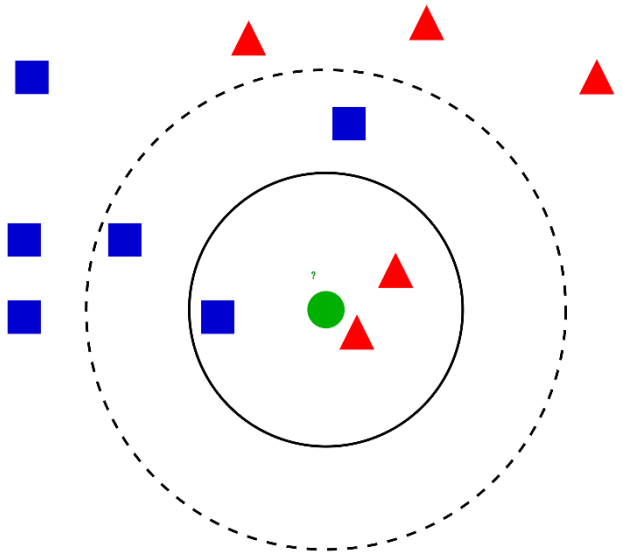
Model Evaluation

Nguyen Ngoc Thao
nnthao@fit.hcmus.edu.vn

Content outline

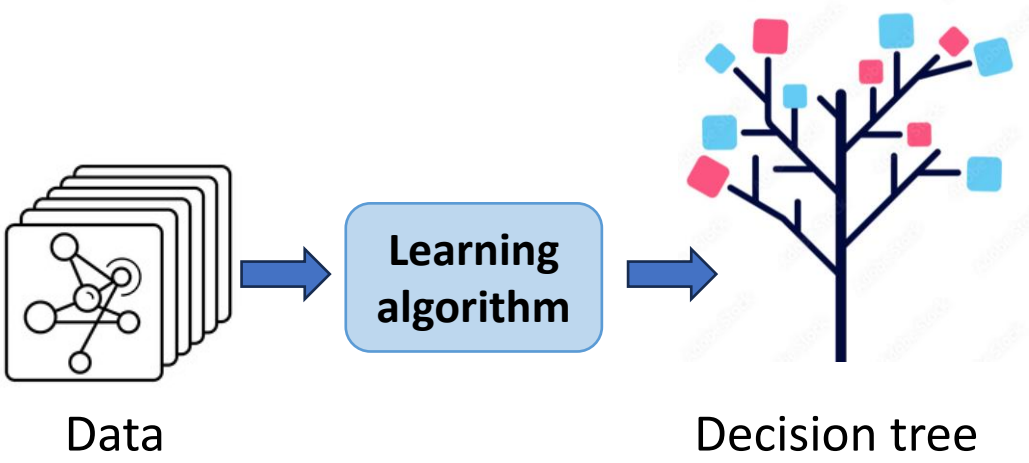
- k-nearest neighbors
- Evaluation metrics for classification
- Regression analysis
- k-means clustering
- Evaluation metrics for clustering

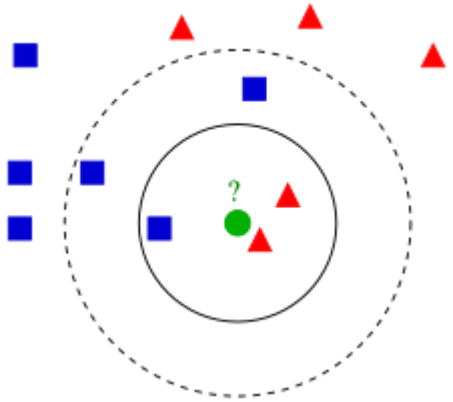
k-nearest neighbors



Lazy learning vs. Eager learning

- Eager learning: Build a classification model from a given set of training examples before classifying new data.
- Lazy learning Simply store the training data (or only minor processing) and delay until a test example comes.
  - Less time in training, yet more time in predicting.





k-nearest neighbors (k-NN)

- Majority voting: Classify an object to the class most common among its k nearest neighbors.
- k is usually a small positive integer, e.g., 1, 3, 5, etc.

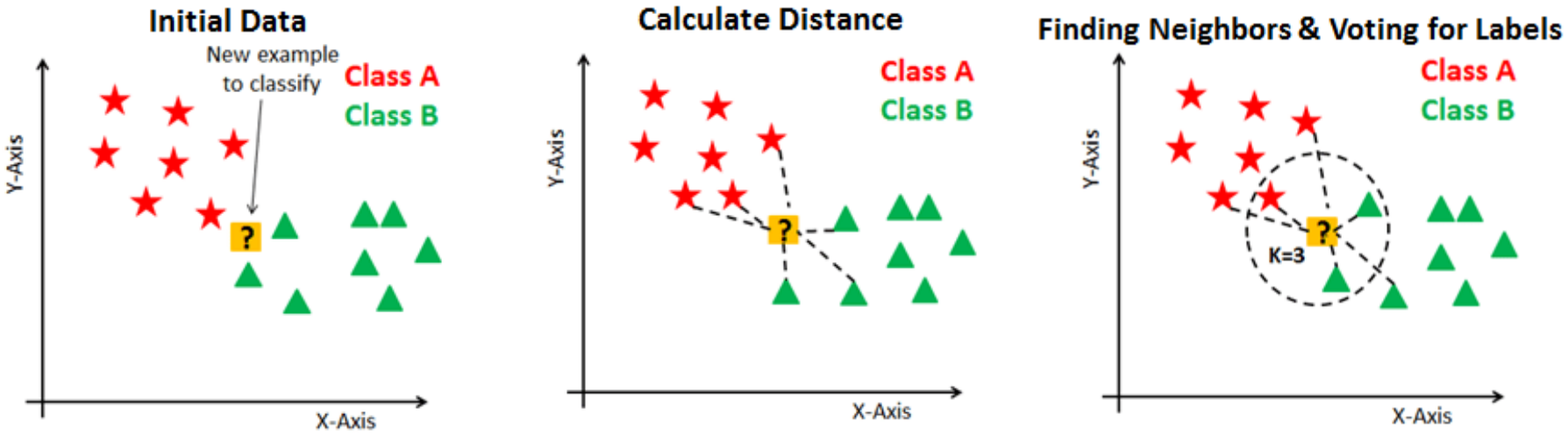
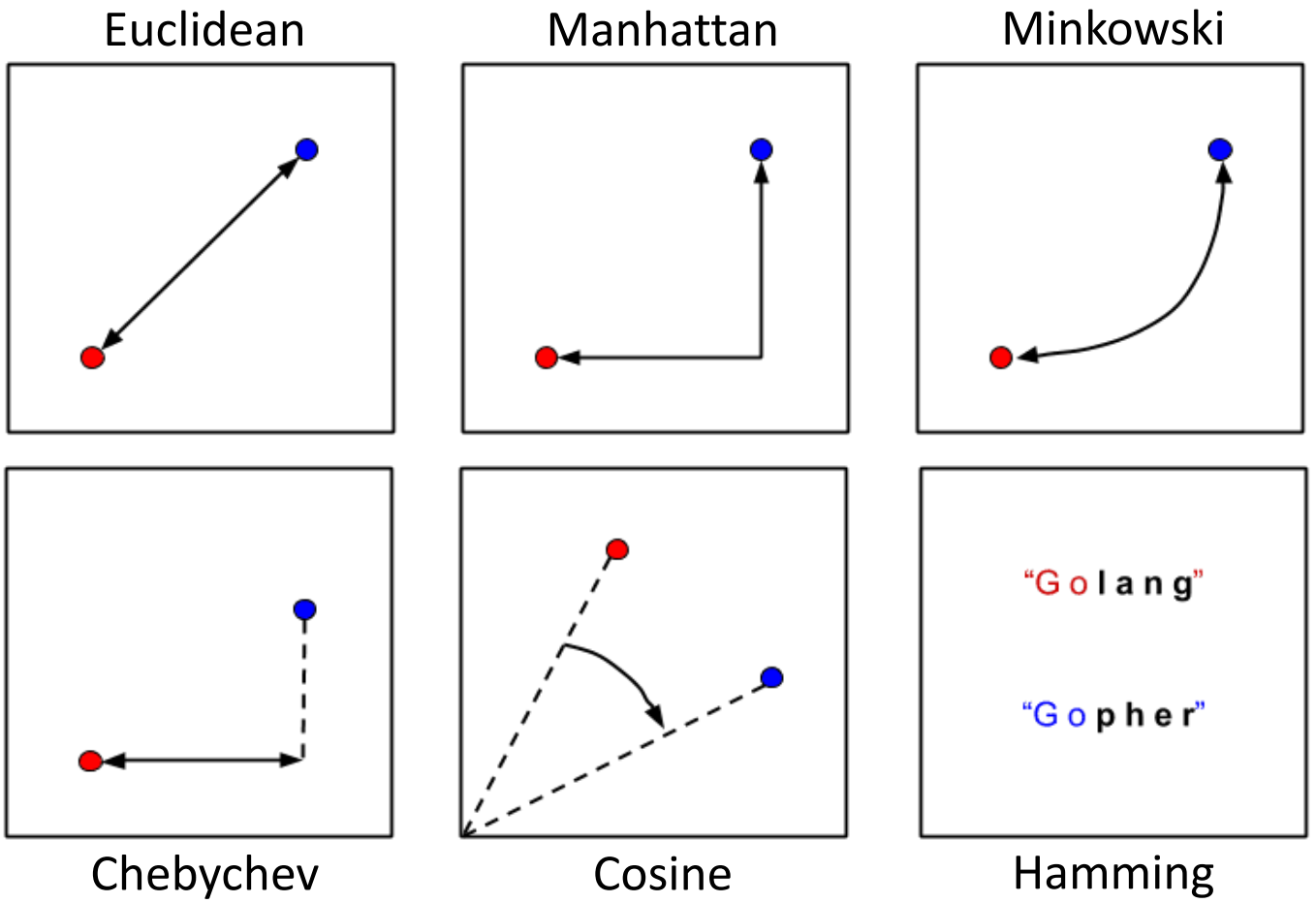


Image credit: kdnuggets

k-nearest neighbors (k-NN)

- The nearest neighbors are defined using a distance metrics.



k-nearest neighbors: An example

ID	Age	Income (K)	No. Cards	Response	L2-dist to unseen record
1	35	35	3	Yes	22.14
2	22	50	2	No	20.9
3	28	40	1	Yes	21.35
4	45	100	2	No	44.11
5	20	30	3	Yes	34.06
6	34	55	2	No	8.12
7	63	200	1	No	145.54
8	55	140	2	Yes	85.01
9	59	170	1	No	115.28
10	25	40	4	Yes	23.37
Unseen	**42**	**56**	**3**	**?**	

- **k = 5**: 3 “Yes” samples and 2 “No” samples → the class assigned is **Yes**

k-nearest neighbors: Normalization

- The attributes in the given data may have different scales, causing direct distance calculations to be inaccurate.
- For example, annual income is in dollars, and age is in years $\rightarrow$ income has a higher influence on the distance calculated

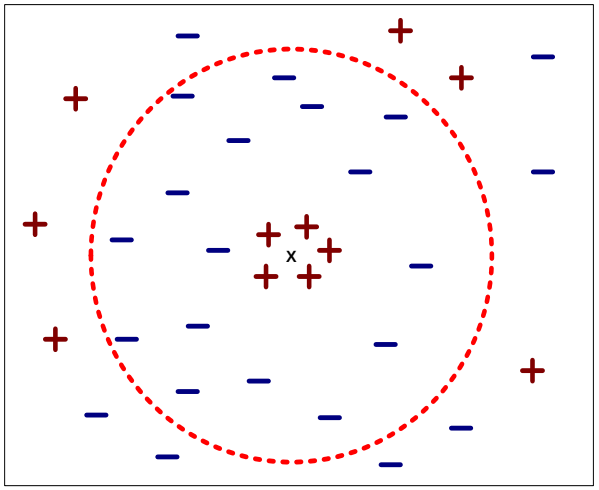
k-nearest neighbors: Normalization

ID	Age	Income (K)	No. Cards	Response	L2-dist to unseen record
1	0.35	0.03	0.67	Yes	0.21
2	0.05	0.12	0.33	No	0.57
3	0.19	0.06	0	Yes	0.75
4	0.58	0.41	0.33	No	0.43
5	0	0	0.67	No	0.53
6	0.33	0.15	0.33	No	0.38
7	1	1	0	No	1.19
8	0.81	0.65	0.33	Yes	0.67
9	0.91	0.82	0	No	1.03
10	0.12	0.06	1	Yes	0.52
Unseen	0.51	0.15	0.67	**?**	

- **k = 5**: 2 “Yes” samples and 3 “No” samples → the class assigned is **No**

k-nearest neighbors: Efficiency

- Easy to implement.
- Robust to noisy data by averaging k nearest neighbors



- All samples are stored $\rightarrow$ the test phase is time-consuming
- The value of k heavily affects the algorithm effectiveness.
  - Too small k: insufficient information for making decision
  - Too large k: noisy values may be included, or the neighborhood violate the areas of other classes

Quiz 01: k-nearest neighbors

1. Given the data set of vegetables and fruits below. We use two features, Sweet and Crunch, to classify food (Food Type).

No.	Object	Sweet	Crunch	Food Type
1	Grape	8	5	Fruit
2	Green bean	3	7	Vegetable
3	Nuts	3	6	Protein
4	Orange	7	3	Fruit

   Apply k-nn to predict the food type for Tomato (Sweet 6, Crunch = 4).

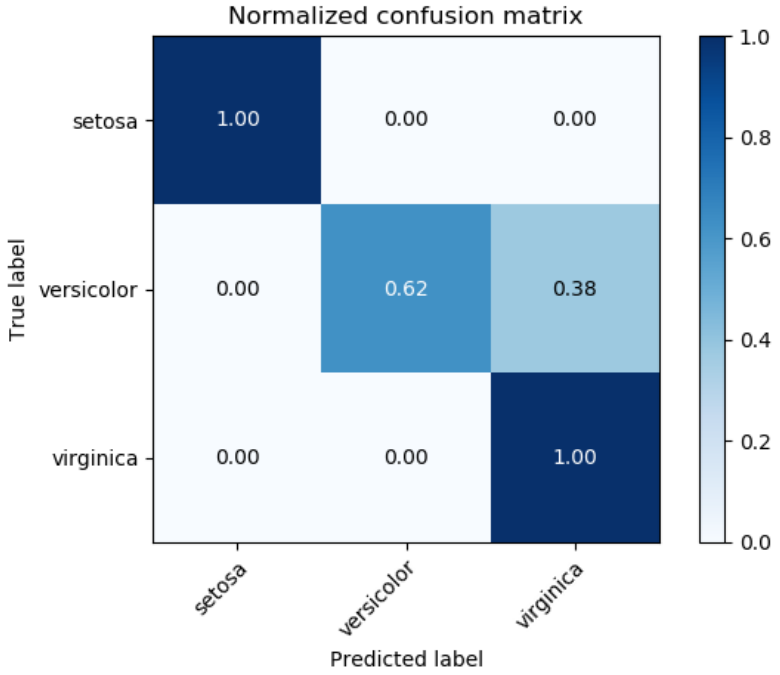
   - What kind of food does Tomato belong to if k = 1? Explain.
   - What kind of food does Tomato belong to if k = 3? Explain.

