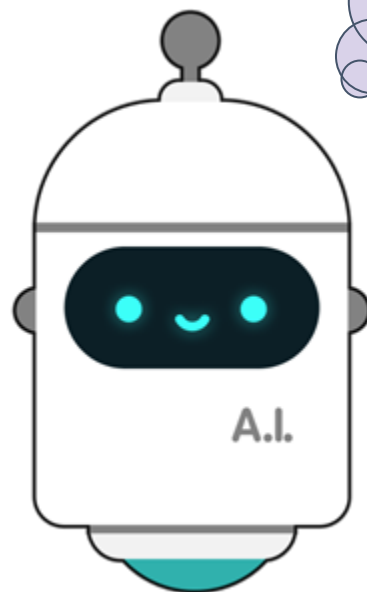


Machine Learning

Mental Model

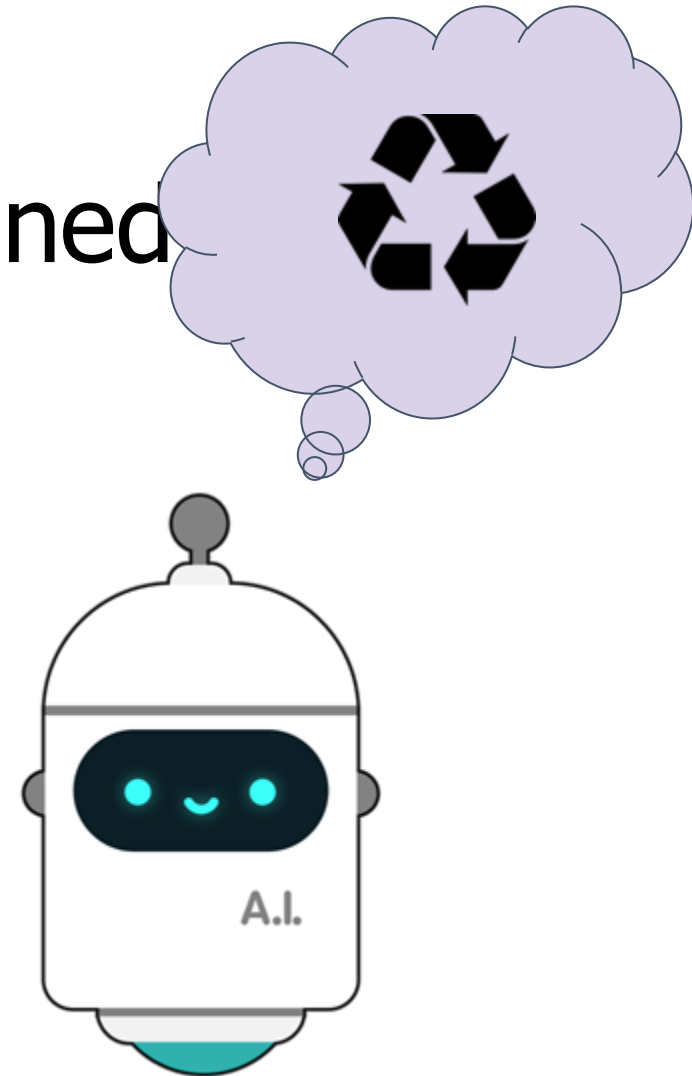


**Machine Learning
Model**



Key Vocabulary

- **Model** - a computer program designed to make a decision



Types of Machine Learning

Activity 1:

Read the following 4 paragraphs, and group similar paragraphs together

Paragraph A

- To many, Messi leading his country of Argentina to a World Cup is like a prerequisite for him to be regarded as the greatest. To some, Messi is called the best club footballer ever but hold the right to reserve the Best Ever talks until he holds that beautiful, shinny golden trophy in his hands.

Paragraph B

- Good recipe but needs lime juice (to be authentic) instead of lemon juice and a clove of crushed garlic. Tomato is optional, guacamole keeps better without the tomato. Also chopped jarred jalapeno works if a fresh pepper isn't on hand.

Paragraph C

- Messi has always had the most intense spotlight shone on him at World Cup finals. He has featured in four World Cups and this will be his fifth (and potentially last), which will see him become the first Argentinian player to do so. Scoring at every World Cup except the 2010, he's scored six goals and provided five assists in 19 appearances, and hopes to add to that record in Qatar.

Paragraph D

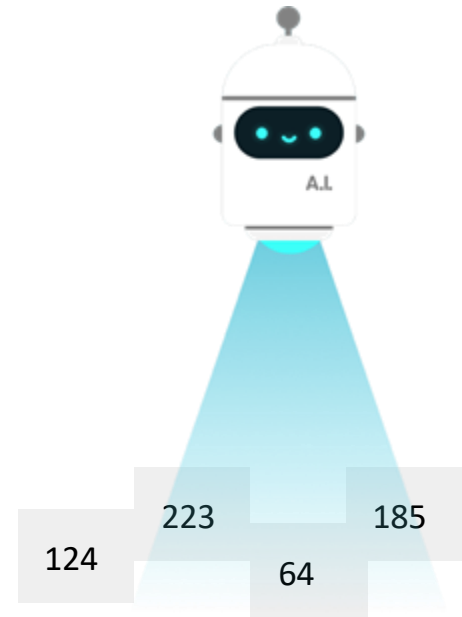
- We love guacamole and this is a good recipe, but definitely use lime juice and add some garlic. A tip to keep your guacamole from turning brown is to put the avocado pits back into the guac until you are ready to serve.

Discuss

What groupings did you create? Is there a pattern?

Key Vocabulary

- **Unsupervised Learning** - finding patterns in data that doesn't have any labels
- **Features** - the inputs that the model uses to make decisions



Types of Machine Learning

Activity 2:

Classify the words into the corresponding categories

To many, **Messi** leading his country of Argentina to a World Cup is like a prerequisite for him to be regarded as the greatest.

