

Quick Primer on Unsupervised Learning

Vineet Kumar

August 31, 2020

Notes on K-means Clustering

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

- There are multiple variants of K-means

Notes on K-means Clustering

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

- There are multiple variants of K-means
- Finding the optimal solution to the k-means is NP-hard

Notes on K-means Clustering

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

- There are multiple variants of K-means
- Finding the optimal solution to the k-means is NP-hard
- Fix `max_iter` and ϵ beforehand for stopping

Notes on K-means Clustering

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

- There are multiple variants of K-means
- Finding the optimal solution to the k-means is NP-hard
- Fix `max_iter` and ϵ beforehand for stopping
- Not possible to obtain **non-convex cluster**

Handling Outliers (noise) in K-means

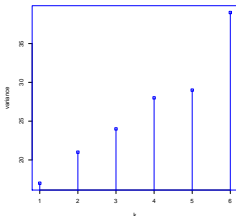
Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

- Consider cluster with unusually high variance

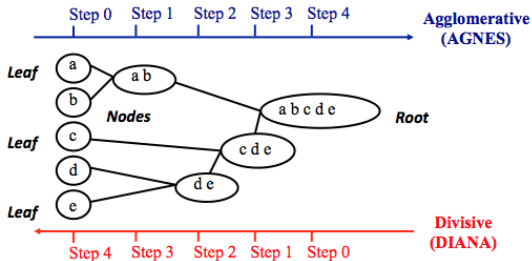


- Now in selected cluster **6** consider each variable, compute its mean
- for every point of cluster if its value $>$ mean put that point in cluster **6.I** otherwise in **6.II**

Hierarchical clustering

Possible to obtain non-convex cluster
Essentially we have two versions –

- Agglomerative clustering (AGNES)
- Divisive Clustering Analysis (DIANA)



Distance Measures for Items

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

All clusters are valid but they have different meanings. Hence, applying clustering **may not always** provide meaningful results.

Distance Measures for Items

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

All clusters are valid but they have different meanings. Hence, applying clustering **may not always** provide meaningful results.

Metric	Description
manhattan	Absolute distance between the two vectors (l_1 norm)
euclidean	Usual square distance between the two vectors (l_2 norm)
minkowski	l_p norm
maximum	Maximum distance between x and y i.e. $\ a - b\ _\infty = \max_i a_i - b_i $
correlation	$1 - r$ where r is the Pearson or Spearman correlation

Table: Distance Measures for distance between items

Distance Measure for Clusters – Linkage Function

Quick Primer
on
Unsupervised
Learning

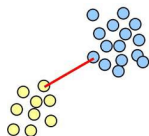
Kmeans

Hierarchical
Clustering

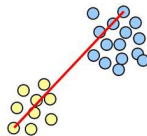
Recommender
Systems

The linkage function tells the distance between clusters.

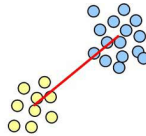
Linkage Type	Description	Type of Clusters	Transformation
Single	$f = \min(d(x, y))$	Long and loose clusters	Invariant
Complete	$f = \max(d(x, y))$	More compact clusters	Invariant
Average	$f = \text{average}(d(x, y))$	Similar to complete linkage	Affected



single-link



complete-link



average-link

Figure: Schematics of three linkage types of hierarchical clustering

Single and Complete

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

Single linkage

Single linkage defines the distance between clusters as the distance between the closest two points. It will have a tendency to combine, at relatively low thresholds, observations linked by a series of close intermediate observations (**chaining**). The clusters produced **violate** the **compactness** property – all observations within each cluster should be similar.

Complete linkage

This is opposite extreme. Two groups G and H are considered close only if all of the observations in their union are relatively similar. It will tend to produce compact clusters with small diameters may **violate** the **closeness** property. That is, observations assigned to a cluster can be much closer to members of other clusters than their own cluster.