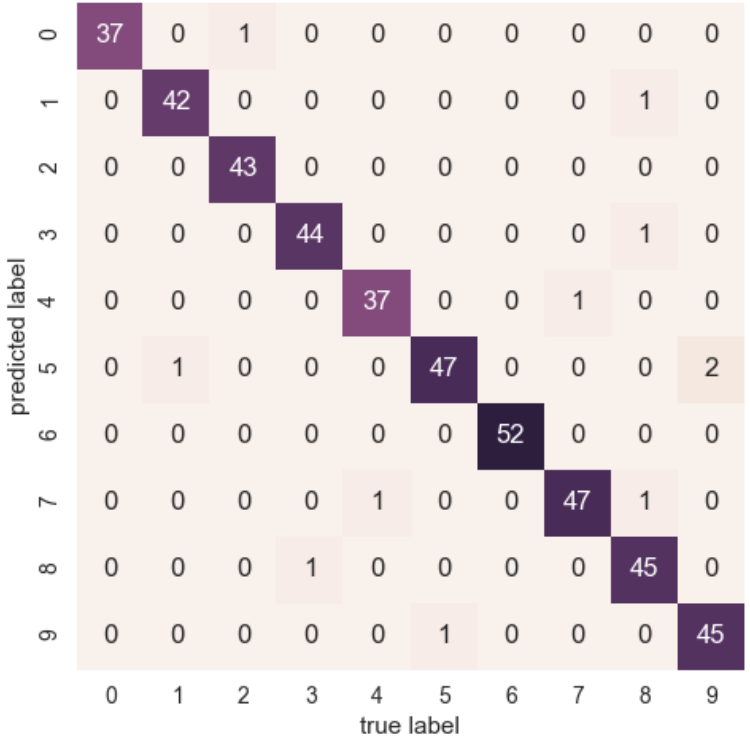
2. How to run k-nn in **scikit-learn**?

Evaluation metrics for classification

Confusion matrix

- An entry $[i, j]$ indicates the number of tuples in class i that were labeled by the classifier as class j.





Accuracy, Error rate, Sensitivity

		Predicted class		
		C_1	$\neg C_1$	
Actual Class	C_1	**True Positives (TP)**	**False Negatives (FN)**	**P**
	$\neg C_1$	**False Positives (FP)**	**True Negatives (TN)**	**N**
		P'	**N'**	**All**

- $Accuracy = (TP + TN)/All$
- $Error\ rate = 1 - Accuracy = \frac{FP+FN}{All}$
- $Sensitivity = TP/P$ (TP recognition rate)
- $Specificity = TN/N$ (TN recognition rate)

Precision, Recall, and F-measure

- **Precision:** exactness – % of tuples that the classifier labeled as positive are actually positive.

$$precision = \frac{TP}{TP + FP}$$

- **Recall:** completeness – % of positive tuples that the classifier labeled as positive.

$$recall = \frac{TP}{TP + FN}$$

- **F1-score**: harmonic mean of precision and recall

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

Evaluation metrics: An example

- Accuracy = (90+9560)/10000 = 0.964
- Error rate = 1 – 0.964 = 0.036
- Sensitivity = 90 / 300 = 0.3
- Specificity = 9560 / 9700 = 0.986
- Precision = 90/230 = 0.391
- Recall = 90/300 = 0.3

		Predicted class		**Total**
		cancer = yes	cancer = no	
Actual Class	cancer = yes	**90**	**210**	300
	cancer = no	**140**	**9560**	9700
Total		230	9770	10000

Quiz 02: Evaluation metrics

1. Given a confusion matrix that shows the classification results on three classes. Compute the evaluation metrics for each class.

		Actual Class			Total
		Cat	Dog	Monkey	
Predicted Class	Cat	15	10	5	**30**
	Dog	10	20	20	**50**
	Monkey	20	10	10	**40**
Total		**45**	**40**	**35**	**120**

2. How to compute the following metrics, accuracy, precision, recall, and F1, in **scikit-learn**?

Statistical tests of significance

- Given the two classifiers, M_1 and M_2, whose mean error rates (10-fold cross-validation) are $\overline{err}(M_1)$ and $\overline{err}(M_2)$.
- Which model is better?
- These mean error rates are just estimates of error on the true population of future data cases.
- What if the difference between the two error rates is just attributed to chance?
- Use a **test of statistical significance** to obtain **confidence limits** for error estimates.

Student's t-test

- Assume that the samples follow a t-distribution with $k-1$ degrees of freedom (here, $k=10$).
- Student's t-test null hypothesis $\boldsymbol{H_0}$: The two models, M_1 and M_2, have a zero difference in mean error rate.
- If the null hypothesis is rejected, the difference between M_1 and M_2 is statistically significant.

  $\rightarrow$ choose model with lower error rate.

t-test with a single test set

- The same test set can be used for both M_1 and M_2.
- Pairwise comparison
  - For i^{th} round of 10-fold cross-validation, the same cross partitioning is used to obtain $err(M_1)_i$ and $err(M_2)_i$
  - Average over 10 rounds to get $\overline{err}(M_1)$ and $\overline{err}(M_2)$
- ***t*-test** computes t-statistic with $k - 1$ degrees of freedom.

$$\boldsymbol{t = \frac{\overline{err}(M_1) - \overline{err}(M_2)}{\sqrt{var(M_1 - M_2)/k}}}$$

where

$$var(M_1 - M_2) = \frac{1}{k}\sum_{i=1}^{k}[err(M_1)_i - err(M_2)_i - (\overline{err}(M_1) - \overline{err}(M_2))]^2$$

t-test with two different test sets

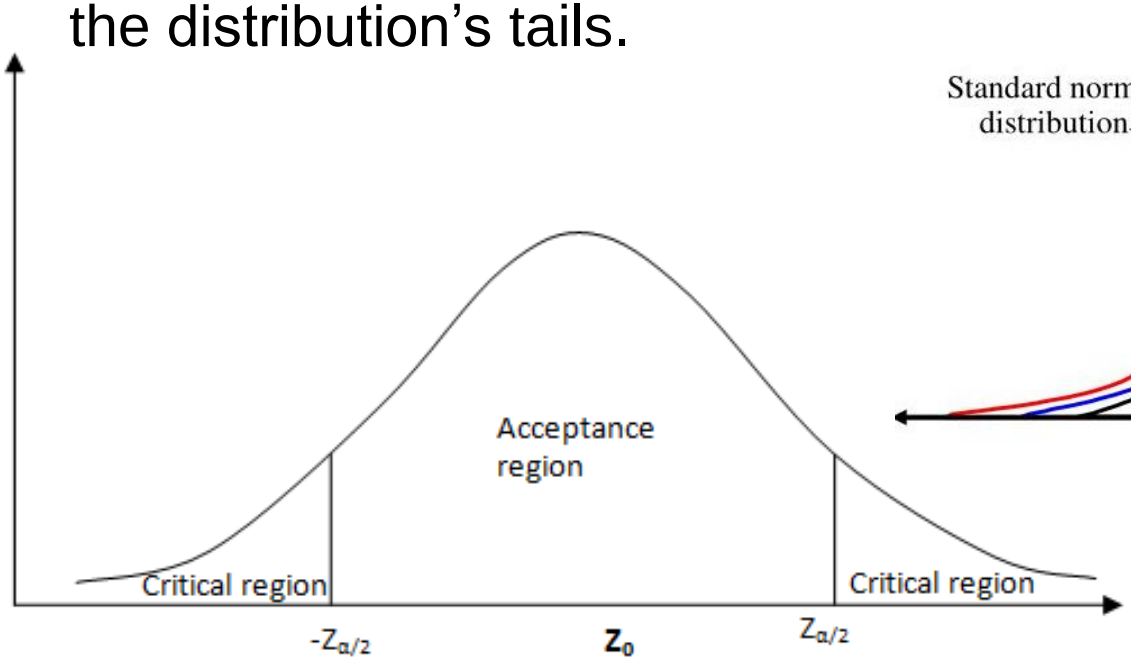
- Non-paired t-test: The variance between the means of the two models is estimated as

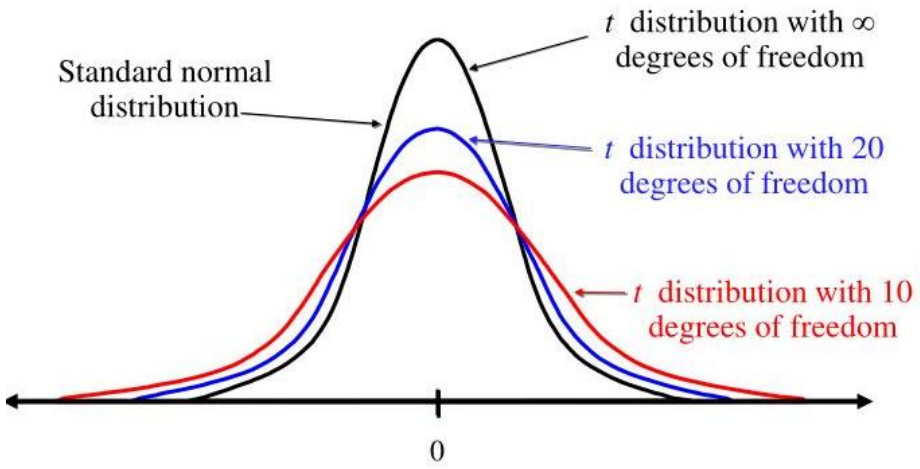
$$\boldsymbol{var(M_1 - M_2) = \sqrt{\frac{var(M_1)}{k_1} + \frac{var(M_2)}{k_2}}}$$

  - where k_1 and k_2 are the number of cross-validation samples (in our case, 10-fold rounds) used for M_1 and M_2, respectively.

t-distribution

- A significance level, e.g., 5%, indicates that there is a 95% confidence that your decision (reject H_0 or not) is correct, assuming all assumptions of the test are met.
- **Confidence limit**, $\boldsymbol{z = sig/2}$, e.g., 0.025 in this case.
- If $t > z$ or $t < -z$, the value of t lies in the rejection region, within the distribution's tails.



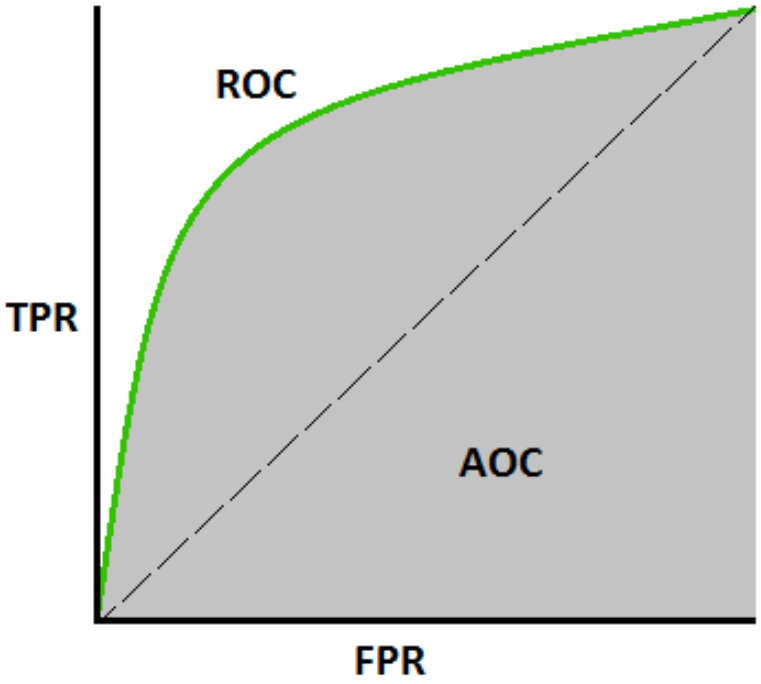


Costs and Benefits

- The cost associated with a false negative is sometimes far greater than those of a false positive.
  - E.g., incorrectly predicting a cancerous patient as not cancerous vs. incorrectly labeling a noncancerous patient as cancerous
- **Solution:** Assign a different cost to each type to outweigh one type of error over another.
  - E.g., the danger to the patient, financial costs of resulting therapies, and other hospital costs, etc.
- Similarly, the benefits associated with a true positive may be different than those of a true negative.

Receiver Operating Characteristics

- The ROC curve show the trade-off between the true positive rate (TPR) and the false positive rate (FPR).



Vertical axis represents the TPR while the horizontal axis for FPR

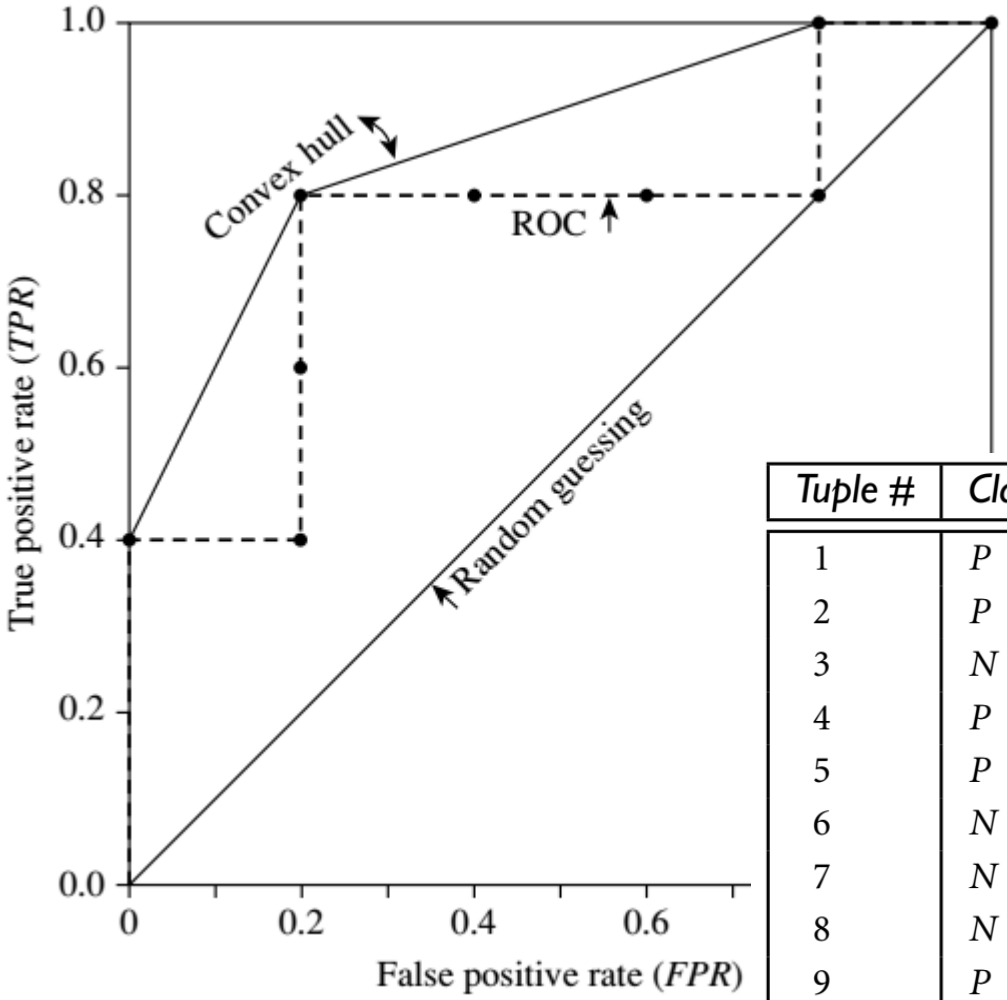
A diagonal line presents random guessing.