Students line up to ask Brock Purdy questions during a communication studies class in early November.

Key Vocabulary

- **Supervised Learning** - When a human trains a model to learn with examples.

Key Vocabulary

- **Training** - the process of giving examples to a model so it can learn
- **Testing** - the process of giving a new example to a model to predict

Key Vocabulary

- **Label** - the output you are trying to decide or predict with a model. It is also the outcomes that the model learns to associate with the input features during training.

Key Vocabulary


- **Classification** - task of predicting a discrete class label for an example.
- **Regression** - task of predicting a continuous quantity for an example.

Summary

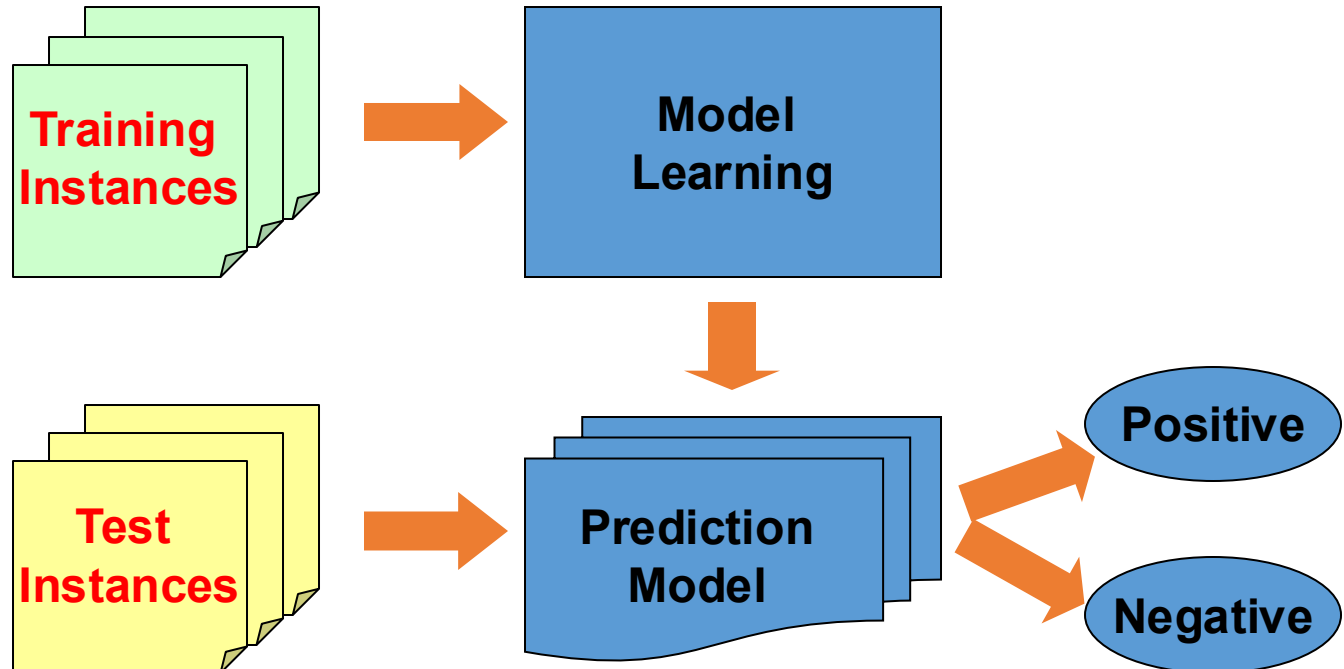
- Supervised learning

- Supervision: The training data such as observations or measurements are accompanied by **labels** indicating the classes which they belong to
- New data is classified based on the models built from the training set

Training Data with class label:

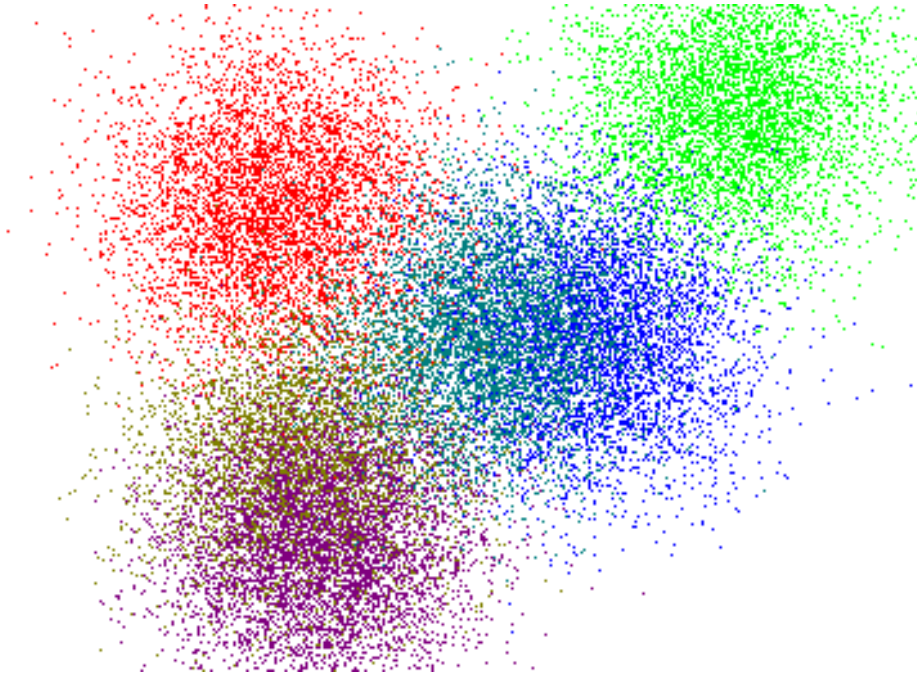


Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No



Summary

- Unsupervised learning (clustering)
 - The class labels of training data are **unknown**
 - Given a set of observations or measurements, establish the possible existence of classes or clusters in the data



