and is predicted to be positive.
- **False Negative (FN)**: Observation is positive, but is predicted negative.
- **True Negative (TN)**: Observation is negative, and is predicted to be negative.
- **False Positive (FP)**: Observation is negative, but is predicted positive.

Terminology

		Reality	
		Positive	Negative
Prediction	Positive	True Positive Correct!	False Positive Incorrect!
	Negative	False Negative incorrect!	True Negative Correct!

False Positives and False Negatives



Measures of Performance

- **Error:** What percent of your predictions were incorrect?
- **Accuracy:** What percent of your predictions were correct?
- **Precision:** What percent of positive predictions were correct?
- **Recall:** What percent of positive cases did you catch?

Can you come up with a formula for each in terms of false positive, false negative, true positive, and true negative?

Error

Error (E) is the proportion of all predictions that are incorrect.

The formula for error is given by

$$E = (FP + FN) / (FP + FN + TP + TN).$$

Accuracy

Accuracy (A) is the proportion of all predictions that are correct. The formula for accuracy is given by

$$A = (TP + TN)/(FP + FN + TP + TN).$$

Precision

Precision (P) is the proportion of all positive predictions that are correct.

$$P = (TP)/(TP + FP).$$

Recall

Recall (R) is the proportion of all positive observations that are predicted correctly.

$$R = (TP)/(TP + FN).$$

F_1 Score

The **F_1 score** is the harmonic mean of precision and recall, giving us an average of both precision and recall. The formula for F_1 is given by

$$F_1 = 2PR/(P + R).$$

We use the harmonic mean since it is more robust to extreme values.

Arithmetic vs. Harmonic Mean

Find the arithmetic and harmonic mean of the numbers 1, 2, 3, 4, 5, 100 using the formulas below.

$$A = \frac{1}{n} \sum_{i=1}^n a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} = \left(\frac{\sum_{i=1}^n x_i^{-1}}{n} \right)^{-1}.$$

Arithmetic vs. Harmonic Mean

In this case, using the harmonic mean provides a better estimate of the center of the data.

$$A = (1 + 2 + 3 + 4 + 5 + 100)/6 \approx 19.17$$

$$H = 6/(1 + 1/2 + 1/3 + 1/4 + 1/5 + 1/100) \approx 2.62$$

Error, Accuracy, Precision, and Recall

Compute the error, accuracy, precision, recall, and F_1 using the table below.

	Positive Cases	Negative Cases
Predicted Positive	60	140
Predicted Negative	40	9760

Error, Accuracy, Precision, and Recall

$$A = (9760 + 60)/(9760 + 140 + 40 + 60) \approx .98$$

$$E = (40+140)/(9760+140+40+60) \approx .02$$

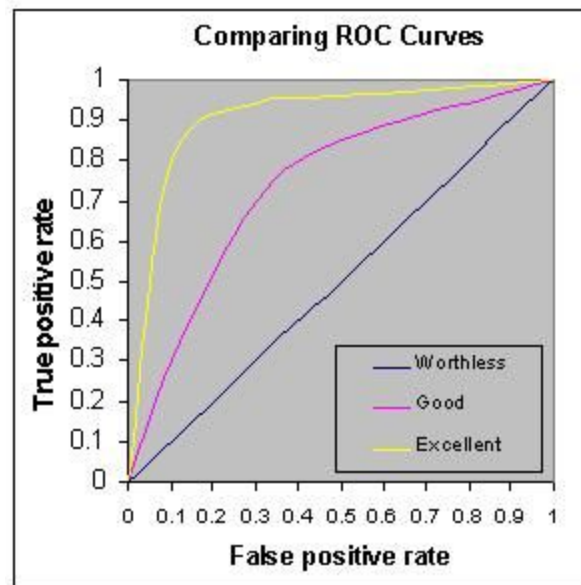
$$P = 60/(60+140) = .30$$

$$R = 60/(40+60) = .60$$

$$F1 = 2(.3)(.6)/(.3 + .6) = .40$$

Area Under ROC Curve

The **area under the ROC curve** is another way to measure model performance. This quantity ranges between 0 and 1 , where values less than $.50$ represents poor model performance and values close to 1 represent excellent model performance.



Matrices and Vectors

Matrices

A **matrix** is a rectangular array of numbers. We denote a matrix with n rows and m columns as an n by m matrix. We typically write matrices as capital letters, say A .

Vectors

A **vector** is a matrix with either one row one column. Vectors with one row are called **row vectors** and vectors with one column are called **column vectors**.

Adding and Subtracting Matrices

- Suppose A and B are both m by n matrices.
- Then $A + B$ is the m by n matrix whose entries are the sums of the corresponding entries of A and B .
- Then $A - B$ is the m by n matrix whose entries are the differences of the corresponding entries of A and B .
- Since vectors are just special type of matrices, we can add and subtract vectors the same way.

Representing Data as a Vector

Going back to our movie example, each movie can be represented as a vector in the form of

(Action (1-5), Romance (1-5), Length, Harrison Ford (Y or N),
Year).

Here we will code a Y as a *1* and a N as a *0*.

Representing Data as a Vector

So we have

$$SW = (5, 1, 121, 1, 1977)$$

$$RLA = (4, 2, 115, 1, 1981)$$

$$C = (2, 4, 102, 0, 1924)$$

$$SR = (1, 3, 103, 0, 1952)$$

Categorical Data

In the green apples and orange oranges example, we looked at mass and color. However, what if we also wanted to sort red apples and purple plums. How would we code these colors?

What if we included the state where the fruit was grown as a feature. How would we code locations?

Norms and Distance

Norms

The **norm** of a vector is a quantity used to describe its magnitude. We will discuss the following three norms.

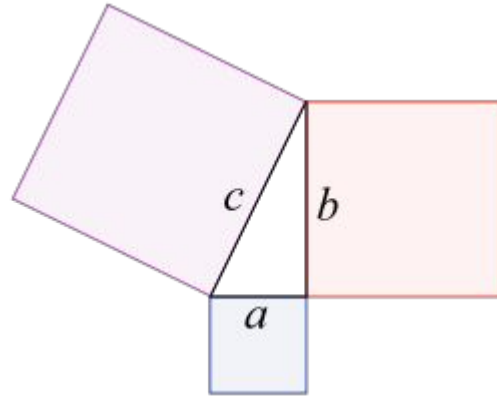
- Euclidean Norm: 2-Norm
- Taxicab (Manhattan) Norm: 1-Norm
- Maximum (Chebyshev) Norm: ∞ -Norm

Euclidean Norm

The **Euclidean Norm** is the ordinary distance from the origin to the point $x = (x_1, x_2, \dots, x_n)$.

Find the Euclidean Norm of

$SW = (5, 1, 121, 1, 1977)$.