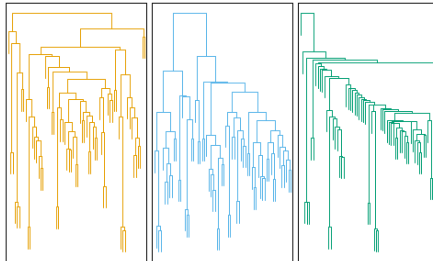
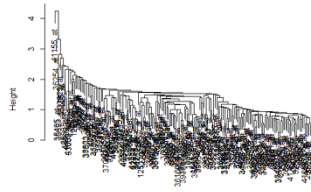
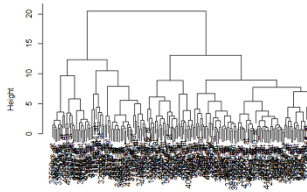
Visual Inspection

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems



Visual Inspection

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

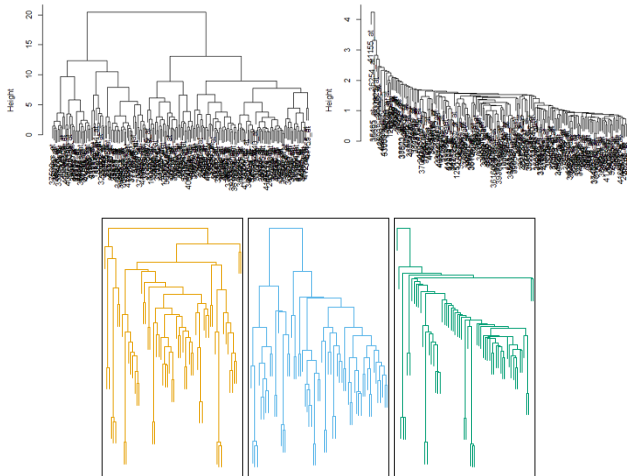


Figure: complete linkage → compact clusters • single linkage → long stringy clusters

No Free Lunch

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

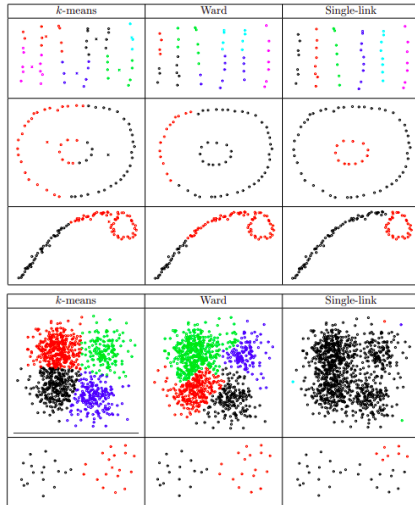


Figure: Some cases where k-means and where single-link method perform better

Hierarchical clustering – numerical example

Quick Primer
on
Unsupervised
Learning

clustering of distances (in kilometers) between some Italian cities using single-linkage ¹

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0



¹Example Credits: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html

Kmeans

Hierarchical
Clustering

Recommender
Systems

Hierarchical clustering – numerical example

Quick Primer
on
Unsupervised
Learning

clustering of distances (in kilometers) between some Italian cities using single-linkage ¹

Kmeans

Hierarchical
Clustering

Recommender
Systems

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

	BA	FI	MI/TO	NA	RM
BA	0	662	877	255	412
FI	662	0	295	468	268
MI/TO	877	295	0	754	564
NA	255	468	754	0	219
RM	412	268	564	219	0



¹Example Credits: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html

Hierarchical clustering – numerical example – II

Quick Primer
on
Unsupervised
Learning

clustering of distances (in kilometers) between some Italian cities using single-linkage

Kmeans

Hierarchical
Clustering

Recommender
Systems

	BA	FI	MI/TO	NA/RM
BA	0	662	877	255
FI	662	0	295	268
MI/TO	877	295	0	564
NA/RM	255	268	564	0



Hierarchical clustering – numerical example – II

Quick Primer
on
Unsupervised
Learning

clustering of distances (in kilometers) between some Italian cities using single-linkage

Kmeans

Hierarchical
Clustering

Recommender
Systems

	BA	FI	MI/TO	NA/RM
BA	0	662	877	255
FI	662	0	295	268
MI/TO	877	295	0	564
NA/RM	255	268	564	0

	BA/NA/RM	FI	MI/TO
BA/NA/RM	0	268	564
FI	268	0	295
MI/TO	564	295	0



Hierarchical clustering – numerical example – III

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

clustering of distances (in kilometers) between some Italian cities using single-linkage