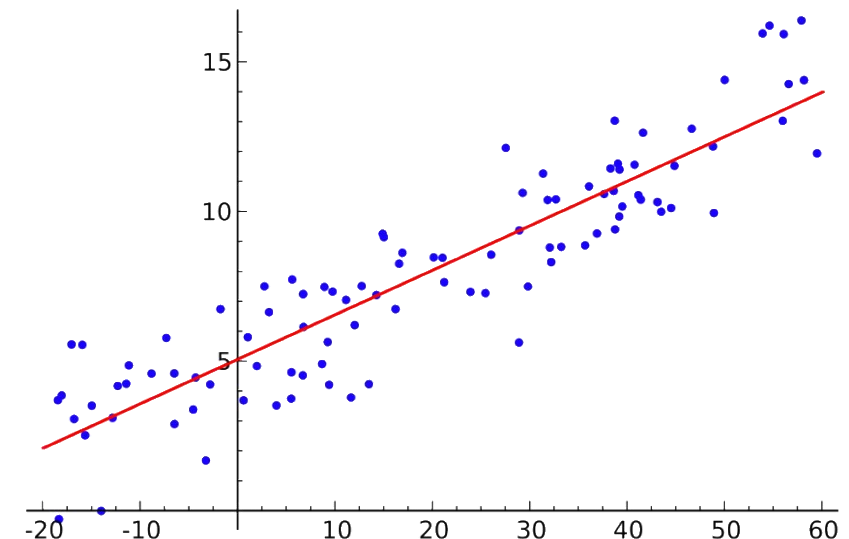
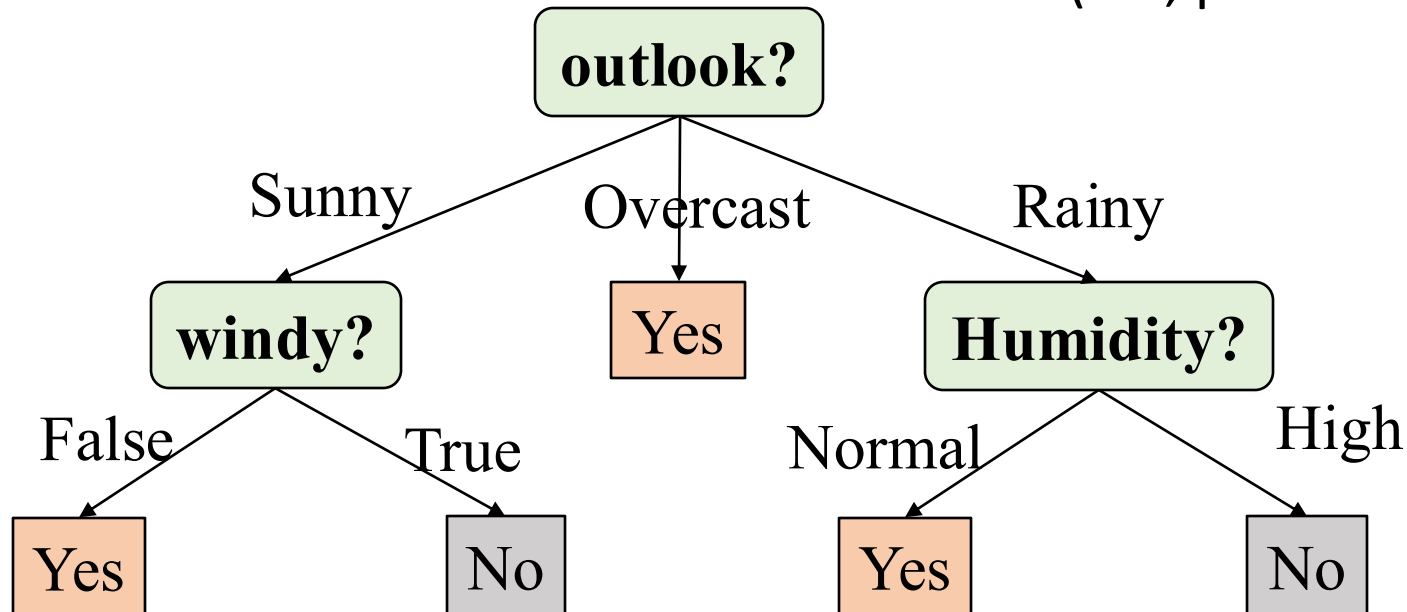
Summary

- **Classification**

- Predict **categorical class labels** (discrete or nominal)
- Construct a model based on the training set and the **class labels** (the values in a classifying attribute) and use it in classifying new data

- **Regression (Numeric prediction)**

- Model **continuous-valued functions** (i.e., predict unknown or missing values)



How to train AI to see and to write

- <https://youtu.be/y2BVTW09vck>
- <https://www.youtube.com/watch?v=d2UccTPnl4w&t=1s>

Supervised or Unsupervised? (or either?)

Next word prediction

Anomaly Detection

Image segmentation

Recommendation

Searching (like google)

Key Vocabulary

- **Generative** – models that can generate new data instances
- **Discriminative** – models that discriminate between different kinds of data instances

Generative or Discriminative? (or both?)