The area under the ROC curve is a measure of the accuracy of the model

ROC curves: Calculation

- The tuples are sorted in descending order of their likelihoods of belonging to the positive class.
  - The model M has to return a probability of the predicted class for a test tuple (e.g., naïve Bayesian, backpropagation classifiers, etc.).
- Let the value that a probabilistic classifier returns for a given tuple $\boldsymbol{X}$ be $f(\boldsymbol{X}) \rightarrow [0,1]$.
- For a binary problem, a threshold t is typically selected so that tuples where $f(\boldsymbol{X}) \geq t$ are considered positive

ROC curves: An example



There are five positive tuples and five negative tuples.

$P = 5$ and $N = 5$.

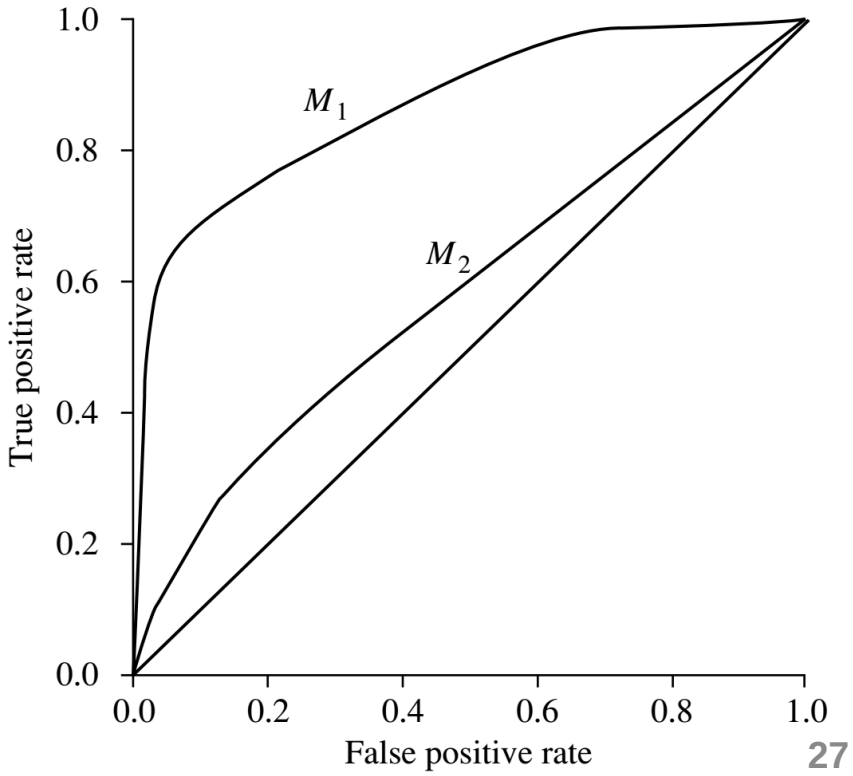
Tuple #	Class	Prob.	TP	FP	TN	FN	TPR	FPR
1	P	0.90	1	0	5	4	0.2	0
2	P	0.80	2	0	5	3	0.4	0
3	N	0.70	2	1	4	3	0.4	0.2
4	P	0.60	3	1	4	2	0.6	0.2
5	P	0.55	4	1	4	1	0.8	0.2
6	N	0.54	4	2	3	1	0.8	0.4
7	N	0.53	4	3	2	1	0.8	0.6
8	N	0.51	4	4	1	1	0.8	0.8
9	P	0.50	5	4	0	1	1.0	0.8
10	N	0.40	5	5	0	0	1.0	1.0

ROC curves: Model selection

- We measure the area under the curve to assess the model.
  - A model with perfect accuracy has an area of 1.0.
- The closer the ROC curve of a model is to the diagonal line, the less accurate the model.

Which one, M_1 or M_2, is more accurate?

The closer the area is to 0.5, the less accurate the model is.



Quiz 03: ROC curves

The aside table shows the ten tuples in a test set, sorted by decreasing probability order. There are 4 positive tuples and 6 negative tuples in the test set. Each tuple is shown in a separate row.

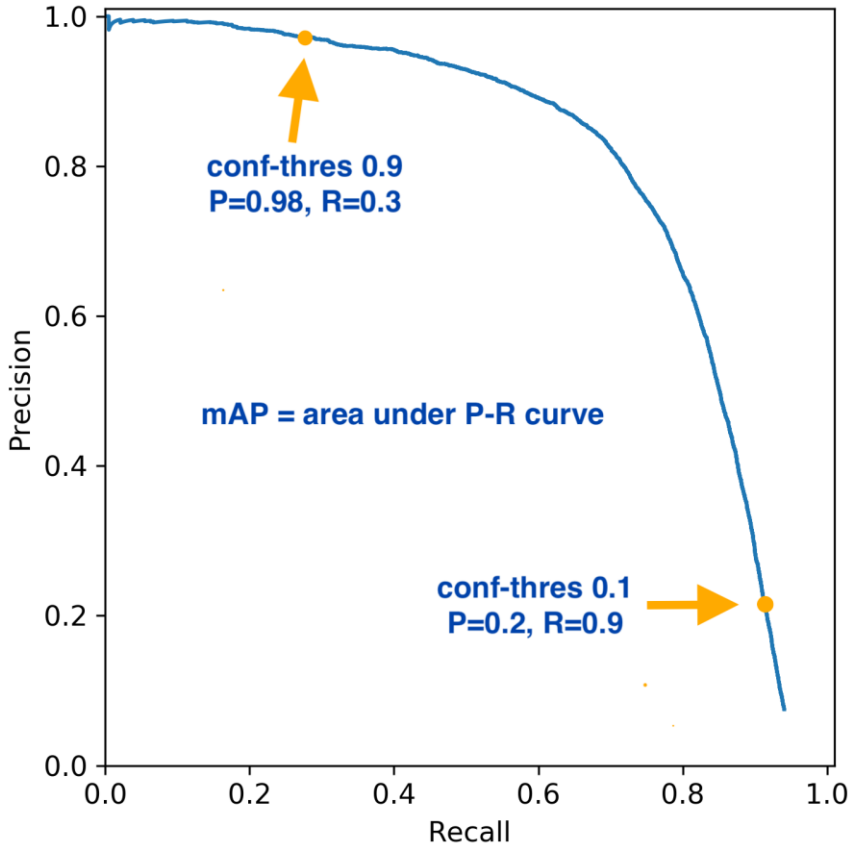
For each tuple, the first column denotes the ranking, the second column shows the actual class label of the tuples, and the third column is the probability returned by a probabilistic classifier.

1. Draw the ROC curve.
2. How to draw the ROC curve in **scikit-learn**?

Rank#	Class	Prob
1	**1**	0.91
2	**1**	0.86
3	0	0.85
4	**1**	0.57
5	0	0.54
6	0	0.26
7	**1**	0.18
8	0	0.16
9	0	0.14
10	0	0.13

Precision – Recall (PR) Curve

- The PR curve show the trade-off between the precision and the recall values.

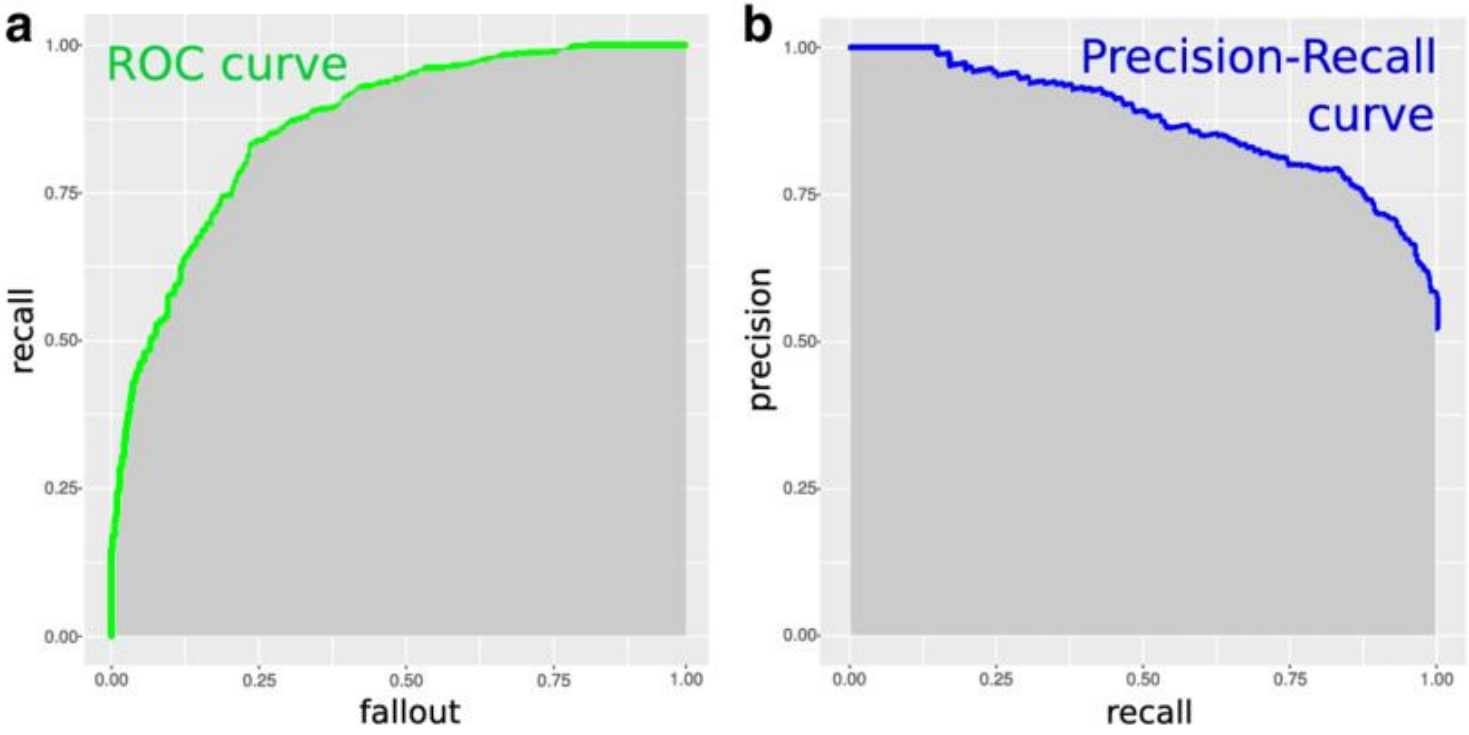


Vertical axis represents the precision while the horizontal axis for recall.

The area under the PR curve is a measure of the accuracy of the model

PR curve: Model selection

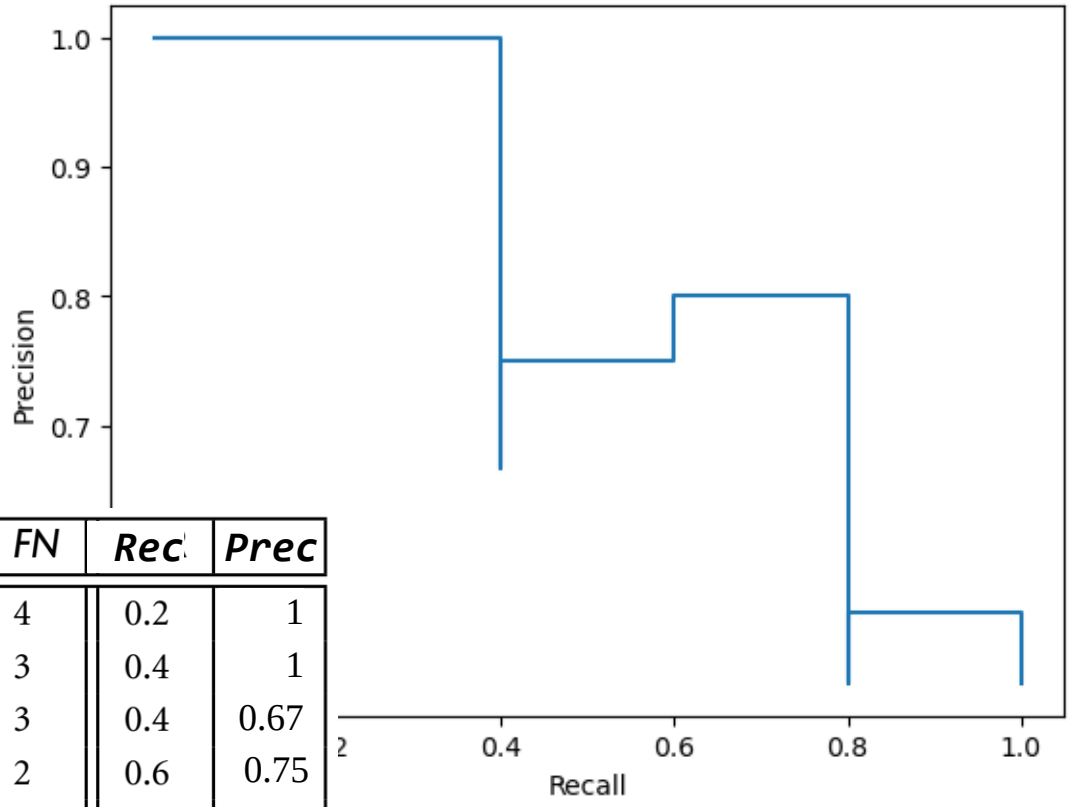
- Similar to ROC curves, we measure the area under the curve to assess the model.
- A model with perfect accuracy has an area of 1.0.