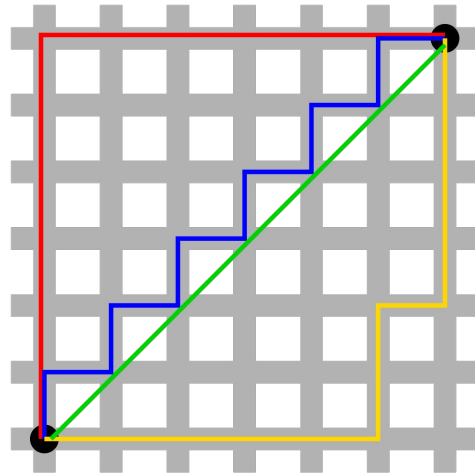


Taxicab Norm

The **Taxicab Norm** is the sum of the absolute values of the entries of $x = (x_1, x_2, \dots, x_n)$.

Find the Taxicab Norm of

$$SW = (5, 1, 121, 1, 1977).$$



Maximum Norm

The **Maximum Norm** is the maximum of the absolute values of the entries of $x = (x_1, x_2, \dots, x_n)$.

Find the Infinity Norm of $SW = (5, 1, 121, 1, 1977)$.

Norms of Star Wars

$$\|SW\|_2 = (5^2 + 1^2 + 121^2 + 1^2 + 1977^2)^{1/2} \approx 145.12$$

$$\|SW\|_1 = |5| + |1| + |121| + |1| + |1977| = 2105$$

$$\|SW\|_\infty = 1977$$

Feature Selection

How does feature selection affect the magnitude of the norm?
Would it be better to remove a feature? How else could we deal with this?

Feature Selection

Suppose we change movie length to hours rounded down to the nearest hour and code year in terms of the decade (for example 1970s are assigned 7). Then we have the following vectors.

$$SW = (5, 1, 2, 1, 7)$$

$$RLA = (4, 2, 2, 1, 8)$$

$$C = (2, 4, 1, 0, 2)$$

$$SR = (1, 3, 1, 0, 5)$$

Computing Distance Using Norms

To find the distance between two vectors,

$$d(v_1, v_2) = \|v_1 - v_2\|$$

Any distance has the following properties.

1. $d(v_1, v_2) \geq 0$
2. $d(v_1, v_2) = 0 \Leftrightarrow v_1 = v_2$
3. $d(v_1, v_2) = d(v_2, v_1)$
4. $d(v_1, v_3) \leq d(v_1, v_2) + d(v_2, v_3)$

Do our three norms satisfy these properties?

Computing Distance Using Norms

Find the distance between the the following vectors using all three norms.

1. $SW = (5, 1, 2, 1, 7)$ and $RLA = (4, 2, 2, 1, 8)$
2. $C = (2, 4, 1, 0, 2)$ and $SR = (1, 3, 1, 0, 5)$
3. $RLA = (4, 2, 2, 1, 8)$ and $C = (2, 4, 1, 0, 2)$

Star Wars vs. Raiders of the Lost Arc

$$SW - RLA = (1, -1, 0, 0, -1)$$

$$\|SW - RLA\|_2 = (1^2 + (-1)^2 + 0^2 + 0^2 + (-1)^2)^{1/2} \approx 1.7$$

$$\|SW - RLA\|_1 = |1| + |-1| + |0| + |0| + |-1| = 3$$

$$\|SW - RLA\|_\infty = 1$$

Casablanca vs. Singin' in the Rain

$$C - SR = (1, 1, 0, 0, -3)$$

$$\|C - SR\|_2 = (1^2 + 1^2 + 0^2 + 0^2 + (-3)^2)^{1/2} \approx 3.5$$

$$\|C - SR\|_1 = |1| + |1| + |0| + |0| + |-3| = 5$$

$$\|C - SR\|_\infty = 3$$

Raiders of the Lost Arc vs. Casablanca

$$RLA - C = (2, -2, 1, 1, 6)$$

$$\|SW - RLA\|_2 = (2^2 + (-2)^2 + 1^2 + 1^2 + 6^2)^{1/2} \approx 6.5$$

$$\|SW - RLA\|_1 = |2| + |-2| + |1| + |1| + |6| = 12$$

$$\|SW - RLA\|_\infty = 6$$

Computing Distance Using Norms

Find the distance between the the modified vectors using all three norms.

1. *SW* and *RLA*
2. *C* and *SR*
3. *RLA* and *C*

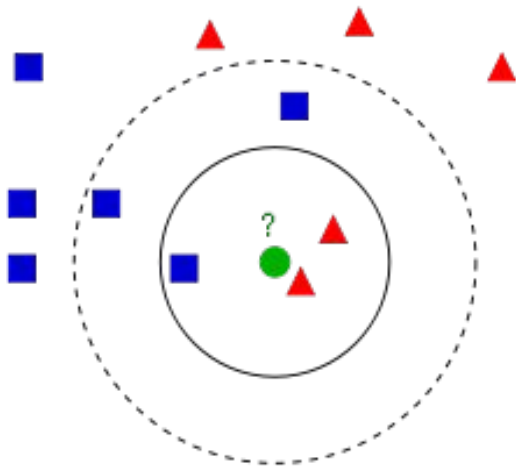
Which feature vectors gave better results?

K-Nearest Neighbor

K-Nearest Neighbors

Suppose $k = 3$. For each point, v , in the validation data

1. Find the 3 points (t_1, t_2, t_3) from the training data that are the closest to v .
2. Determine the most common label of the 3 points t_1, t_2, t_3 .
3. Label v with the most common label.



Football or Basketball?



K-Nearest Neighbors Practice

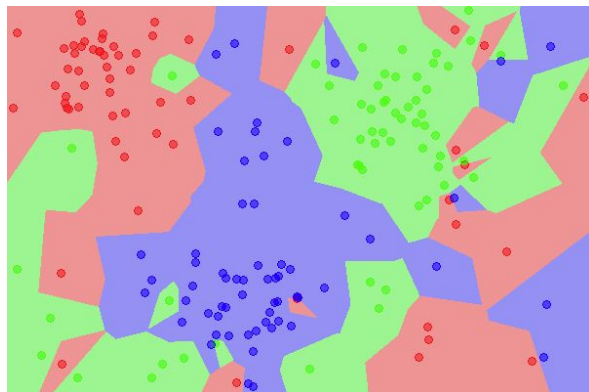
- Based on your results, compute the error, accuracy, precision, recall, and F_1 .
- Do you think your results be different for a different k ?
- Do you think your results would be different if the test data had more basketball players than football players (or vice versa)?
- Suppose we added one more basketball player. What would happen if we use $k = 21$?
- What would you be classified as based on your height and weight?