Translation

Face recognition

Plagiarism detection

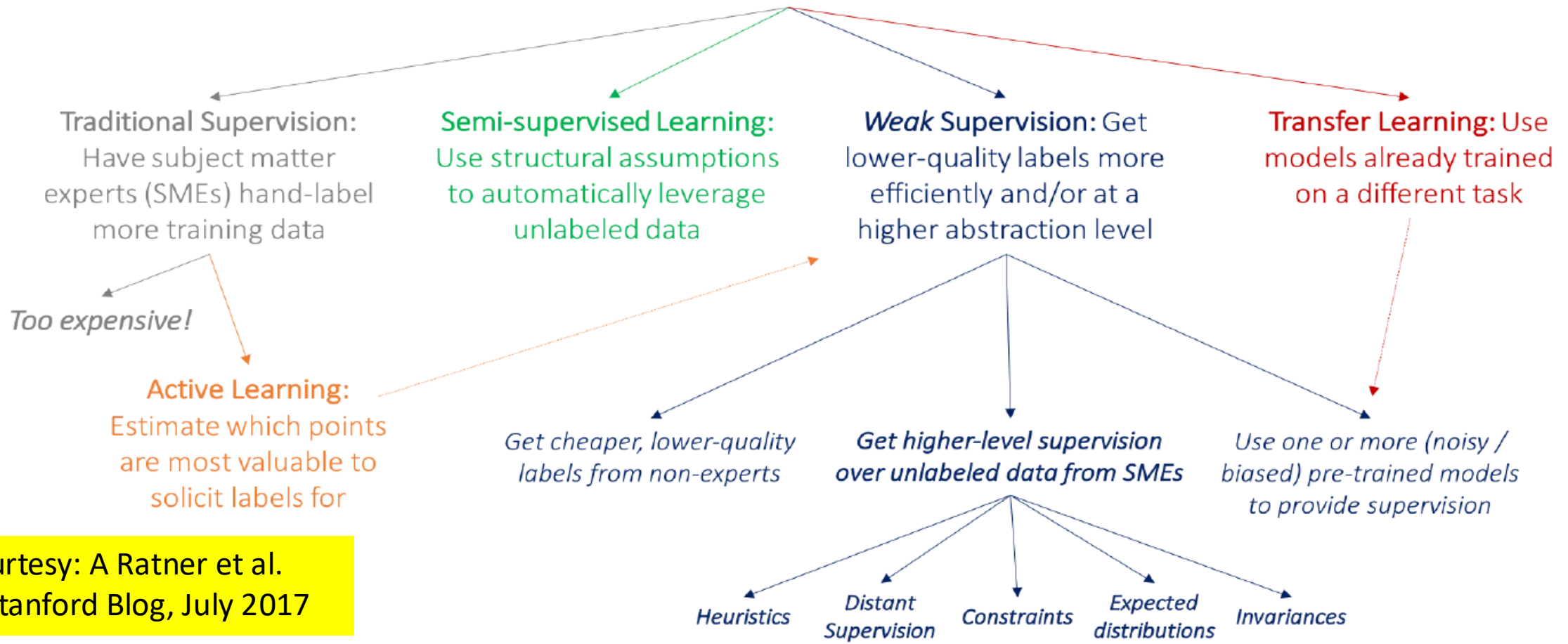
Function prediction

Generative vs. Discriminative Classifiers

- X: observed variables (features)
- Y: target variables (class labels)
- A generative classifier models $p(Y, X)$
 - It models how the data was "generated", "what is the likelihood this or that class generated this instance?" and pick the one with higher probability
- A discriminative classifier models $p(Y|X)$
 - It uses the data to create a decision boundary

Relationships Among Different Kinds of Supervisions

How to get more labeled training data?

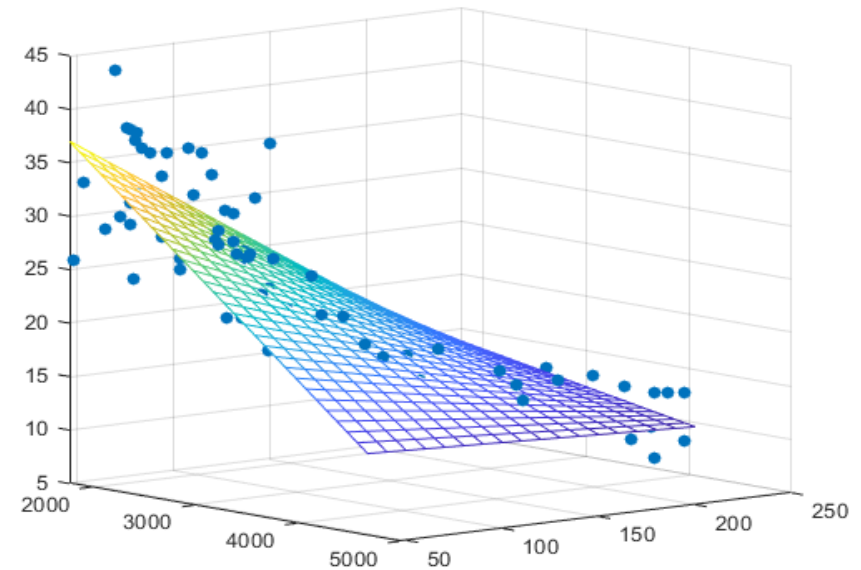
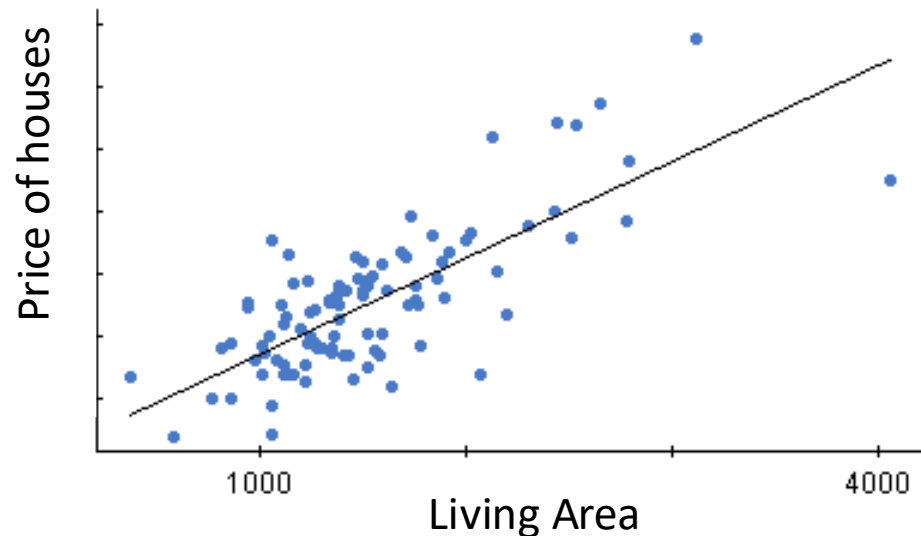


Courtesy: A Ratner et al.
@Stanford Blog, July 2017

Many areas of machine learning are motivated by the bottleneck of labeled training data, but are divided at a high-level by what information they leverage instead.