PR curves: An example

There are five positive tuples and five negative tuples.

$P = 5$ and $N = 5$.



Tuple #	Class	Prob.	TP	FP	TN	FN	Rec	Prec
1	P	0.90	1	0	5	4	0.2	1
2	P	0.80	2	0	5	3	0.4	1
3	N	0.70	2	1	4	3	0.4	0.67
4	P	0.60	3	1	4	2	0.6	0.75
5	P	0.55	4	1	4	1	0.8	0.8
6	N	0.54	4	2	3	1	0.8	0.67
7	N	0.53	4	3	2	1	0.8	0.57
8	N	0.51	4	4	1	1	0.8	0.5
9	P	0.50	5	4	0	1	1.0	0.55
10	N	0.40	5	5	0	0	1.0	0.5

ROC curves vs. PR curves

- ROC curve is best for balanced datasets.
  - The proportion of positive and negative classes is roughly equal.
- It evaluates performance across all decision thresholds and considers both classes equally.

- PR curve is ideal for imbalanced datasets where the positive class is rare.
- It focuses on the performance for the positive class and is more sensitive to changes in predicting positives.

Issues affecting model selection

- Classifier accuracy when predicting the class label
- Time to construct the model (training time) and time to use the model (classification / prediction time).
- Robustness: handling noise and missing values
- Scalability: efficiency in disk-resident databases
- Interpretability: the insights provided by the model
- Other measures: rules' goodness and compactness, and decision tree's size, etc.

Quiz 04: PR curves

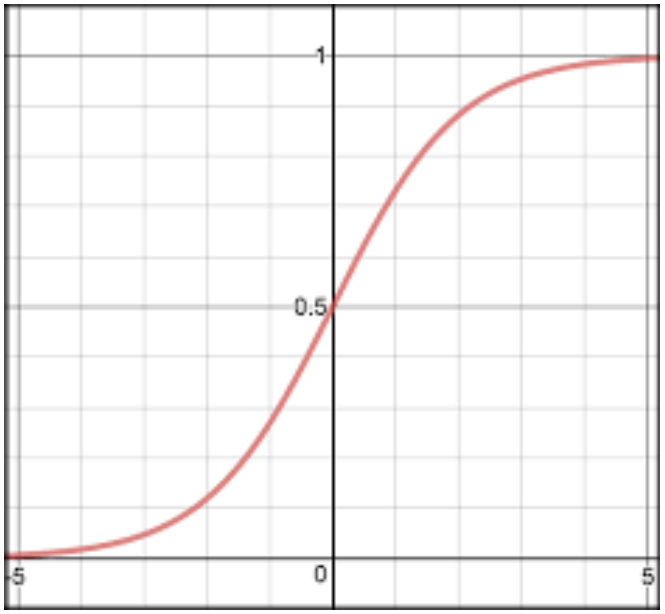
The aside table shows the ten tuples in a test set, sorted by decreasing probability order. There are 4 positive tuples and 6 negative tuples in the test set. Each tuple is shown in a separate row.

For each tuple, the first column denotes the ranking, the second column shows the actual class label of the tuples, and the third column is the probability returned by a probabilistic classifier.

1. Draw the PR curve.
2. How to draw the PR curve in **scikit-learn**?

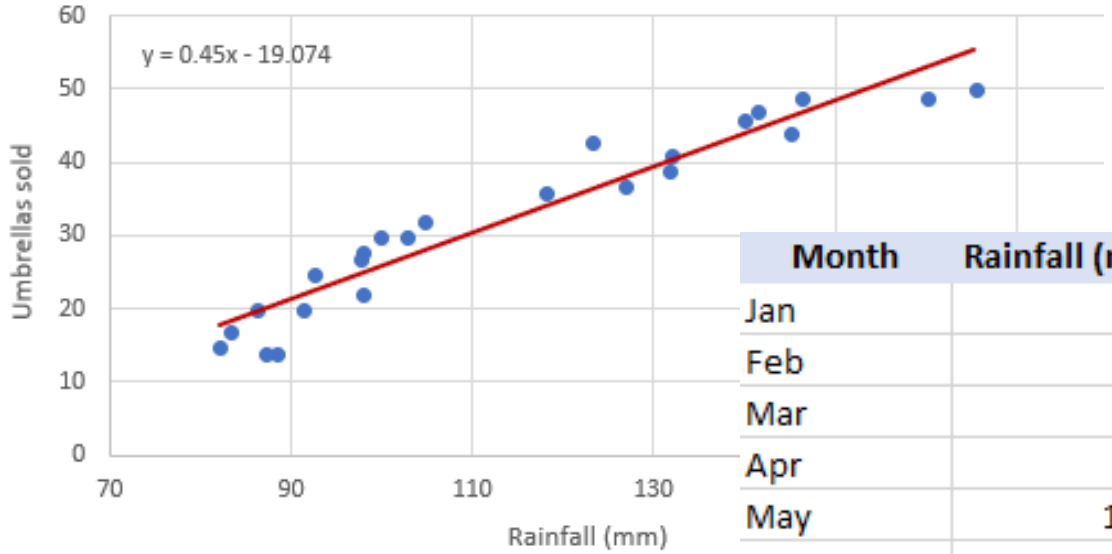
Rank#	Class	Prob
1	**1**	0.91
2	**1**	0.86
3	0	0.85
4	**1**	0.57
5	0	0.54
6	0	0.26
7	**1**	0.18
8	0	0.16
9	0	0.14
10	0	0.13

Regression analysis



Regression analysis

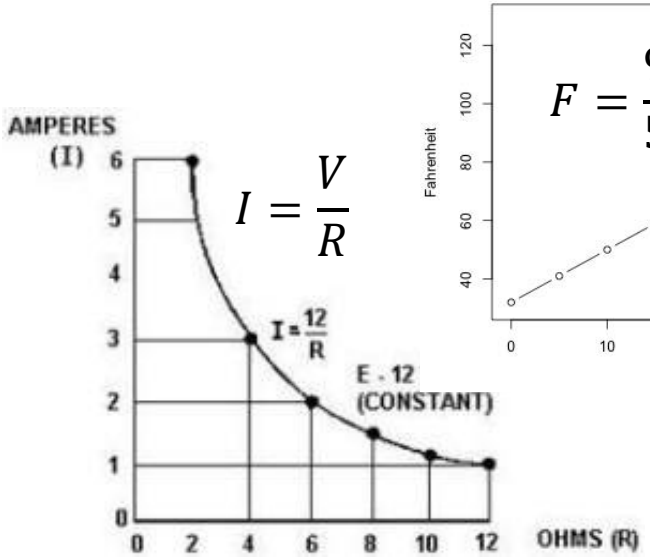
- A set of statistical processes for modelling the relationships between a dependent variable and one or more independent variables in a non-deterministic manner.
- The need occurs frequently in engineering and science.

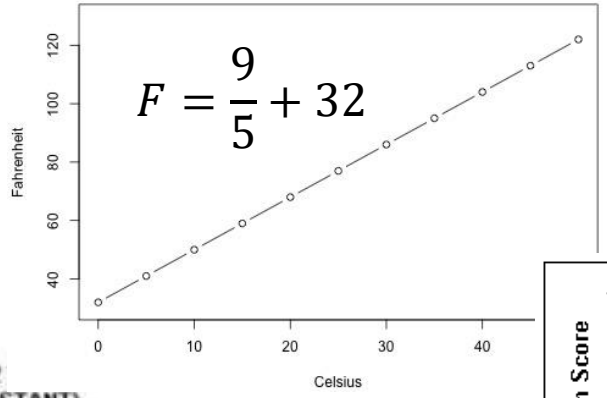


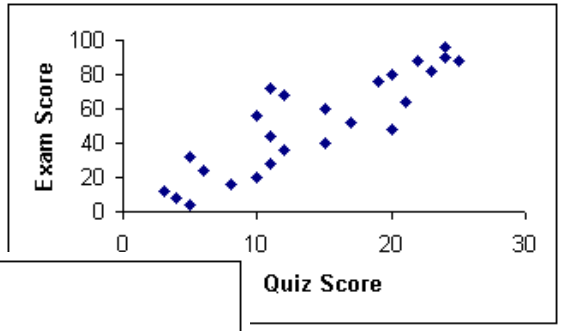
Month	Rainfall (mm)	Umbrellas sold
Jan	82	15
Feb	92.5	25
Mar	83.2	17
Apr	97.7	28
May	131.9	41
Jun	141.3	47
Jul	165.4	50

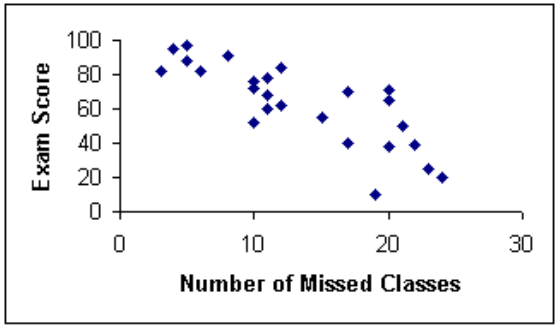
Deterministic relationship?

- **Deterministic relationship:** The equation precisely characterizes the relationship between the two variables.
  - The observed (x, y) data points fall directly on a line.





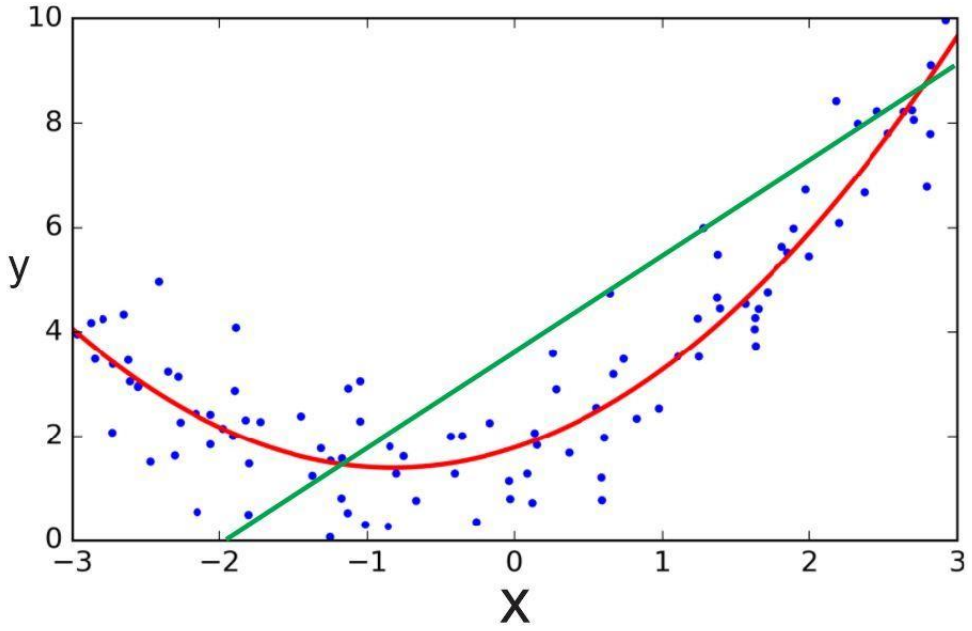




- Non-deterministic relationship: two variables do not correlate perfectly.
  - The plot exhibits some "trend," but it also exhibits some "scatter."

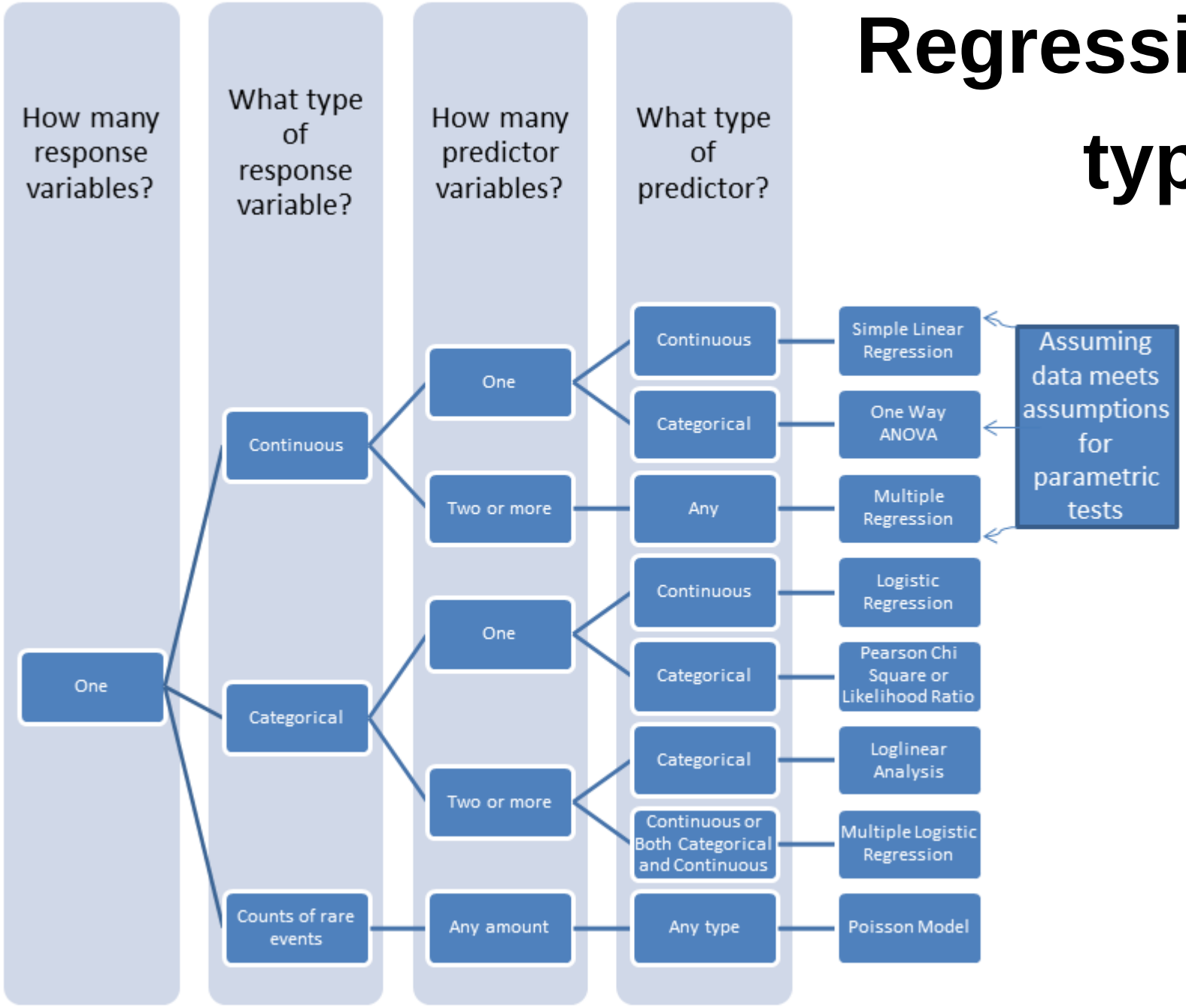
Regression: Basic concepts

- We observe a response (dependent) variable $\boldsymbol{Y}$, and many predictor (independent) variables, $\{\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}\}$.
- **Goal:** find the mathematical relationship between responses and predictors, $\boldsymbol{Y = h(X_1, \ldots, X_n)}$