K-Nearest Neighbor in Orange



K-Nearest Neighbors Practice

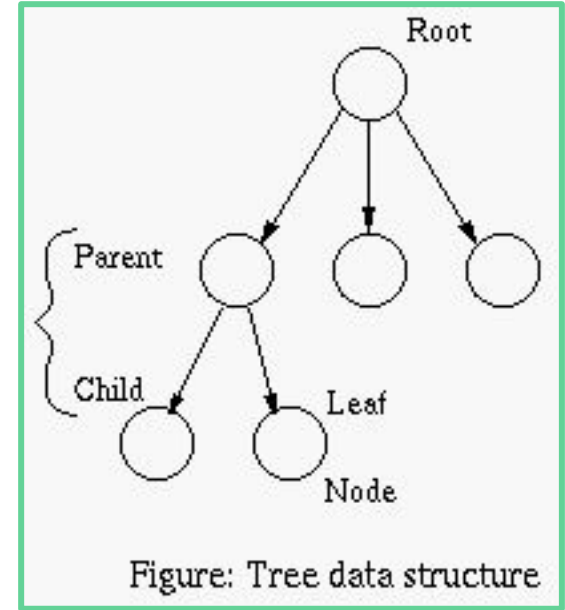
1. Find or create a data set.
2. Separate the data into training and validation sets.
3. Use the *k*-nearest neighbor algorithm to classify the validation data.
4. Calculate the accuracy of the algorithm.
5. Try multiple values of *k* and see how the algorithm performs.



Classification Trees

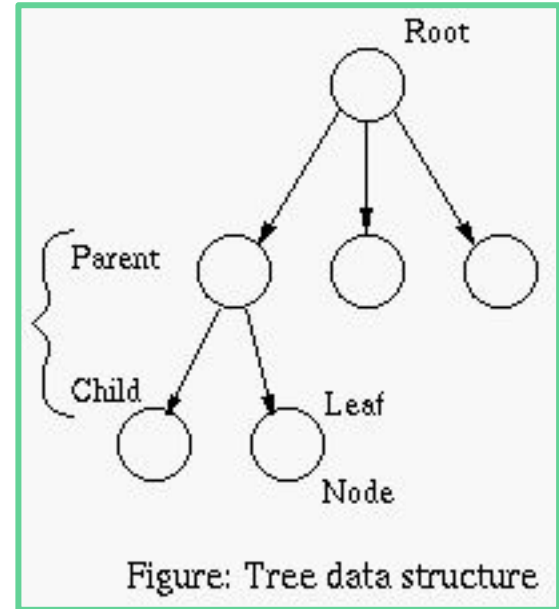
Trees

- A **tree** is an abstract data structure that consist of
 - **nodes**,
 - one directional **branches**,
 - no **cycles**, and
 - a single **root**.
- The top node of the tree is called the **root**. All nodes must be accessible from the root following the direction of the branches.

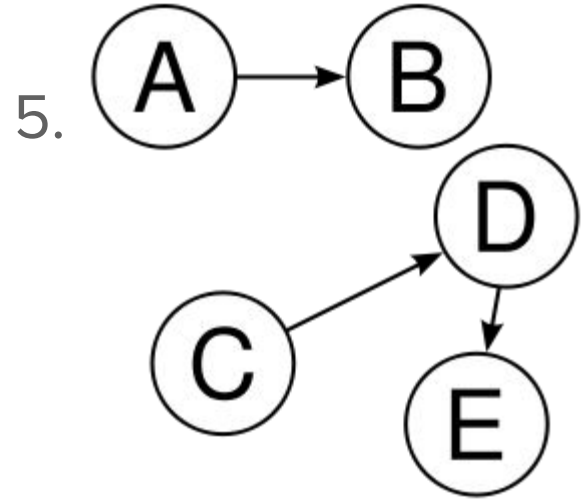
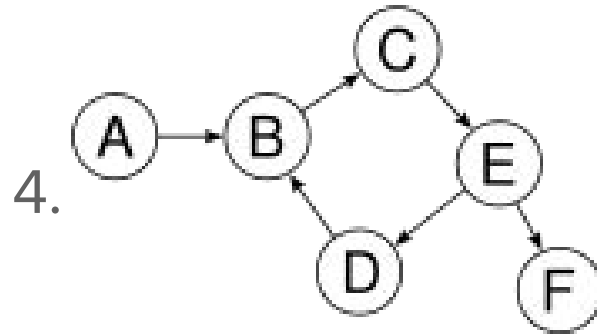
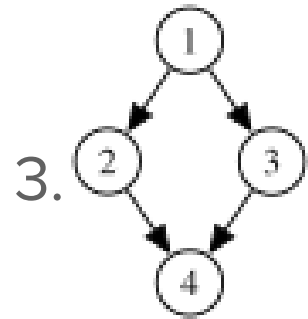
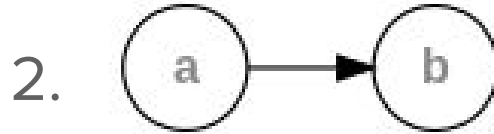


Trees

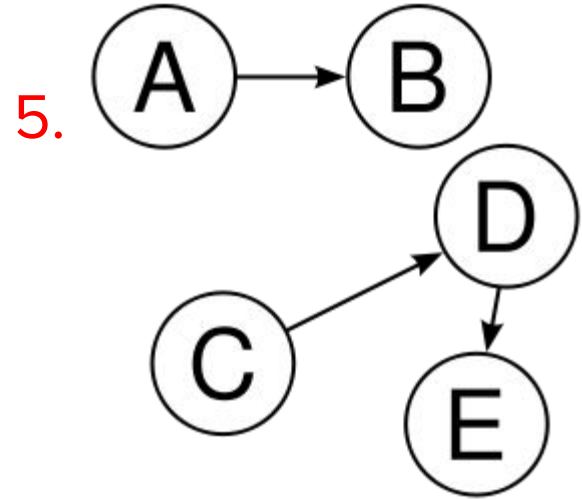
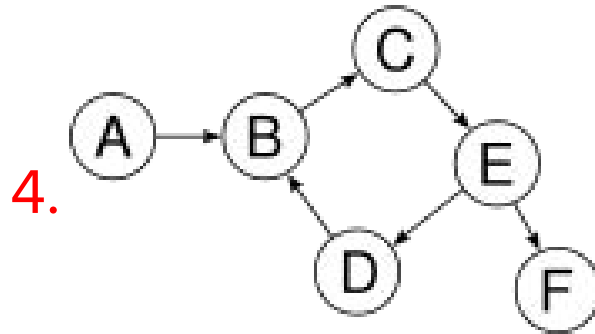
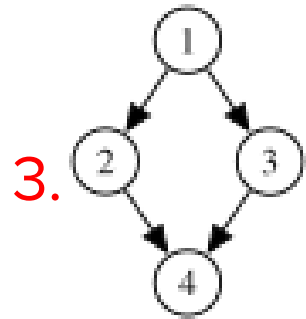
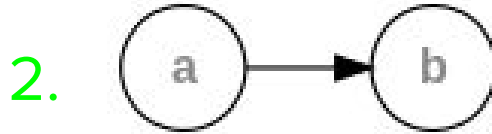
- A **parent** is connected to another node called a **child**. Each child can only have 1 parent.
- **Siblings** are nodes with the same parent.
- A **leaf** is a node with no children.
- The **height** of a tree is the number of branches on the longest path from leaf to root.
- A **forest** is a collection of disjoint trees.