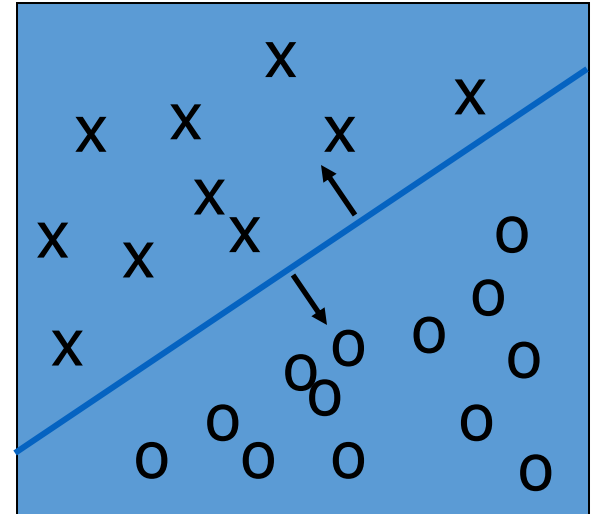
Linear Regression Problem: Example

- Mapping from independent attributes to **continuous value**: $x \Rightarrow y$
- {living area} \Rightarrow Price of the house
- {college; major; GPA} \Rightarrow Future Income



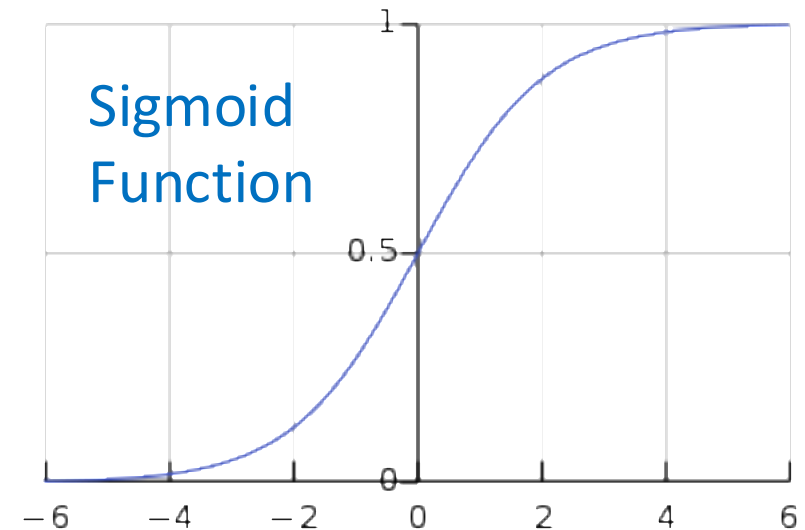
Linear Classifier: General Ideas

- Binary Classification
- $f(x)$ is a linear function based on the example's attribute values
 - The prediction is based on the value of $f(x)$
 - Data above the blue line belongs to class 'x' (i.e., $f(x) > 0$)
 - Data below blue line belongs to class 'o' (i.e., $f(x) < 0$)
- Classical Linear Classifiers
 - Logistic Regression
 - Linear Discriminant Analysis (LDA)
 - Perceptron
 - SVM