	BA/FL/NA/RM	MI/TO
BA/FL/NA/RM	0	295
MI/TO	295	0



Hierarchical clustering – numerical example – III

Quick Primer
on
Unsupervised
Learning

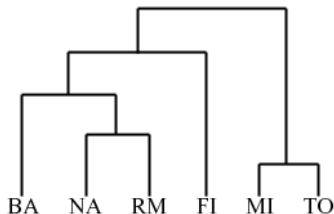
Kmeans

Hierarchical
Clustering

Recommender
Systems

clustering of distances (in kilometers) between some Italian cities using single-linkage

	BA/FI/NA/RM	MI/TO
BA/FI/NA/RM	0	295
MI/TO	295	0



Example of Clustering in Real Life

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

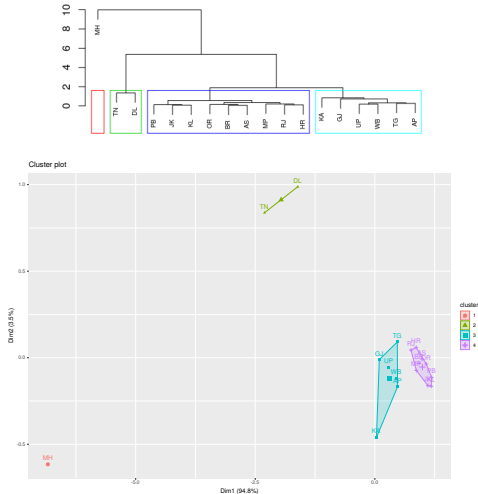


Figure: Clustering # of Covid Cases (statewise)

Algorithmic Analysis of H/C

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

Running Time

Need to build **minimal** spanning tree

If edge weight is known use Kruskal $\mathcal{O}(n \log n)$, if not known use prims $\mathcal{O}(n^2)$ Hence, **MST based approach** – $\mathcal{O}(n^2)$

Space Analysis

Need to store similarity matrix in memory $\mathcal{O}(n^2)$

Applying single linkage to pathological images of $512\text{px} \times 512\text{px}$ (prohibitively high computational cost)

Do not scale well. AGNES can never undo what was done last.

Algorithmic Analysis of H/C

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

Running Time

Need to build **minimal** spanning tree

If edge weight is known use Kruskal $\mathcal{O}(n \log n)$, if not known use prims $\mathcal{O}(n^2)$ Hence, **MST based approach** – $\mathcal{O}(n^2)$

Space Analysis

Need to store similarity matrix in memory $\mathcal{O}(n^2)$

Applying single linkage to pathological images of $512\text{px} \times 512\text{px}$ (prohibitively high computational cost)

Do not scale well. AGNES can never undo what was done last.

you should be **intensely skeptical** of any clustering results

Power Law of Nature

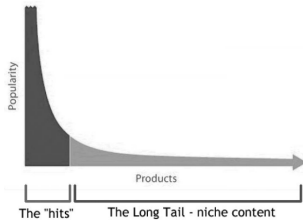
Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

C. Anderson, The Long Tail: <https://www.wired.com/2004/10/tail/>



Open Long-Tailed Datasets

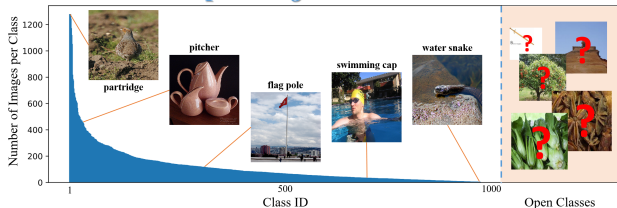


Figure: Long Tailed Distribution in Nature (follows power law)

Formalization & Steps of Recommender Systems

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

X = set of Customers

S = set of Items, we define a utility function u such that

$$u : X \times S \rightarrow R$$

Key steps are:

- 1 Gathering “known” ratings (Ask people to rate items, otherwise Learn ratings from user actions)
- 2 Extrapolating Utilities : Utility matrix is **sparse**. Most people have not rated most items