To tree or not to tree? That is the question.

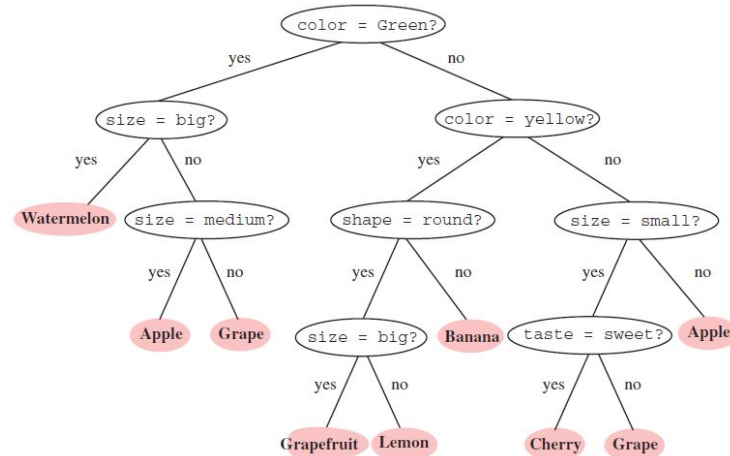
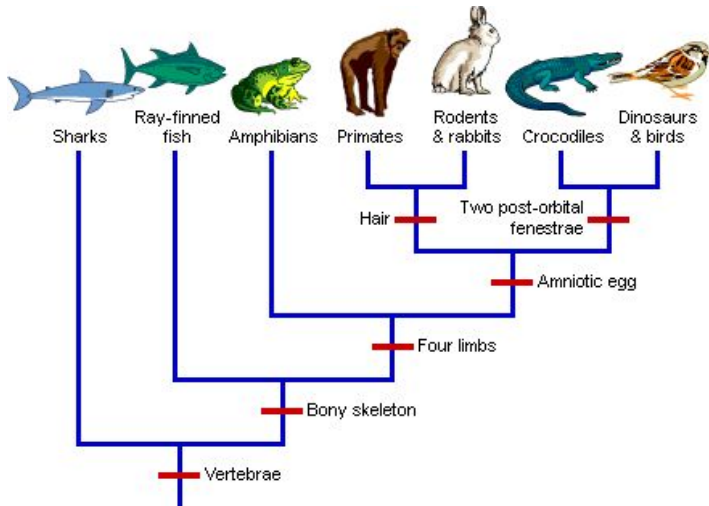


To tree or not to tree? That is the question.



Machine Learning Trees

The algorithm creates a **decision tree** using the training data. Each leaf should only contain data with the same label.



Validation data is then classified by following the tree.

Analyzing Decision Trees

Suppose we have 1024 data points that fall into 2 different categories.

- What is the tallest tree we could have?
- What is the shortest tree we could have?

Suppose we have 1024 data points that each belonged to a different category.

- What is the tallest tree we could have?
- What is the shortest tree we could have?

Analyzing Decision Trees

- If we have 1024 data points that fall into 2 different categories, the tallest tree would have height 1023 and the shortest tree would have height 1.
- If we have 1024 data points that each belonged to a different category the tallest tree would have height 1023 and the shortest tree would have height 10.

Classification Tree Practice

Create a classification tree that could be used to sort pennies, nickels, dimes, and quarters. Assume that these coins are worn and cannot be read. Once, finished, trade with a partner and check to see that their tree is accurate.



Robustness

- Would your method work on all data sets? An algorithm that works on a wide range of data is called **robust**.
- Computers only understand precise and exact statements. They do what you tell them to do, not what you meant for them to do. Could you teach your method to a computer?
- Refine your method so that it is robust and exact.

Teaching a Computer

Think of a simple task and write down precise instructions that a computer could follow.



Greedy Algorithm

1. For each attribute, determine the range that contains the **most** data points from a single category.
2. Find the attribute that separates the **most** points.
3. Use the range found in step 1 for the attribute from step 2 as the next branch for the tree.
4. Remove the data that is separated by your new branch.
5. Repeat the algorithm from step 1 with the remaining data.

Classification Trees in Orange

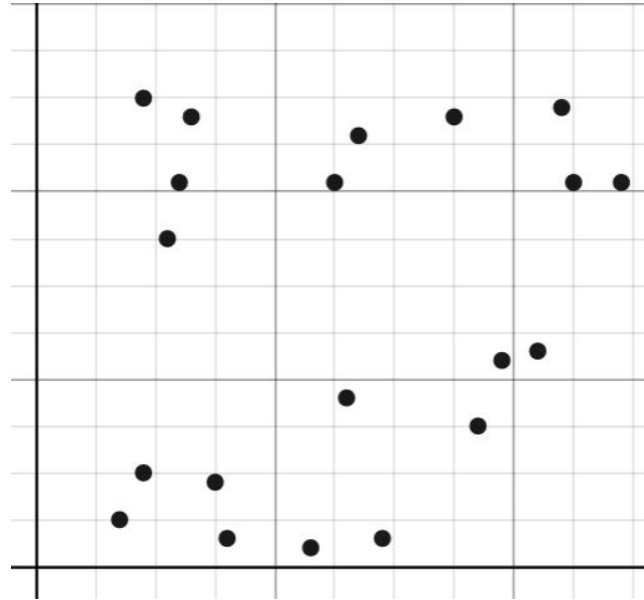
Make a tree by both by hand and using Orange.
Compare and contrast the two trees.