Linear regression (green line)
vs.
Non-linear regression (red line)

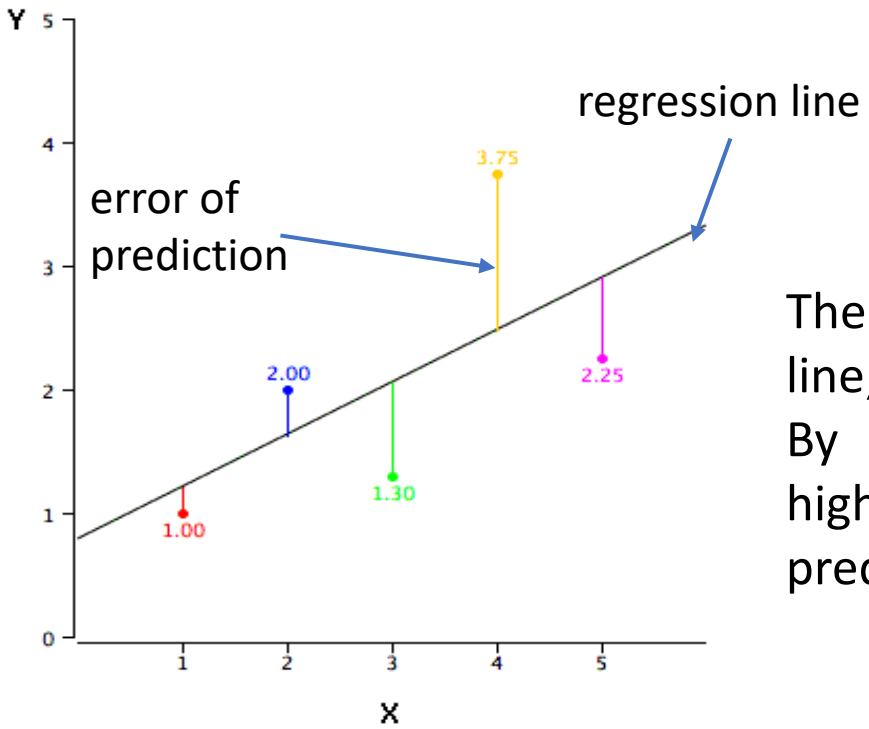
Regression types



Simple linear regression

Simple linear regression

- Study the relationship between two quantitative variables by finding the best-fitting straight line through the points
- Regression line: the best fitting line that includes the predicted score on Y for each possible value of X.

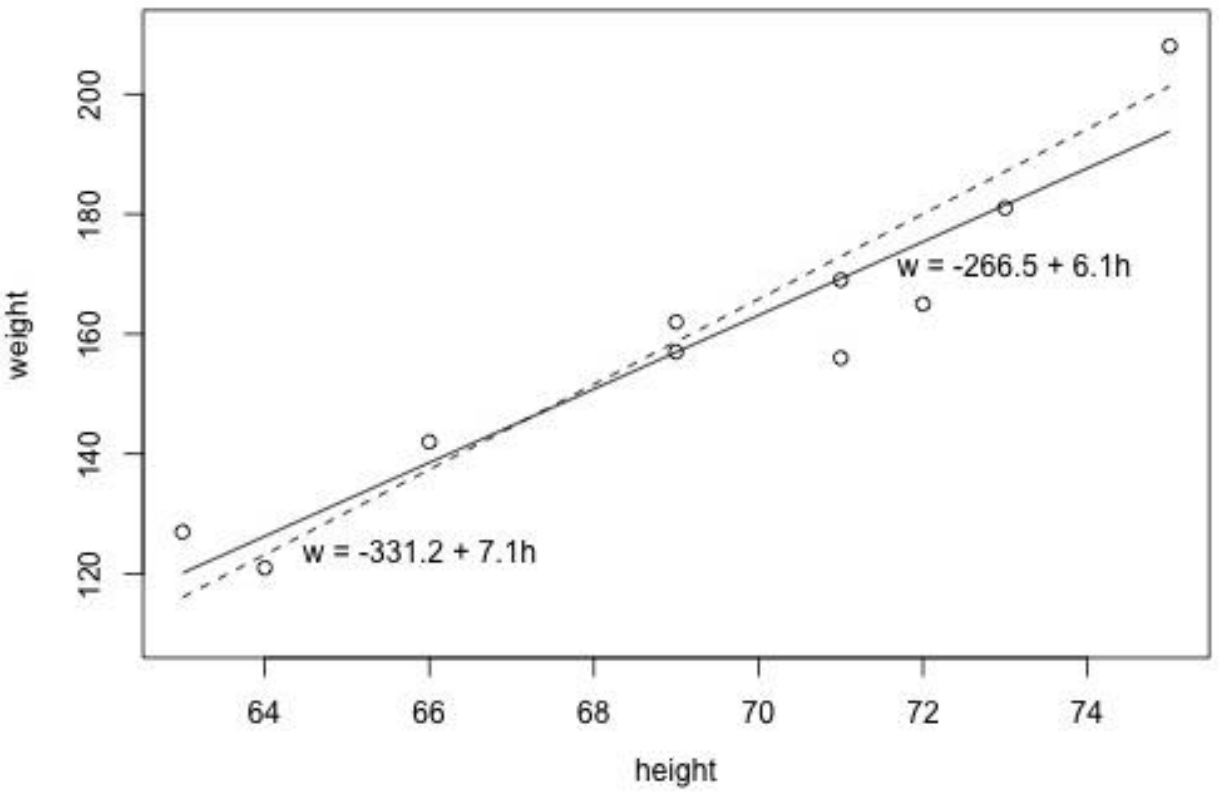


The red point is very near the regression line; its error of prediction is small.
By contrast, the yellow point is much higher than the regression line; its error of prediction is large

What is the "Best fitting line"?

- Given a set of heights (x) and weights (y) of 10 students.

i	x_i	y_i
1	63	127
2	64	121
3	66	142
4	69	157
5	69	162
6	71	156
7	71	169
8	72	165
9	73	181
10	75	208



Which line — the solid line or the dashed line — best summarizes the trend between height and weight?

What is the "Best fitting line"?

- Consider a student i.
- Let x_i , y_i and $\hat{y}_i$ be the predictor value (height), observed response (weight), and the predicted response, respectively.
- Then, the equation for the best fitting line is $\hat{y}_i = \beta_0 + \beta_1 x_i$.

- Let's try the first point with the line $w = 266.53 + 6.1376\ h$
  - If this student's height is 63 inches, how many pounds is his weight?
  - Given $x = 63$, $\hat{y}_i$ is 266.53 + 6.1376 $\times$ 63 = 120.1, while $y = 127$.
  - The difference is 127 – 120.1 = 6.9 pounds.

What is the "Best fitting line"?

- The prediction error (or "residual error") is the difference between the actual value and predicted value of the point.

$$r_i = \hat{y}_i - y_i$$

- The best fitting line gives the minimal total of n prediction errors, one for each observed data point.
- Least square criterion: find the values of β_0 and β_1 that minimize

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

What is the "Best fitting line"?

w = -**331.2** + **7.1** h (the dashed line)					
i	x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	63	127	116.1	10.9	118.81
2	64	121	123.2	-2.2	4.84
3	66	142	137.4	4.6	21.16
4	69	157	158.7	-1.7	2.89
5	69	162	158.7	3.3	10.89
6	71	156	172.9	-16.9	285.61
7	71	169	172.9	-3.9	15.21
8	72	165	180.0	-15.0	225.00
9	73	181	187.1	-6.1	37.21
10	75	208	201.3	6.7	44.89
					766.5

w = -**266.53** + **6.1376** h (the solid line)					
i	x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	63	127	120.139	6.8612	47.076
2	64	121	126.276	-5.2764	27.840
3	66	142	138.552	3.4484	11.891
4	69	157	156.964	0.0356	0.001
5	69	162	156.964	5.0356	25.357
6	71	156	169.240	-13.2396	175.287
7	71	169	169.240	-0.2396	0.057
8	72	165	175.377	-10.3772	107.686
9	73	181	181.515	-0.5148	0.265
10	75	208	193.790	14.2100	201.924
					597.4

Least squares estimates

- There are formulas for the intercept β_0 and the slope β_1 to minimize the sum of the squared prediction errors.
- Least squares estimates: take the derivative with respect to β_0 and β_1, set to 0, and solve for β_0 and β_1

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- In practice, let statistical software find those least squares lines for you!

Estimated regression coefficients

- The estimated regression coefficients, β_0 and β_1, allow for prediction of future responses.
- If the "scope of the model" includes $x = 0$, then β_0 is the predicted mean response when $x = 0$; otherwise, β_0 is not meaningful.
  - E.g., a person of 0 inches tall is predicted to weigh -266.53 pounds $\rightarrow$ this prediction is nonsense $\rightarrow$ extrapolate beyond the "scope of the model".
- The mean response is expected to increase or decrease by β_1 units for one unit increase in x.
  - E.g., when subtracting the predicted weight of 66"-inch-tall people from the predicted weight of 67"-inch-tall people, we obtain 144.69 - 138.55 = 6.14 pounds -- the value of β_1.
- More discussion can be found here.

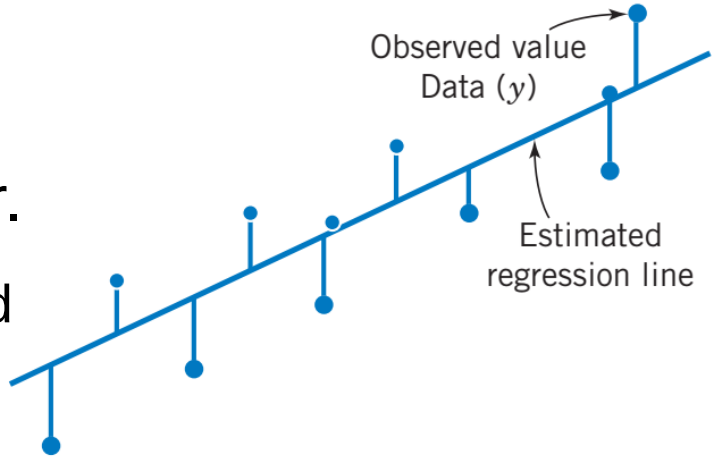
Simple linear regression model

- The simple linear regression model can be written as

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

  - where $i = 1,2,\ldots,n$ and $Var(\varepsilon_i) = \sigma^2$

- The estimates are subjected to error.
  - Any response y_i will be the linear trend $\beta_0 + \beta_1 x$ plus some error ε_i.



- Unbiased estimator (MSE: Mean Squared Error)

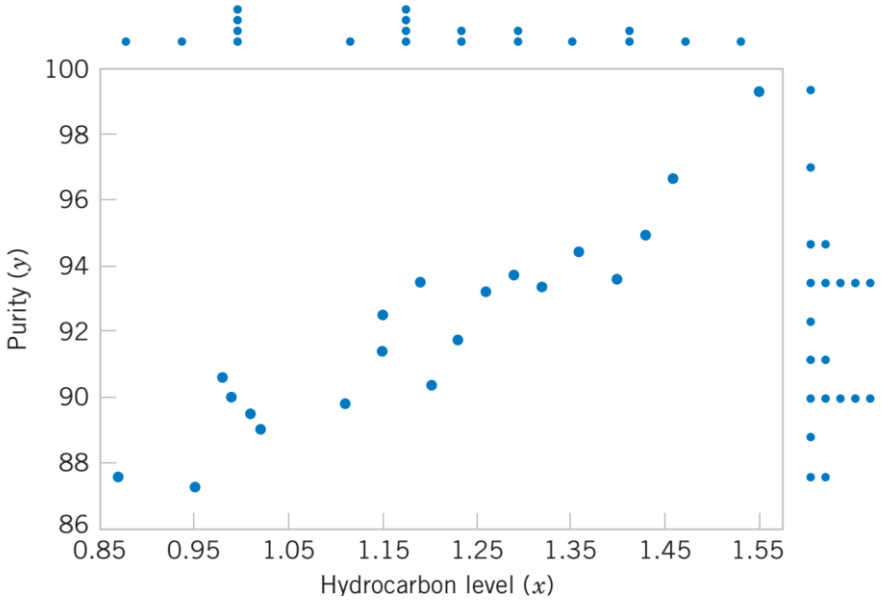
$$\hat{\sigma}^2 = MSE = \frac{\sum_{i=1}^{n} r_i^2}{n-2}$$

  - $\sigma = \sqrt{MSE}$ is the regression standard error or residual standard error.

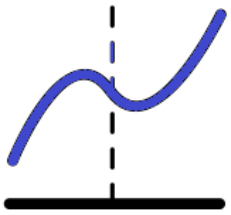
Quiz 05: Simple linear regression

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

- Fit a simple linear regression model to relate purity (y) to hydrocarbon level (x).
- What is the estimate of the error variance?