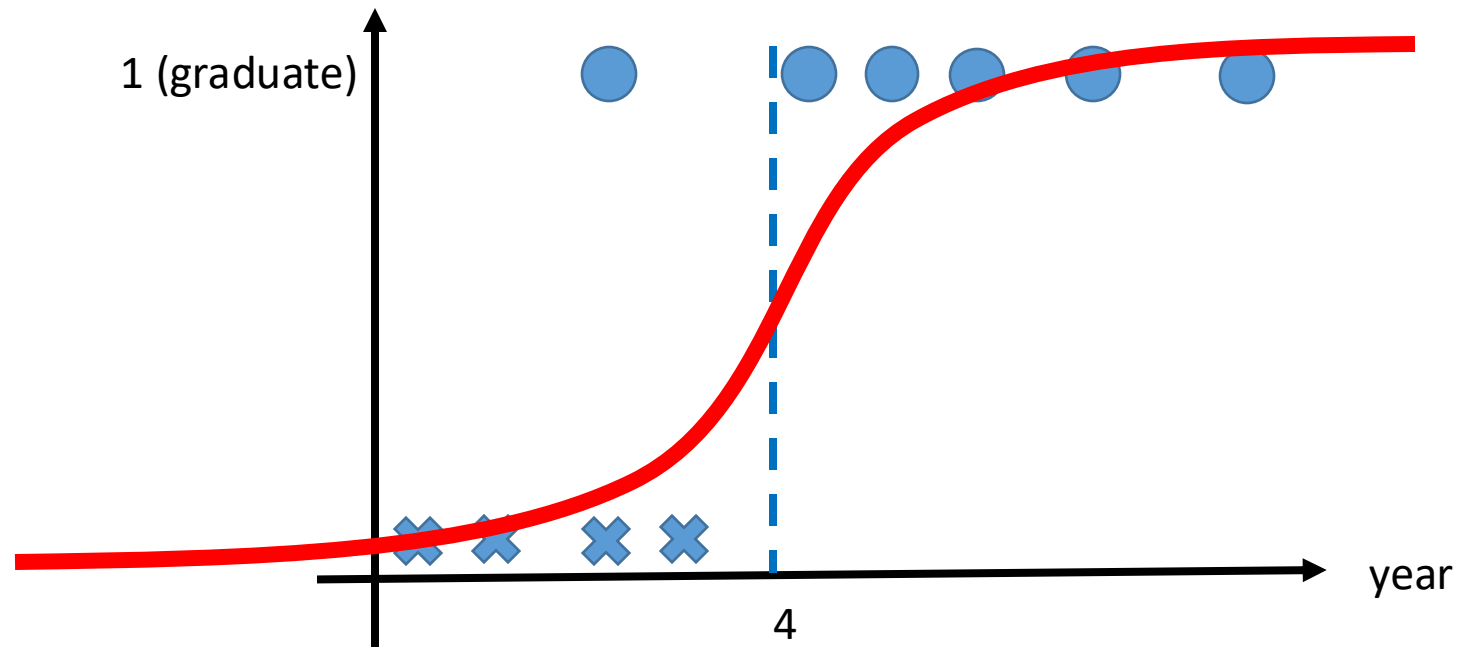
Logistic Regression: General Ideas

- How to solve “classification” problems by regression?
- Key idea of Logistic Regression
 - We need to transform the real value Y into a probability value $\in [0,1]$
- Sigmoid function (differentiable function) :
 - $\sigma(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{e^z+1}$
 - Projects $(-\infty, +\infty)$ to $[0, 1]$
 - Not only LR uses this function, but also neural network, deep learning
- The projected value change sharply around zero point
- $\ln \frac{y}{1-y} = w^T x + b$



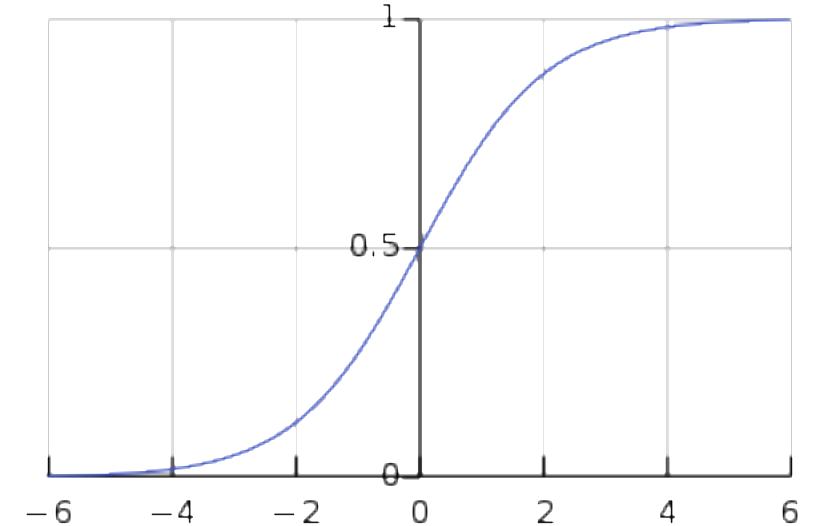
Logistic Regression: An Example

- Suppose we only consider the year as feature
 - Data points are converted by sigmoid function (“activation” function)



Logistic Regression: Model

- The prediction function to learn
- Probability that $Y=1$:
 - $p(Y = 1 | X = x; \mathbf{w}) = \text{Sigmoid}(w_0 + \sum_{i=1}^n w_i \cdot x_i)$
 - $\mathbf{w} = (w_0, w_1, w_2, \dots, w_n)$ are the parameters
- A single data object with attributes x_i and class label y_i
 - Suppose the probability of $p(\hat{y}_i = 1 | x_i, w) = p_i$, then $p(\hat{y}_i = 0 | x_i, w) = 1 - p_i$
 - $p(\hat{y}_i = y_i) = p_i^{y_i} (1 - p_i)^{1-y_i}$
- Maximum Likelihood Estimation
 - $L = \prod_i p_i^{y_i} (1 - p_i)^{1-y_i} = \prod_i \left(\frac{\exp(w^T x_i)}{1 + \exp(w^T x_i)} \right)^{y_i} \left(\frac{1}{1 + \exp(w^T x_i)} \right)^{1-y_i}$



Linear Regression VS. Logistic Regression

- Linear Regression

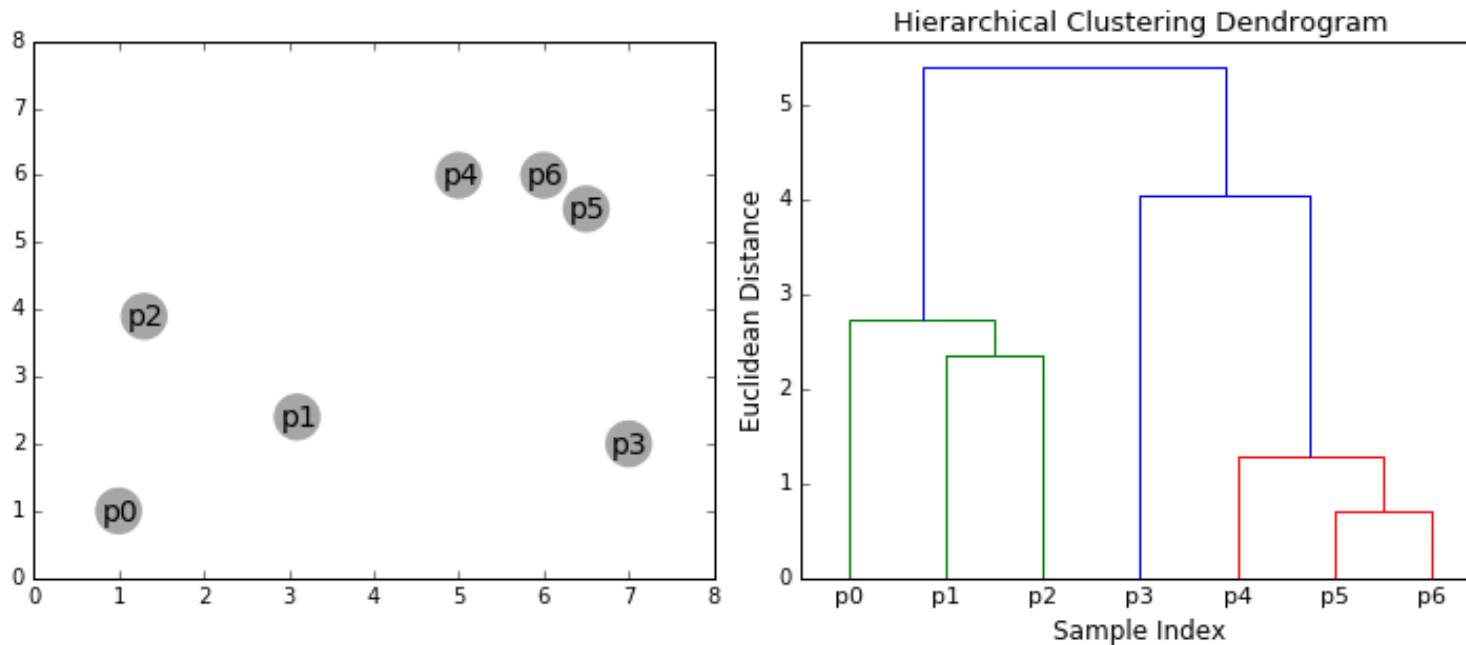
- Y: Continuous Value $\in [-\infty, +\infty]$
- $Y = W^T X + b$
- Often used in value prediction problems