K-Means

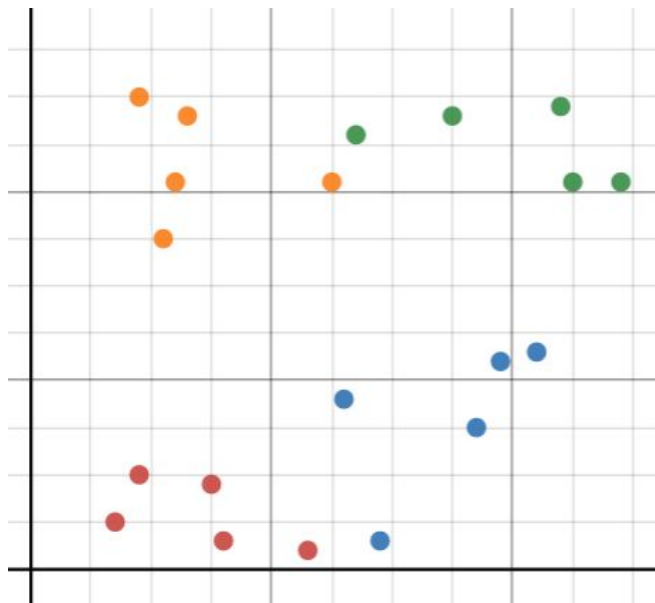
Clustering

The scatterplot shown to the right contains four groups. Sort these data into the four groups.



Clustering

- How did you do it?
- What if we had more data?
- What if we had more than two features?

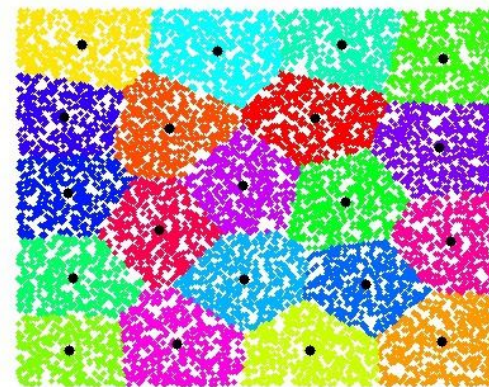


K-Means Algorithm

The ***k*-means algorithm** is an iterative method that automatically groups data into k clusters.

An **iterative method** is a repeated procedure that generates an improving approximate solution for some problem.

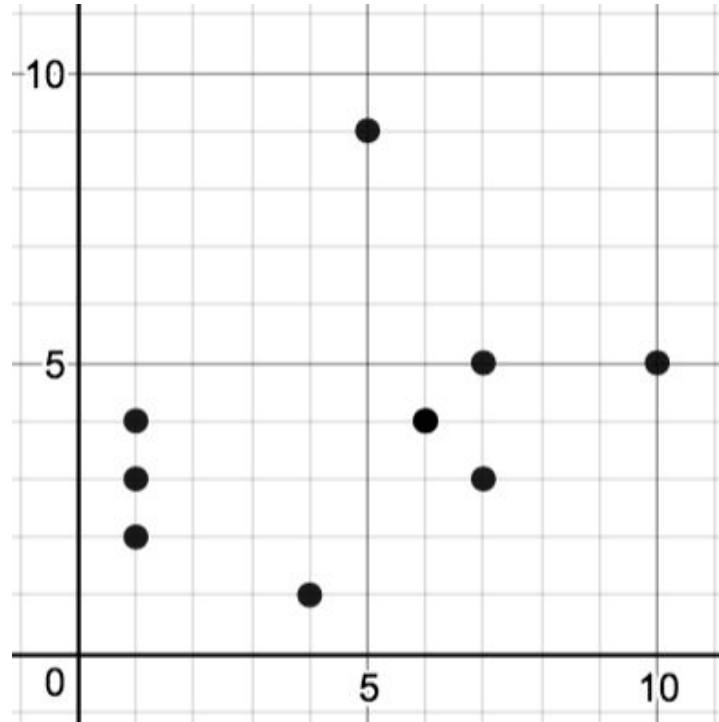
The **termination criterion** determines when the algorithm stops. This is often done when the approximate solution no longer changes or after a set number of iterations.



K-Means Algorithm

1. Choose k random starting points called **centroids**.
2. For each point, calculate the distance to each centroid.
3. Label each point according to the nearest centroid.
4. Recalculate the mean of the coordinates of points with the same label.
5. Move the centroid to the mean coordinates calculated above.
6. Repeat steps 1-5 until the centroid no longer moves.

K-Means Handout

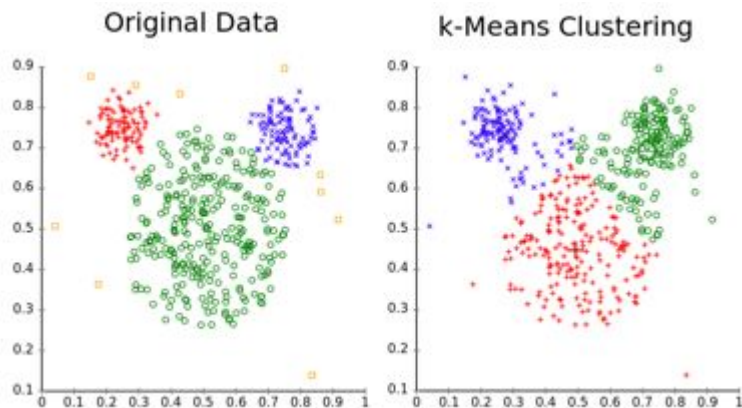


K-Means in Orange



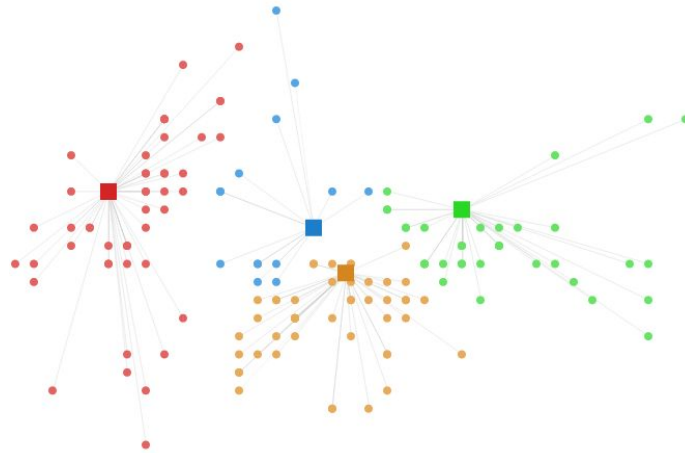
Caveats of *K*-Means Algorithm

- k must be known before hand.
- Clusters tend to be the same size.



Class Clustering

We want separate our class into four groups during breaks, based on their interests. As a class, determine who will be in each of the four groups.



Class Clustering

1. Determine two questions (1-5 scale) to ask each member of the class. The purpose of these questions is to separate students based on interests.
2. Enter this data into a spreadsheet.
3. Use the k -means algorithm to determine who is in each of the four groups.