The scatter diagram of oxygen purity versus hydrocarbon level



Logistic regression

Binomial logistic regression

- Consider a model with two predictors, X_1 and X_2, and one binary response variable Y, with parameter $P = P(Y = 1)$.
- Assume that a linear relationship between the predictors and the log-odds (or logit) of the event that $Y = 1$.

$$l = \log_b\left(\frac{P}{1-P}\right) = \beta_o + \beta_1 X_1 + \beta_2 X_2$$

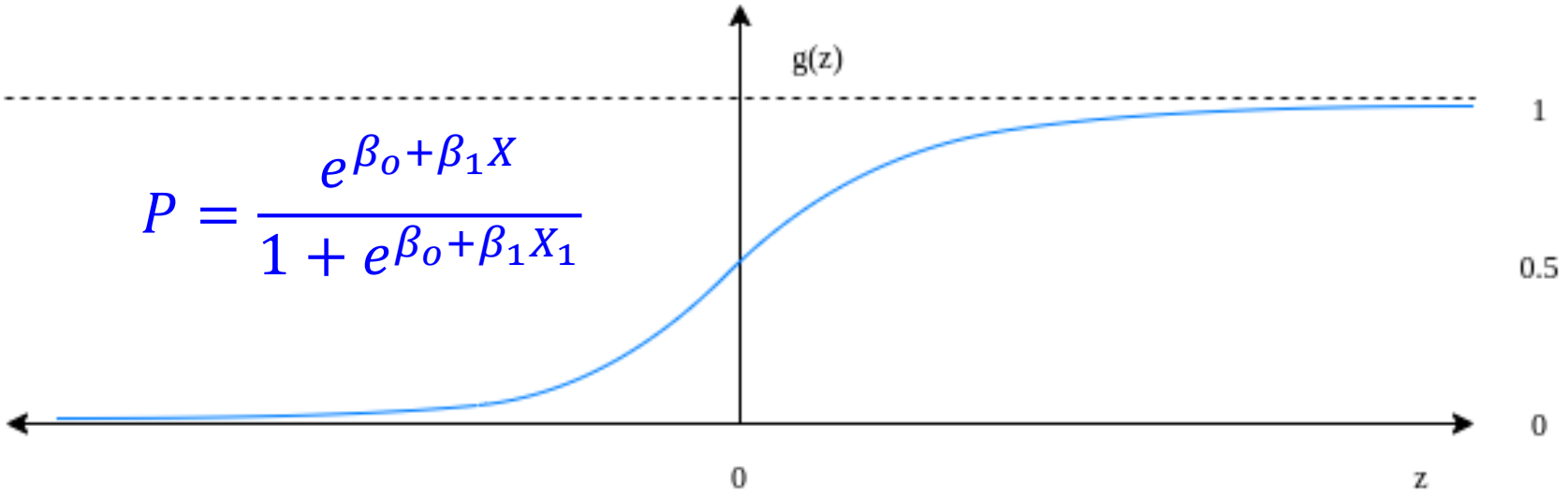
Log odds or Logit Intercept

  - l is the log-odds, b is the logarithm base, and β_i are coefficients of the equation on the right-hand side.
- Then, solve for the probability P by taking the exponential on both sides of the equation

Binomial logistic regression

- The probability of a categorical response based on one or more predictor variables can be estimated as

$$P = \frac{e^{\beta_o + \beta_1 X}}{1 + e^{\beta_o + \beta_1 X_1}}$$



- Use stochastic gradient descent (SGD) to find the coefficients

Logistic regression: An example

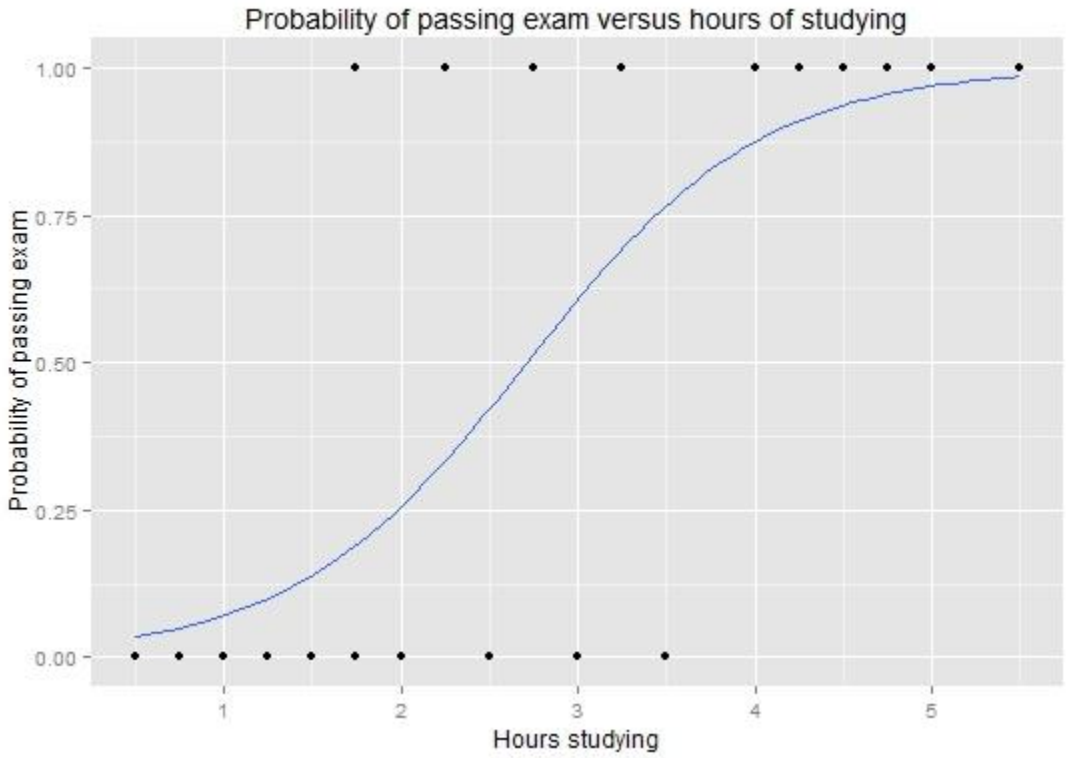
- A group of 20 students spends 0-6 hours studying for an exam.
- The table shows the number of hours each student spent studying, and whether they passed (1) or failed (0).

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50
Pass	0	0	0	0	0	0	1	0	1	0
Hours	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	1	0	1	0	1	1	1	1	1	1

- How does the number of hours spent studying affect the probability of the student passing the exam?

Logistic regression: An example

- Logistic regression analysis



- The probability of passing the exam for a student who studies 2 hours is
  - $l = -4.078 + 1.505 \cdot 2 = -1.069$ (β_o and β_1 are given example values)
  - $P = \frac{1}{1+e^{-l}} = 0.26$

Quiz 06: Logistic regression

The task is to predict if a given city has a risk of a disease epidemic or not. The data is defined using two input features or variables.

- logarithm of size of the city
- distance to the nearest city with epidemic

City #	Log Size of City	Distance	Risk?
1	0.09	7.2	0
2	-0.48	10	0
3	0.62	2.7	0
4	0.57	2.8	1
5	0.44	0.01	1

Quiz 06: Logistic regression

Suppose that we use the following coefficients for the logistic regression model: b0 = 1.05, b1 = -0.52, and b2 = 0.85.

a) Use the model above (with probability cutoff 0.5) and classify the given examples as "Risk" or "No risk", if p > 0.5 then the case as a "Risk". For each example, present its logit and predicted class.

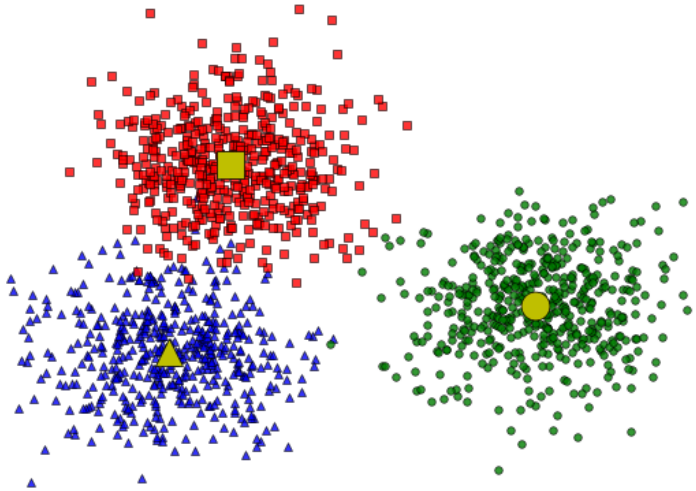
b) Present the results in a confusion matrix, i.e., identify the terms TP, FP, TN, and FN.

Quiz 06: Logistic regression

The training data set is as follows.

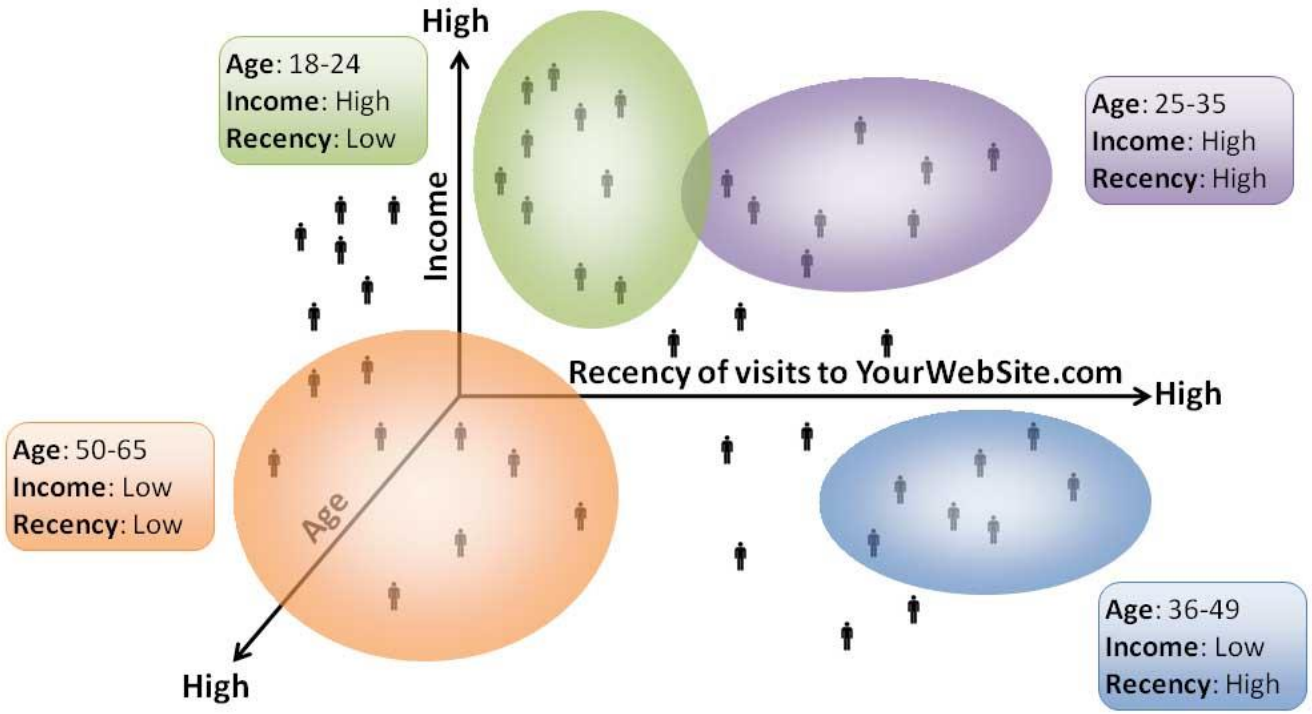
City #	Log Size of City	Distance	Risk?
1	0.09	7.2	0
2	-0.48	10	0
3	0.62	2.7	0
4	0.57	2.8	1
5	0.44	0.01	1

k-means clustering



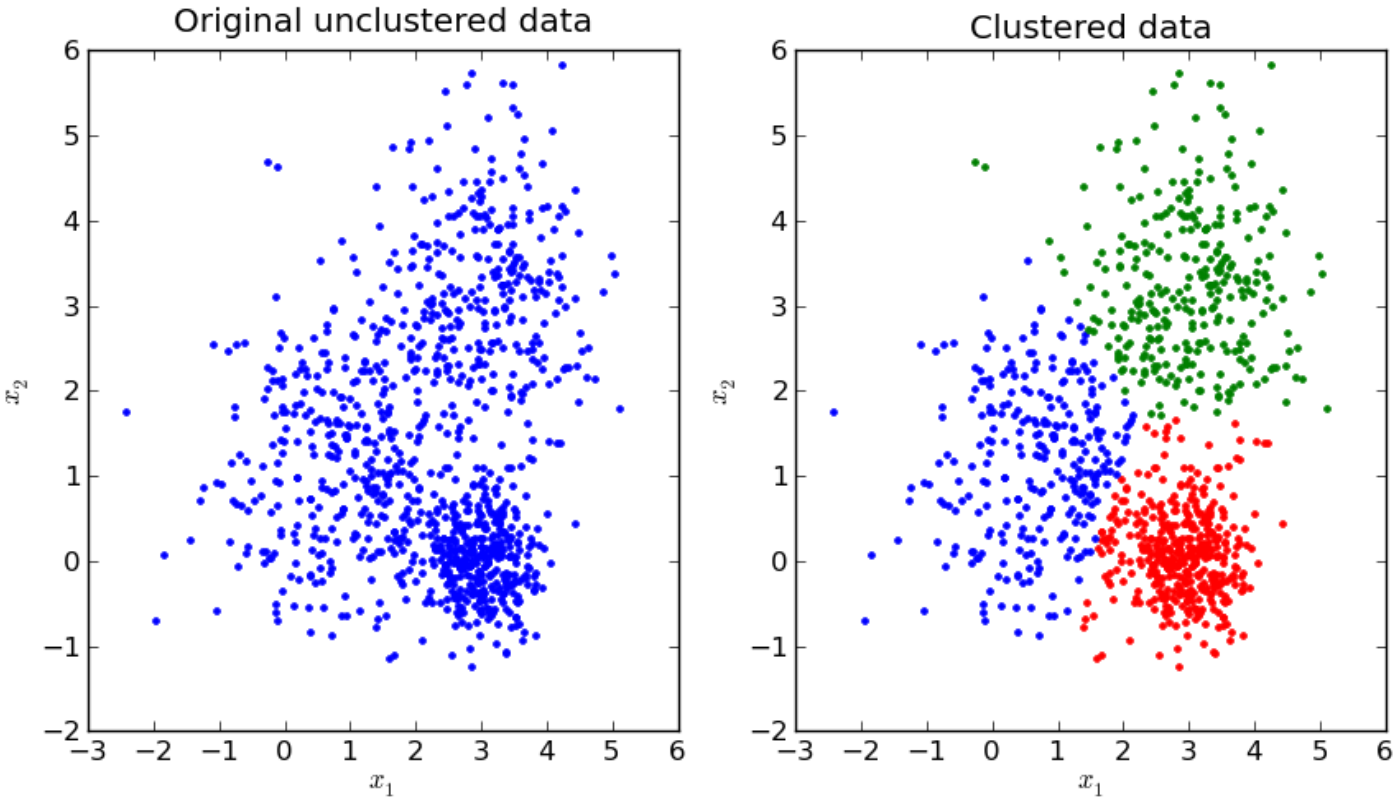
The clustering problem

- Group data objects into clusters by analyzing the similarities between objects following the features found in the data
- Keyworks: cluster analysis or data segmentation



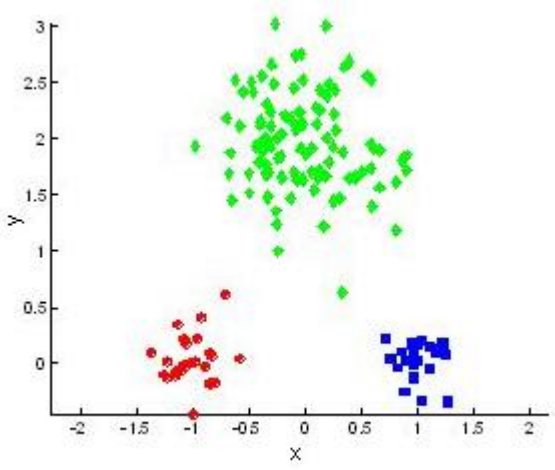
What is a cluster?