- Logistic Regression

- Y: A discrete value from m classes
- $P(Y = C_i) \in [0,1]$ and $\sum_{i=1}^m P(Y = C_i) = 1$
- Often used in classification problems

Dendrogram: Shows How Clusters are Merged

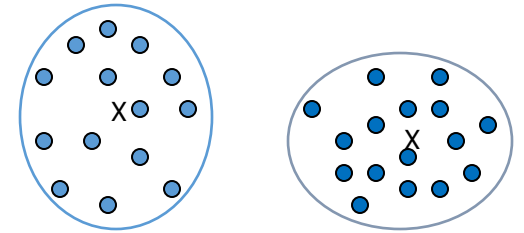
- Dendrogram: Decompose a set of data objects into a tree of clusters by multi-level nested partitioning
- A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster



Hierarchical clustering generates a dendrogram (a hierarchy of clusters)

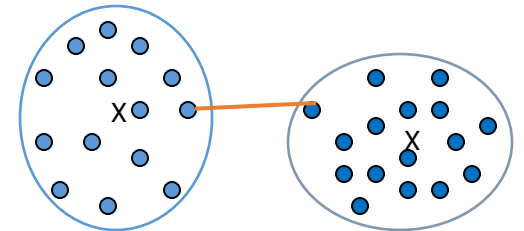
Agglomerative Clustering Algorithm

- AGNES (AGglomerative NESTing) (Kaufmann and Rousseeuw, 1990)
 - Continuously merge nodes that have the least dissimilarity
 - Eventually all nodes belong to the same cluster



□ Single link (nearest neighbor)

- The similarity between two clusters is the similarity between their most similar (nearest neighbor) members
- Local similarity-based: Emphasizing more on close regions, ignoring the overall structure of the cluster
- Capable of clustering non-elliptical shaped group of objects
- Sensitive to noise and outliers

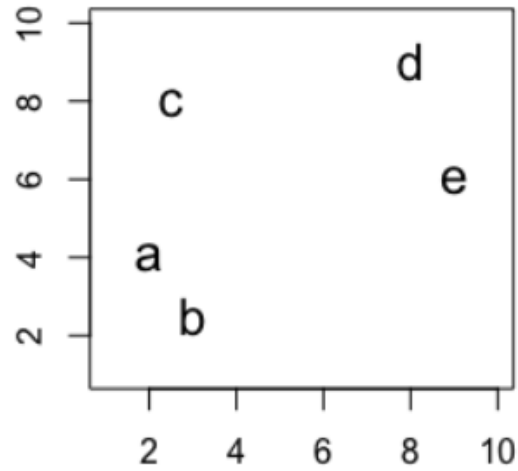


Agglomerative Clustering: Example

Distance Matrix

- 2-D Data points

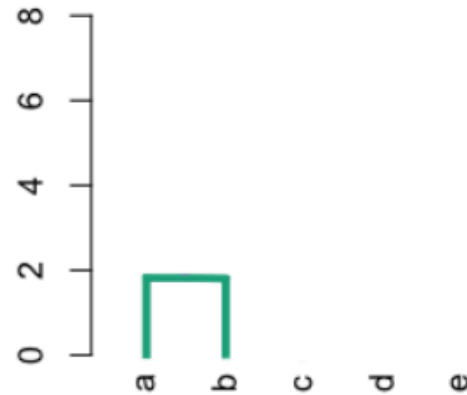
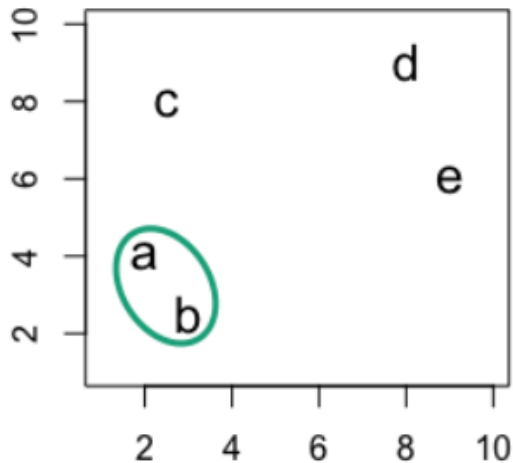
- a(2,4)
- b(3,2)
- c(2,8)
- d(8,9)
- e(9,6)