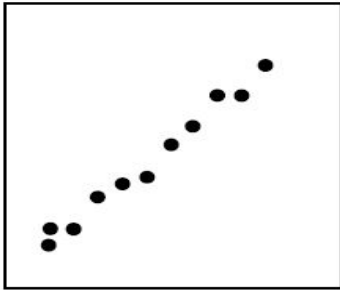
Linear Regression

Correlation

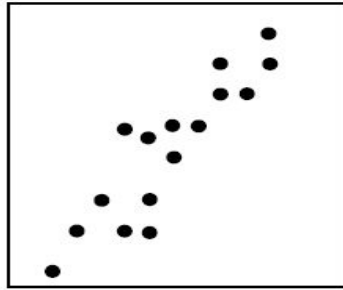
Correlation describes the strength (**strong**, **moderate**, or **weak**) and direction (**positive** or **negative**) of a linear relationship.

Correlation

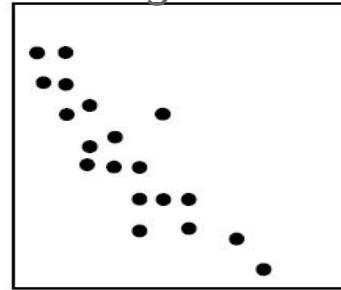
Strong Positive



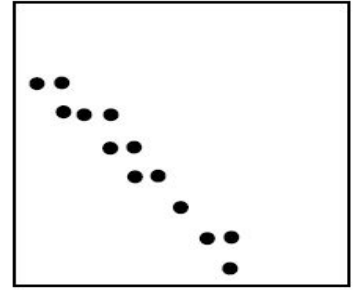
Moderate Positive



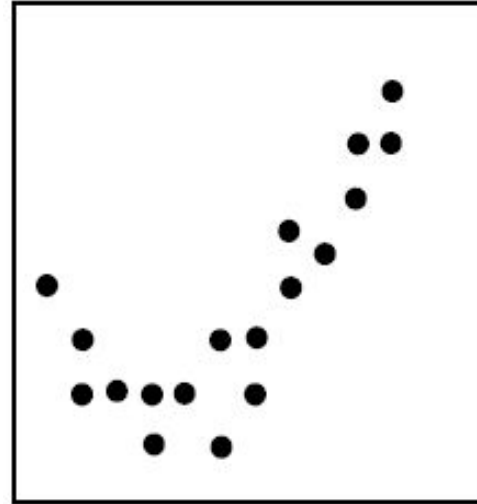
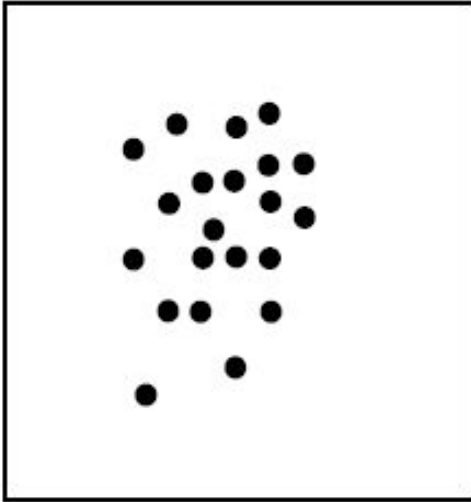
Moderate
Negative



Strong Negative



No Correlation



Correlation

The **correlation coefficient**, r , is a unitless value that quantifies the strength and direction of a linear relationship, given by the following formula.

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

Note that $-1 \leq r \leq 1$, where 1 or -1 represents a perfect relationship and 0 represents no relationship. Positive values of r represent a positive correlation, whereas negative values of r represent a negative correlation.

Correlation

Let's practice determining the value of r by looking at a scatterplot.



Spurious Correlations

Two variables can be strongly correlated even if they have nothing to do with one another. Correlation does not imply causation!

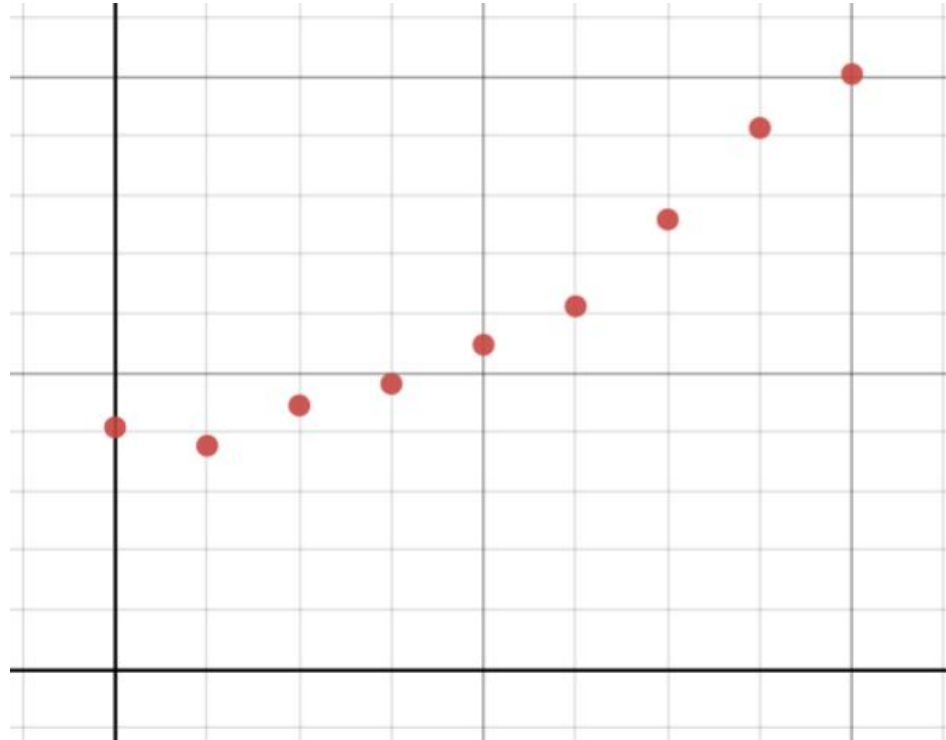
tylervigen.com

Line of Best Fit

Knowing the strength and direction of a relationship is useful, but we would also like to know the equation of the function that best models the data so that we can make predictions about future values.

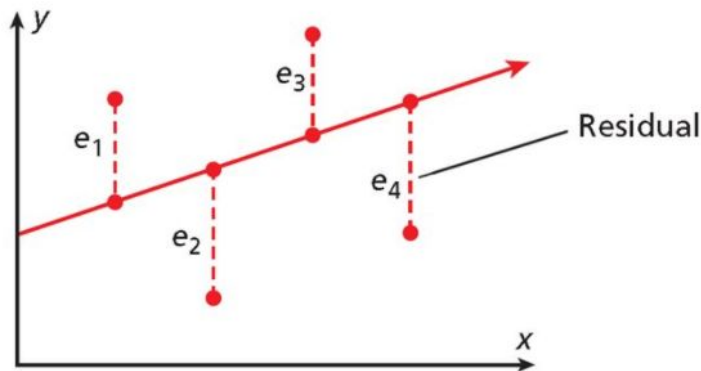
So, how do we determine which line fits the data the best?