- A cluster is a collection of data objects that is
  - similar (or related) to one another within the same group
  - dissimilar (or unrelated) to the objects in other groups

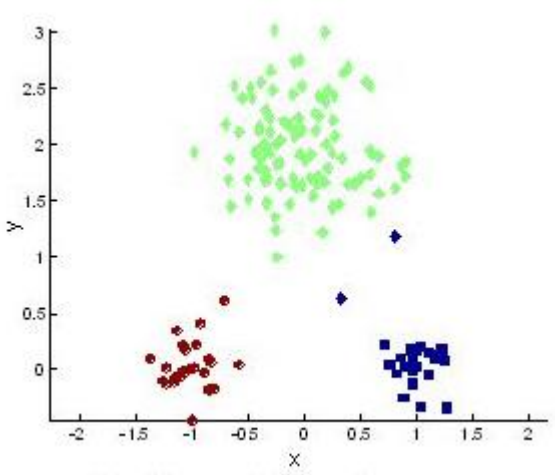


What is good clustering?

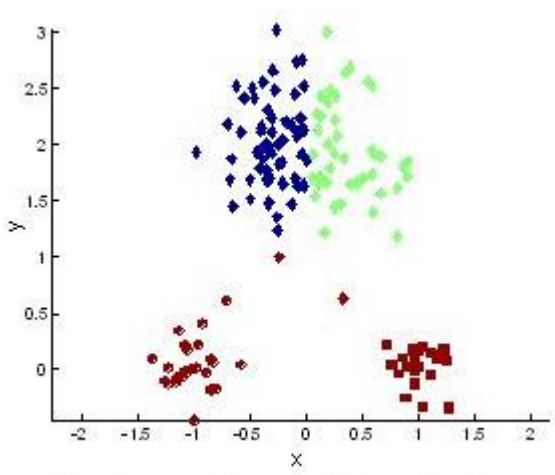
- A good clustering method will produce high quality clusters.
  - High intra-class similarity: cohesive within clusters
  - Low inter-class similarity: distinctive between clusters



Original points



Optimal clustering



Suboptimal clustering

k-means: Algorithm

- Input: A data set of n objects and the number of clusters k
- Output: A set of k clusters

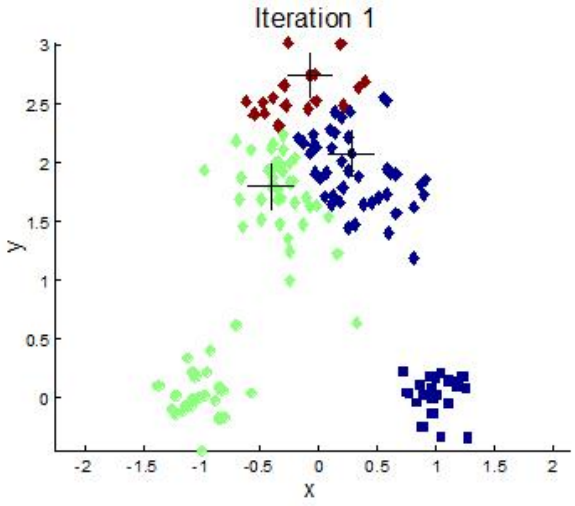
Arbitrarily pick k objects from D as the initial cluster centers

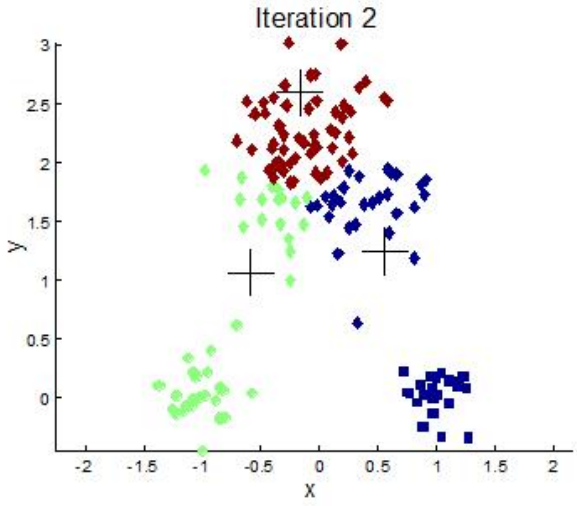
Repeat

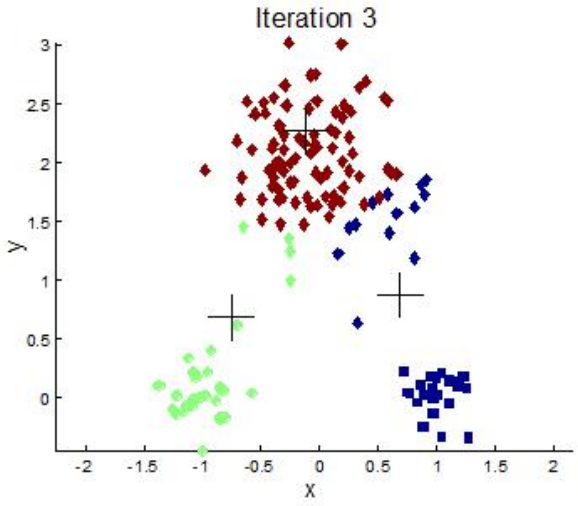
1. (Re)assign each object o to the cluster to which o is most similar, based on the distance from o to the mean value of the objects in that cluster.
2. Update the cluster by computing the mean value of the objects for each cluster.

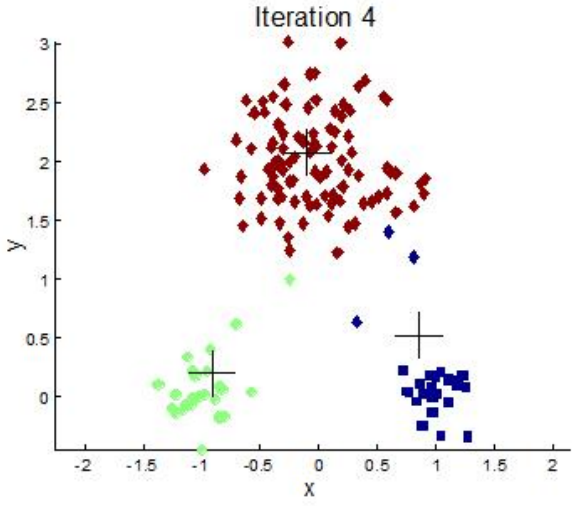
Until no change

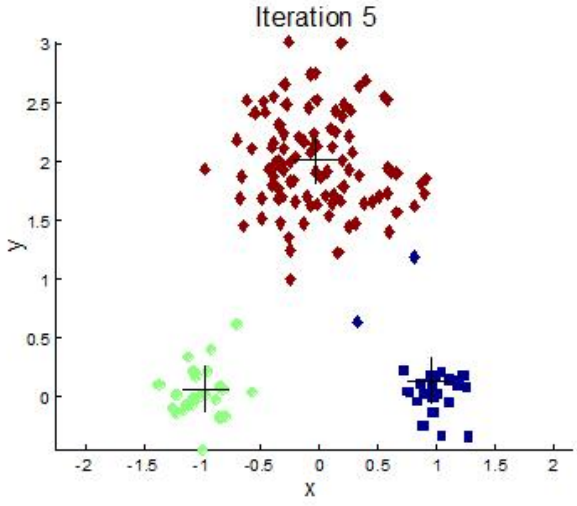
k-means: Group points into clusters

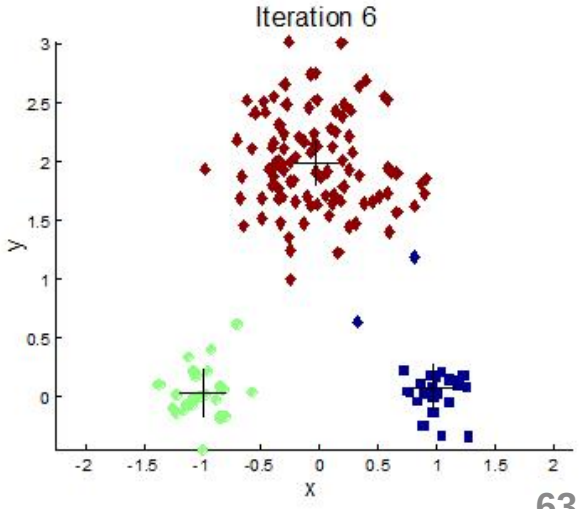








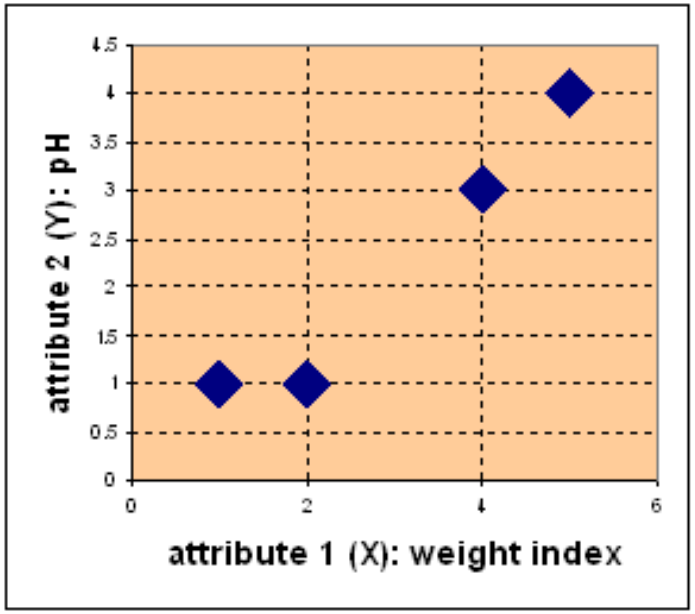




k-means: A numerical example

- Suppose we have several objects (4 types of medicines) and each object have two attributes, weight index and pH, as shown below.

Object	weight index	pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4



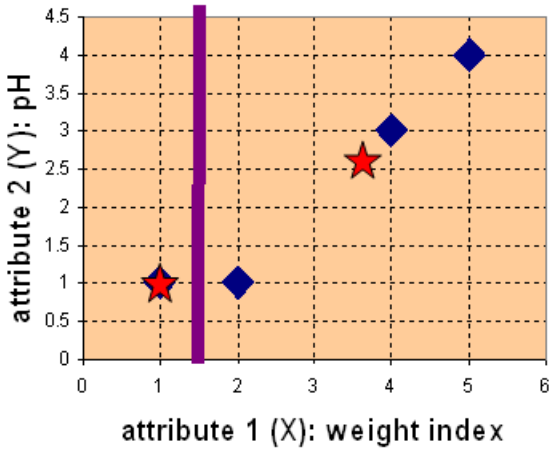
- Group the objects into $k = 2$ clusters based on the two features.
- Distance metric: Euclidean distance

k-means: A numerical example

- Assume that the initial centers are C_1 = A(1,1) and C_2 = B(2,1)

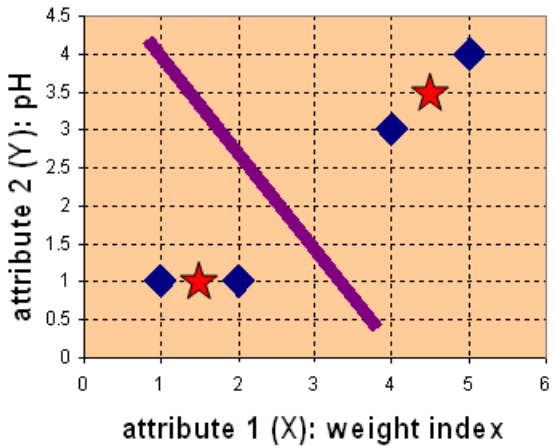
- First iteration
  - Cluster 1 = {A}
  - Cluster 2 = {B, C, D}
  - C_1 = (1, 1)
  - C_2 = (3.67,2.67)

Points	Distance to C_1	Distance to C_2
A(1,1)	0	1
B(2,1)	1	0
C(4,3)	3.606	2.828
D(5,4)	5	4.243



- Second iteration
  - Cluster 1 = {A, B}
  - Cluster 2 = {C, D}
  - C_1 = (1.5, 1)
  - C_2 = (4.5,3.5)

Points	Distance to C_1	Distance to C_2
A(1,1)	0	3.145
B(2,1)	1	2.357
C(4,3)	3.606	0.471
D(5,4)	5	1.886



k-means: A numerical example

- Third iteration
  - Cluster 1 = {A, B}
  - Cluster 2 = {C, D}
  - C_1 = (1.5, 1)
  - C_2 = (4.5,3.5)

Points	Distance to C_1	Distance to C_2
A(1,1)	0.5	4.301
B(2,1)	0.5	3.536
C(4,3)	3.202	0.707
D(5,4)	4.610	0.707

- There is no change in clusters from the second to the third iteration.
- k-means clustering has reached its stability $\rightarrow$ no more iteration required