	a	b	c	d	e
a	0				
b	2.2	0			
c	4	6.1	0		
d	7.8	8.6	6.1	0	
e	7.3	7.2	7.3	3.2	0

Agglomerative Clustering: Example

Distance Matrix



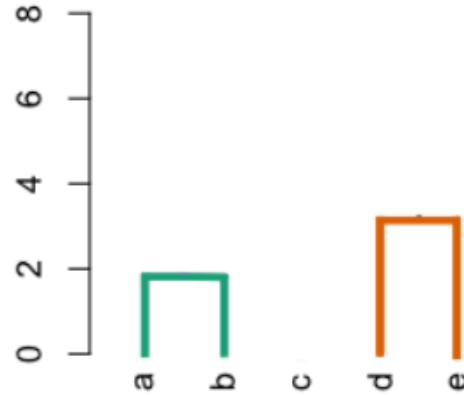
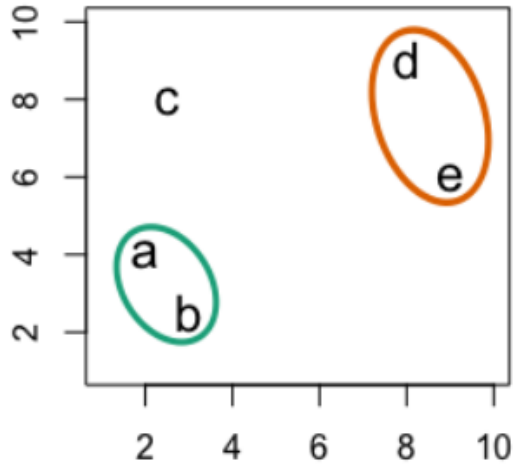
	a,b	c	d	e
a,b	0			
c	4	0		
d	7.8	6.1	0	
e	7.2	7.3	3.2	0



- Update distance
 - $\text{Distance}((a,b), c) = \min(\text{Distance}(a,c), \text{Distance}(b,c)) = \min(4, 6.1) = 4$
 - $\text{Distance}((a,b), d) = \min(\text{Distance}(a,d), \text{Distance}(b,d)) = \min(7.8, 8.6) = 7.8$
 - $\text{Distance}((a,b), e) = \min(\text{Distance}(a,e), \text{Distance}(b,e)) = \min(7.3, 7.2) = 7.2$

Agglomerative Clustering: Example

Distance Matrix



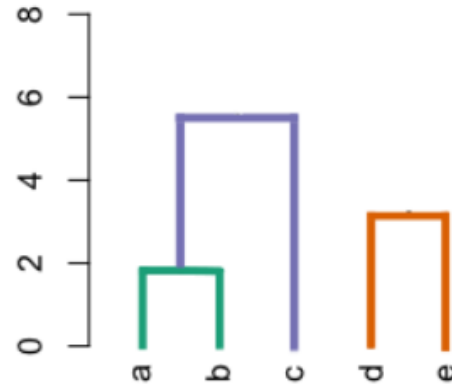
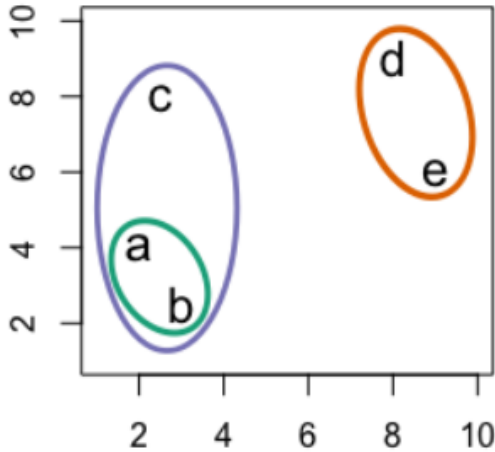
	a,b	c	d,e
a,b	0		
c	4	0	
d,e	7.2	6.1	0



- Update distance
 - $\text{Distance}((d,e), (a,b)) = \min(\text{Distance}(d,(a,b)), \text{Distance}(e,(a,b))) = 7.2$
 - $\text{Distance}((d,e), c) = \min(\text{Distance}(d,c), \text{Distance}(e,c)) = 6.1$

Agglomerative Clustering: Example

Distance Matrix



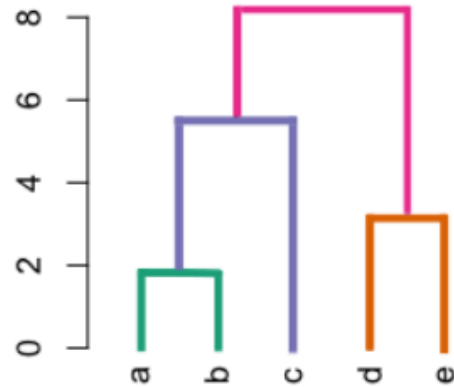
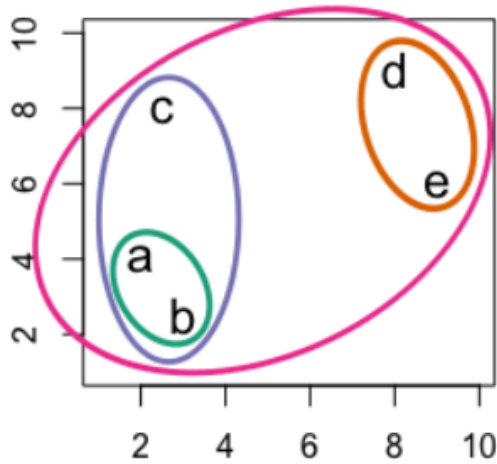
	a,b,c	d,e
a,b,c	0	
d,e	6.1	0



- Update distance
 - $\text{Distance}((d,e), (c,(a,b))) = \min(\text{Distance}((d,e), (a,b)), \text{Distance}((d,e), c)) = 6.1$

Agglomerative Clustering: Example

Distance Matrix



	a,b,c,d,e
a,b,c,d,e	0