Line of Best Fit



Residuals

A **residual**, or **error**, is the distance from a point to the model being used to fit the data. In other words, it is the difference between the observed value and the predicted value.



Line of Best Fit

The **line of best fit** is the line, $y = mx + b$, that minimizes the **sum of the squared residuals (RSS)**. That is, it minimizes the following quantity.

$$RSS = \sum_{i=1}^n (\epsilon_i)^2$$

Why do we square the residuals before summing them?

Model Fit

- Mean Squared Error (MSE)
- Root Mean Squared Error ($RMSE$)
- Mean Absolute Error (MAE)
- Coefficient of Determination (R^2)

Mean Squared Error

The **mean squared error (MSE)** is the average of the squared errors, given by the following formula. Note that this is just the mean of *RSS*.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$$

Root Mean Squared Error

The **root mean squared error (RMSE)** is simply the square root of MSE, which provides us with an estimate of, on average, how far the the model varies from the data.

Mean Absolute Error

The **mean absolute error (MAE)** is the average of the absolute value of residuals, given by

$$\text{MAE} = \frac{\sum_{i=1}^n |y_i - x_i|}{n} = \frac{\sum_{i=1}^n |e_i|}{n}.$$

Total Sum of Squares

The **total sum of squares (TSS)** is the sum of squared differences between each observed value and the mean, given by the following quantity.

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2$$

TSS represents the total variation in the dependent variable.

Model Sum of Squares

The **explained sum of squares (ESS)** is the sum of squared differences between each predicted value and the mean, given by

$$\text{ESS} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

ESS represents the variation in the data that can be explained by the model.

Coefficient of Determination

- The **coefficient of determination**, R^2 , is the proportion of variation that be explained by the model, given by

$$R^2 = ESS/TSS.$$

- $0 \leq R^2 \leq 1$, where 0 represents no fit and 1 represents a perfect fit.
- R^2 can be used to determine goodness of fit for any model, including linear models. For linear models only, $R^2 = r^2$.

Multiple Linear Regression

Instead of using a single predictor, it is likely that we would want to use several features to predict an outcome. Therefore, we will use the multiple linear regression model shown below.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

Multiple Linear Regression

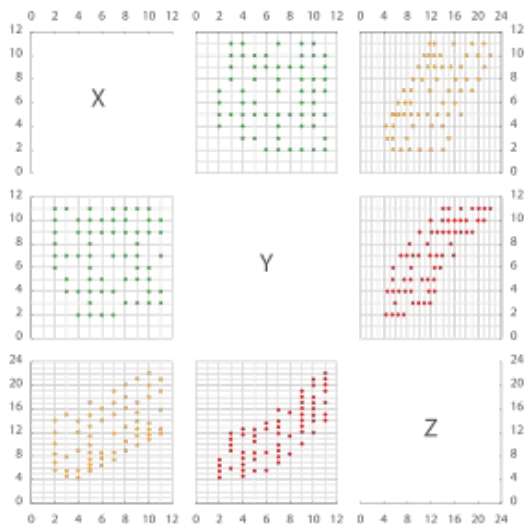
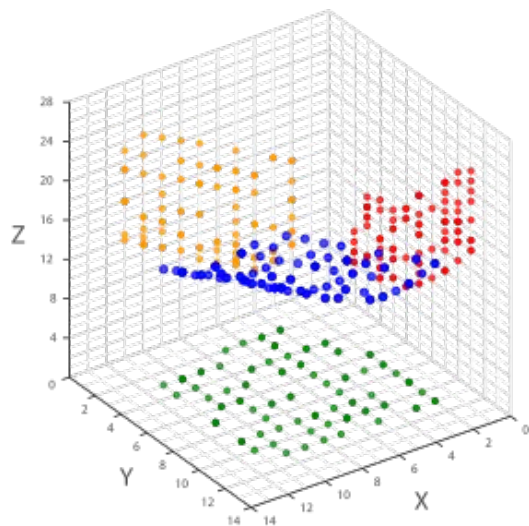
- Rather than finding the line of best fit, we are looking for the plane (3D) or hyperplane (4D, 5D, etc.) of best fit.
- We can still use the same measures of model fit (MSE , $RMSE$, R^2).

Regression in Orange



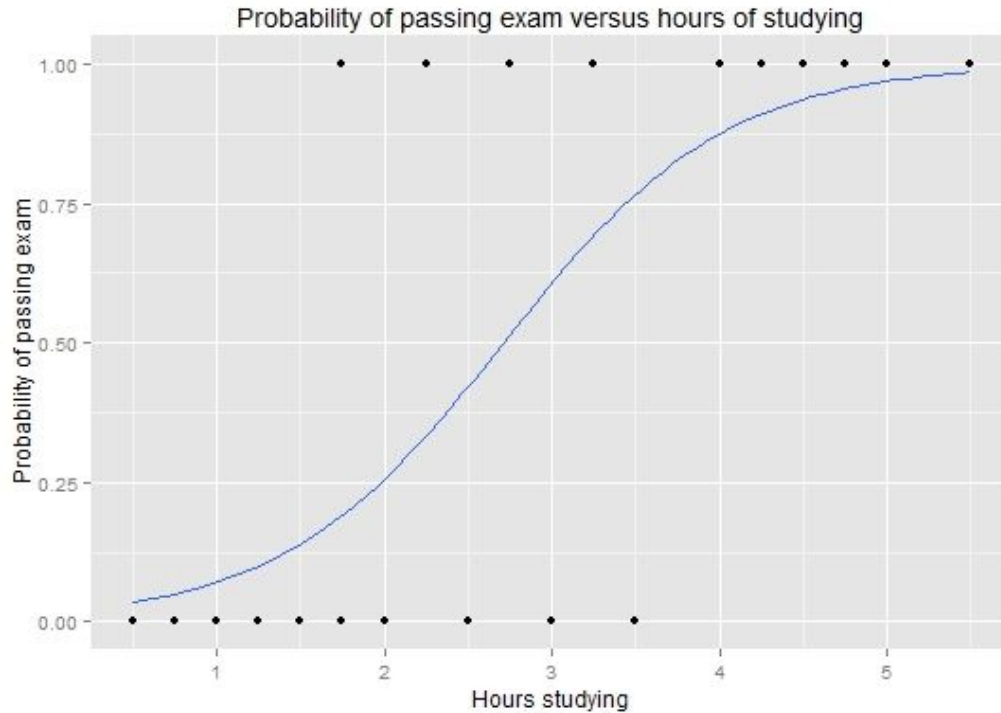
Regression

Find a data set online and determine the best fitting model.
Report the values of MSE , $RMSE$, MAE , and R^2 .