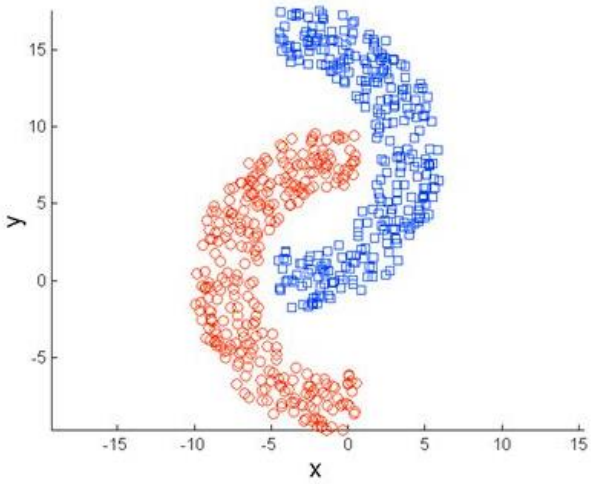
k-means: Algorithm efficiency

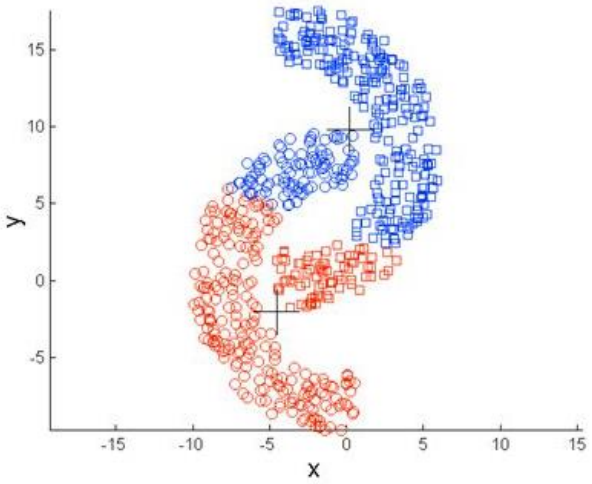
- Each cluster is represented by the center of the cluster.
- Efficiency: the algorithm complexity is $O(tkn)$
  - n: # objects, k: # clusters, and t: # iterations.
  - Normally, $k, t \ll n$.
- Often terminate at a local optimal
- Only applicable to objects in a continuous domain

k-means: Limitations

- Non-convex/non-rounded clusters

Original points

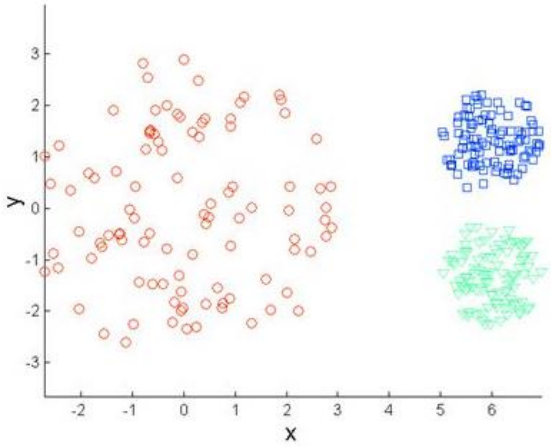


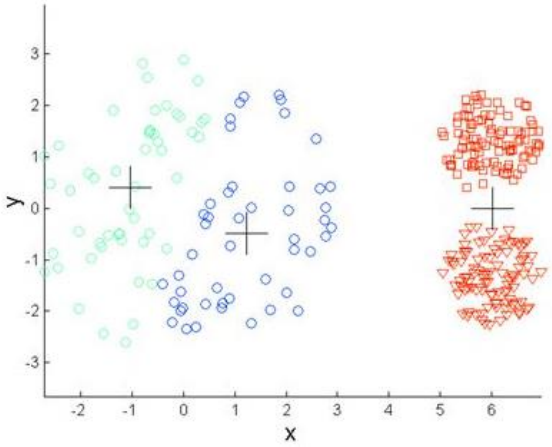


k-means (k = 2)

- Clusters with different densities

Original points

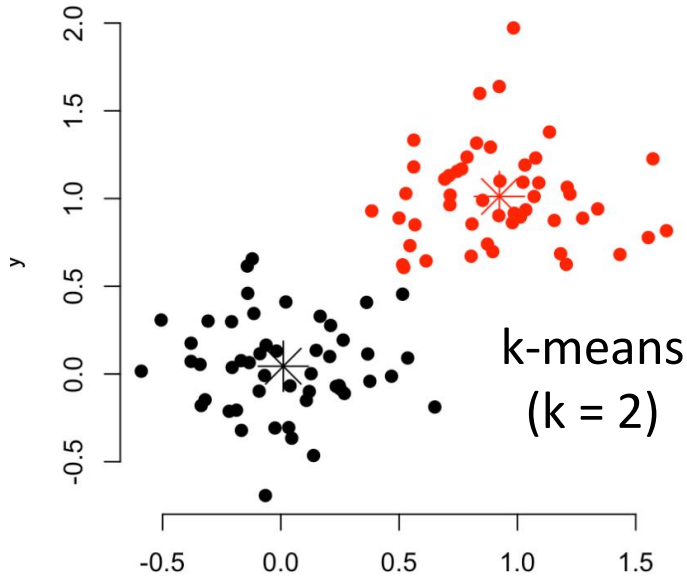


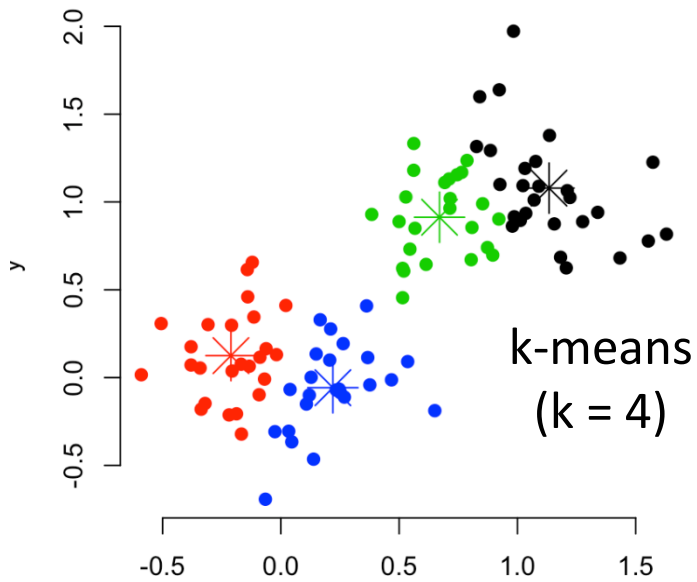


k-means (k = 3)

k-means: Algorithm efficiency

- The number of clusters k needs to be specified in advance.
  - There are ways to find the best k (e.g., Hastie et al., 2009)





- Sensitive to noisy data and outliers
  - An object with an extremely large value may distort the distribution.

Quiz 06: k-means clustering

Given a dataset including 8 types of fruits, which are categorized into 3 classes based on their sweetness and sourness (both from 1 to 10).

ID	Fruit	Sweetness	Sourness	Fruit Type
1	Lemon	1	9	Sour
2	Grapefruit	2	8	Sour
3	Orange	3	7	Sour
4	Cherry	6	4	Sweet
5	Banana	9	1	Sweet
6	Grapes	8	2	Sweet
7	Avocado	1	1	None
8	Strawberry	5	5	Sour

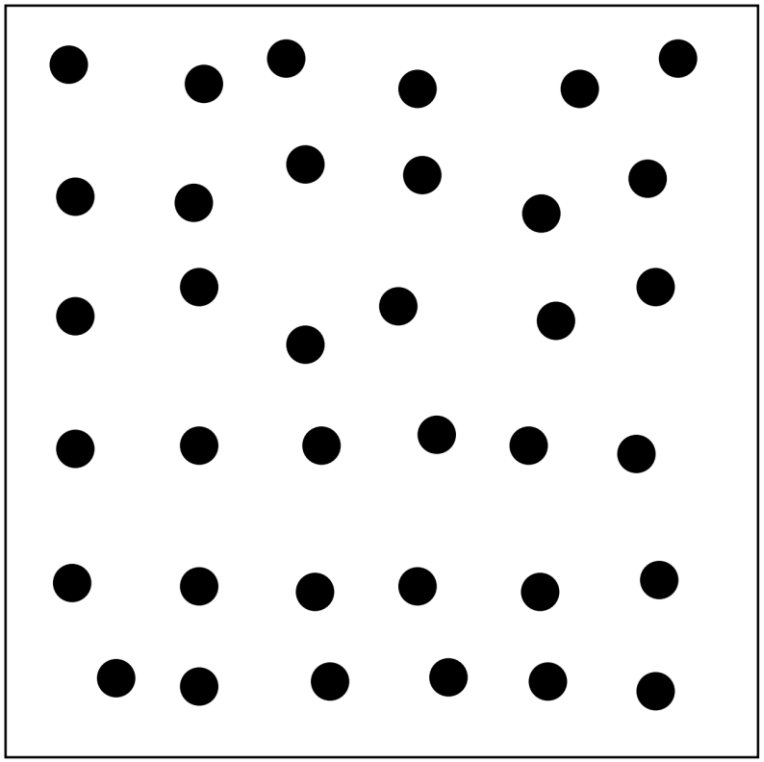
Partition the data into 3 clusters using k-means clustering with squared Euclidean distance and comment on the quality of the clusters.

Evaluation metrics for clustering

Clustering tendency assessment

- This determines whether a given data set has a non-random structure, which may lead to meaningful clusters.

This data set is uniformly distributed in 2-D data space.
A clustering algorithm may partition the points into groups, yet the groups will unlikely mean anything significant due to the uniform distribution of the data.



Hopkins statistics

- Let D be a dataset noted as sample of a random variable o
- Sample n points, $p_1 \ldots p_n$, uniformly from D.
- For each p_i, find its nearest neighbor $(1 \leq i \leq n)$ in D:

$$x_i = \min_{v \in D} \{\mathrm{dist}(p_i, v)\}$$

- Sample n points, $q_1 \ldots q_n$, uniformly from D.
- For each q_i, find its nearest neighbor in $D - q_i$:

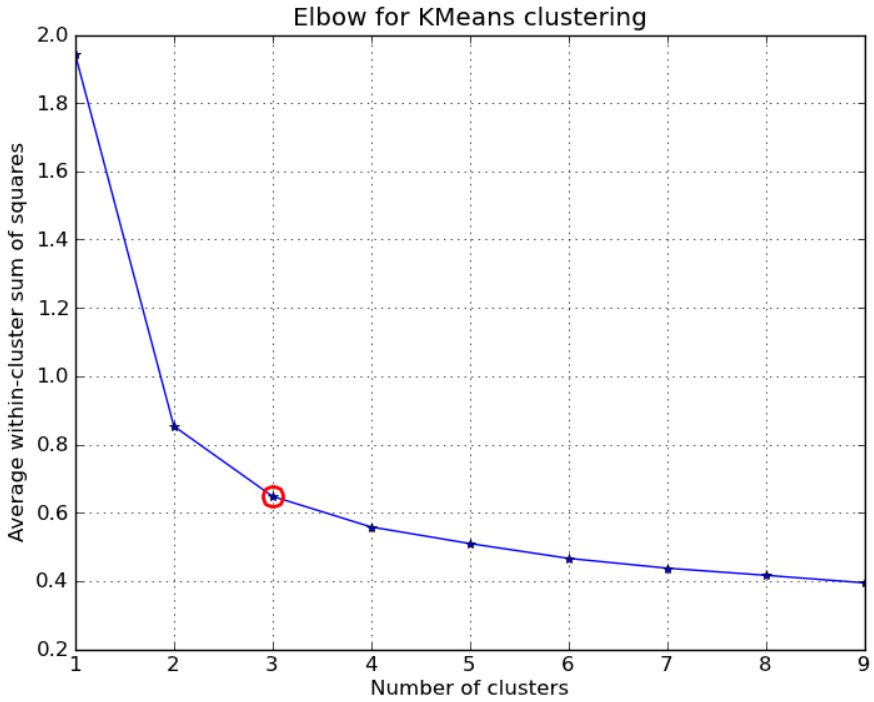
$$y_i = \min_{v \in D,\, v \neq q_i} \{\mathrm{dist}(q_i, v)\}$$

Hopkins statistics

- Hopkins statistics: statistically test spatial randomness, i.e., determine how far away o is from being uniformly distributed in the data space.
- Calculate the Hopkins statistic: $\boldsymbol{H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}}$
  - If D is uniformly distributed, $\sum x_i$ and $\sum y_i$ will be close to each other and H is close to 0.5.
  - If D is highly skewed, H is close to 0

Determine the number of clusters

- Empirical method: The number of clusters $\approx \sqrt{n}/2$ for a dataset of n points



- Elbow method: Use the turning point in the curve of sum of within cluster variance w.r.t the number of clusters

Determine the number of clusters

- Cross validation method
- Divide a given data set into m parts
- Use $m - 1$ parts to obtain a clustering model and the remaining part to test the quality of the clustering
- For each point in the test set, find the closest centroid, and use the sum of squared distance between all points in the test set and their closest centroids to measure how well the model fits the test set
- For any $k > 0$, repeat it m times, compare the overall quality measure to find the number of clusters that fits the data the best

Measure clustering quality

- Extrinsic: supervised, the ground truth is available
  - Compare a clustering against the ground truth using certain clustering quality measure
  - E.g., BCubed precision and recall metrics
- Intrinsic: unsupervised, i.e., the ground truth is unavailable
  - Evaluate the goodness of a clustering by considering how well the clusters are separated, and how compact the clusters are
  - E.g., Silhouette coefficient

Clustering quality: Extrinsic methods

- The measure $Q(C, C_g)$ evaluates the quality of a clustering C given the ground truth C_g.
- Q is good if it satisfies the following 4 essential criteria
  - Cluster homogeneity: the purer, the better
  - Cluster completeness: objects belonging to the same category in the ground truth should be assigned to the same cluster
  - Rag bag: putting a heterogeneous object into a pure cluster should be penalized more than putting it into a rag bag (i.e., “miscellaneous” or “other” category)
  - Small cluster preservation: splitting a small category into pieces is more harmful than splitting a large category into pieces

Clustering quality: Silhouette coefficient

- Assume the data have been clustered via any technique (e.g., k-means), into k clusters.
- For any point $i \in C_I$, compute $a(i) = \frac{1}{|C_I|-1} \sum_{j \in C_I, j \neq i} d(i,j)$
  - Where $|C_I|$ is the number of points belonging to cluster C_I, and $d(i,j)$ is the distance between data points i and j.
- Also for any point $i \in C_I$, define $b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j)$
- $a(i)$ measures how well i is assigned to its cluster
  - The smaller the value, the better the assignment.
- $b(i)$ implies how i matches its neighboring cluster.
  - The larger the value, the better the assignment.

Clustering quality: Silhouette coefficient

- The silhouette (value) of a point i is

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i), b(i)\}} \quad \text{if } |C_I| > 1$$

and $\quad s(i) = 0 \quad \text{if } |C_I| = 1$

- It is clear that $-1 \leq s(i) \leq 1$
- The silhouette coefficient for the maximum value of the mean $s(i)$ over the entire data

$$SC = \max_k \tilde{s}(k)$$

  - where $\tilde{s}(k)$ represents the mean $s(i)$ over all data for a specific number of clusters k.


...the end.