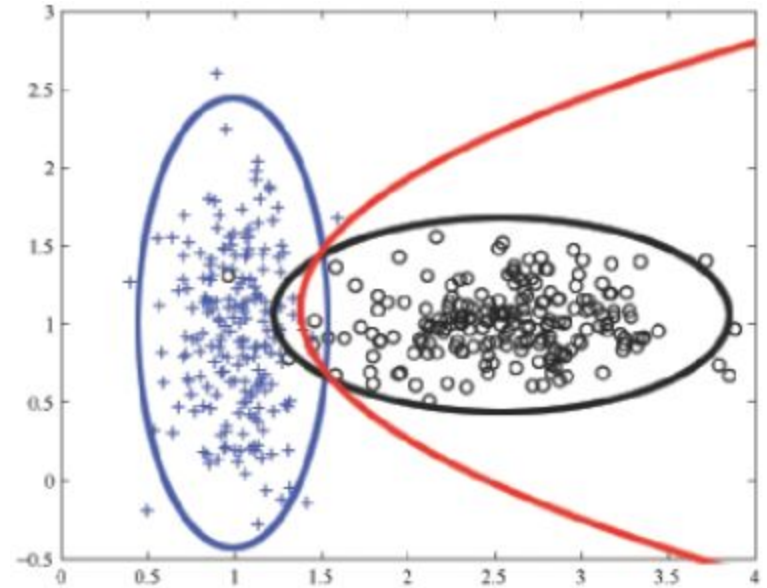
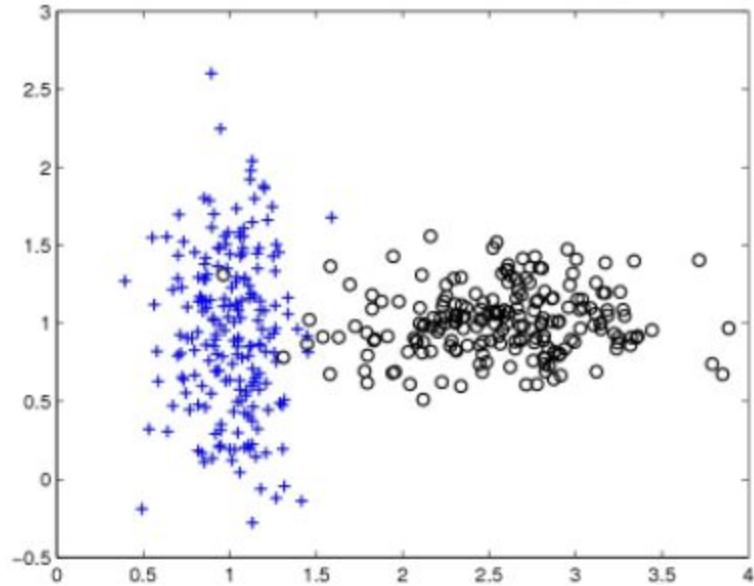


Other Machine Learning Algorithms

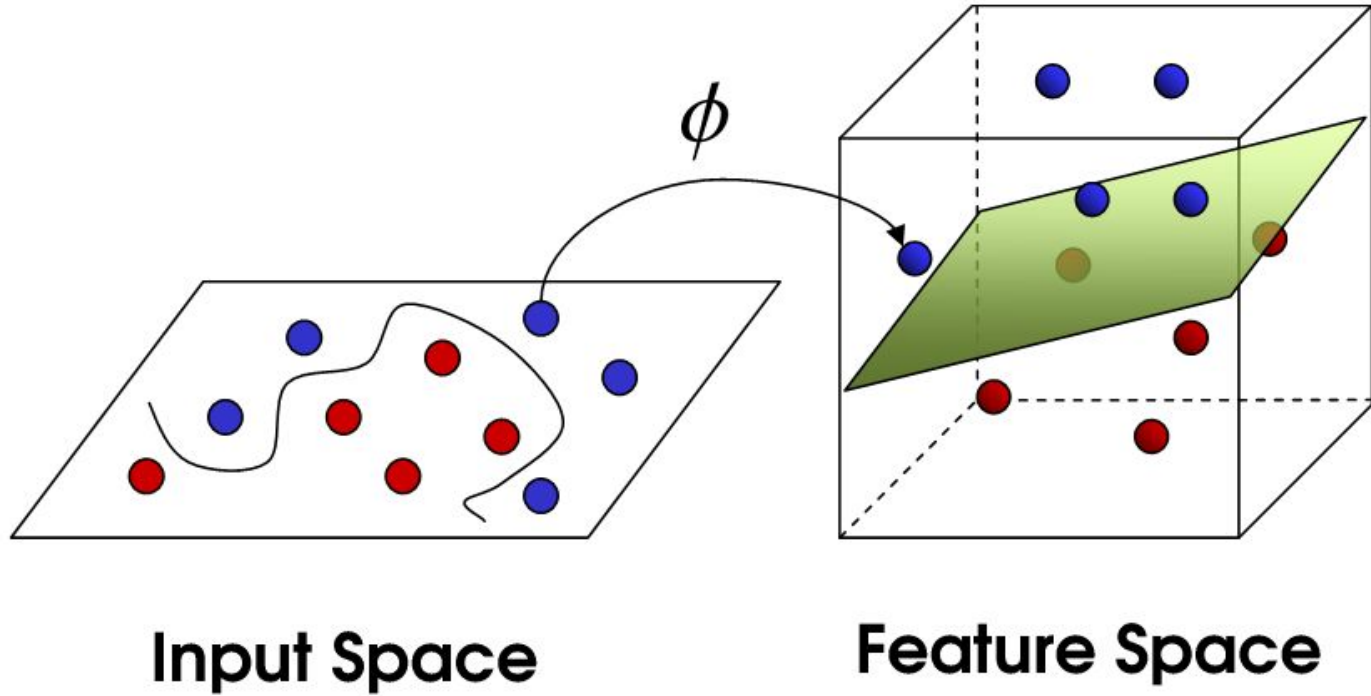
Logistic Regression



Gaussian Naive Bayes



Support Vector Machines



Machine Learning Project

Find a data set online and analyze it using multiple machine learning algorithms (KNN, Trees, K-Means, Regression). Also consider implementing algorithms we did not discuss (Logistic Regression, Naive Bayes, SVM). Be sure to include the following.

- Histogram
- Boxplot
- Scatterplot
- Decision Tree
- Confusion Matrix
- Accuracy
- Error
- Precision
- Recall
- F_1



Congratulations!

1. Evaluations
2. Certificates
3. Celebrate!