Cold Start Problem

New items have no ratings, New users have no history

- 3 Evaluating the systems

Content Based Filtering

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

Recommend items to customer x similar to previous items rated highly by x

Given user profile x and item profile i , estimate

$$\cos(x, i) = \frac{x \cdot i}{||x|| \cdot ||i||}$$

Movie#	1	2	3	4	5
Actor	B	B	B	A	A
Ratings	1	2	4	3	5

Mean Rating:

$$(1+2+4+3+5)/5=3$$

Normalize the ratings to capture negative sentiments:

Actor A normalized rating:

$$(3-3), (5-3) = \{0, 2\}$$

Actor B normalized rating:

$$(1-3), (2-3), (4-3) = \{-2, -1, 1\}$$

Profile Weight

$$\mathbf{A} := (0 + 2)/2 = 1$$

$$\mathbf{B} := (-2 - 1 + 1)/3 = -2/3$$

(Dis)Advantages of Content Based System

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

Advantages

- No need for data on other users No cold-start or sparsity problems
- Able to recommend to users with unique tastes
- Able to recommend new & unpopular items → No first-rater problem

(Dis)Advantages of Content Based System

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

Advantages

- No need for data on other users No cold-start or sparsity problems
- Able to recommend to users with unique tastes
- Able to recommend new & unpopular items → No first-rater problem

Disadvantages

- Finding the appropriate features is hard
- Recommendations for new users (new user profile)
- Cannot recommend items outside user's content profile People might have multiple interests
- **Unable to exploit judgments of other users**

Collaborative Filtering

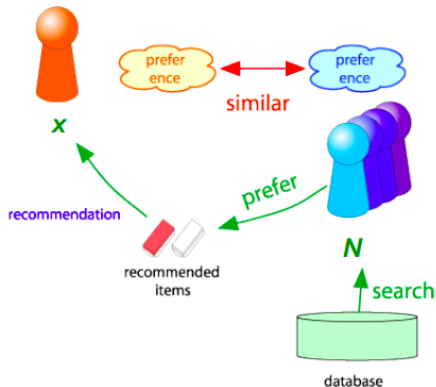
Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

Leveraging information/ratings of other users in own recommendation



Need to find out some quantification of similarity

Quantification of Similarity

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

$$r_x = [* , _ , _ , * , ***] \text{ and } r_y = [* , _ , ** , ** , _]$$

Jaccard Similarity – Set

$$r_{a,b} = \frac{|r_A \cap r_B|}{|r_A \cup r_B|} = \frac{1}{5}. \text{ **Problem:** Ignores the value of the rating}$$

Cosine Similarity – Point

$$\begin{aligned} r_{a,b} &= \cos(r_A, r_B) \\ \text{vec1} &= c(1,0,0,1,3) \text{ \& } \text{vec2} = c(1,0,2,2,0) \\ > \cosine(\text{vec1}, \text{vec2}) &:= 0.3015113 \end{aligned}$$

Problem: Treats missing ratings as negative

Centered Cosine similarity – Pearson correlation

Subtract the mean of each user and recompute cosine similarity

Understanding Similarity

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>
<i>A</i>	4	5		5	1		3	2
<i>B</i>		3	4	3	1	2	1	
<i>C</i>	2		1	3		4	5	3

Figure: Utility Matrix

The ratings, on a 1–5 star scale, of eight items, a through h, by three users A, B, and C. Answer the following:

- 1 Compute the Jaccard distance between each pair of users.
 $SIM(A,B) = 4/8$; $SIM(A,C) = 3/8$; $SIM(B,C) = 4/8$

Understanding Similarity

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

- Compute cosine distance between each pair of users.

```
U = np.array([4,5,0,5,1,0,3,2,0,3,4,3,1,2,1,0,2,0,1,3,0,4,5,3]).reshape(3,8)
```

```
# user ratings are rows of Utility matrix  
A = U[0]  
B = U[1]  
C = U[2]
```

```
def cosine(X,Y):  
    return np.dot(X,Y)/(np.sqrt(np.dot(X,X)*np.dot(Y,Y)))
```

```
print 'A,B:', cosine(A,B)  
print 'A,C:', cosine(A,C)  
print 'B,C:', cosine(B,C)
```

```
A,B: 0.601040764009  
A,C: 0.614918693812  
B,C: 0.513870119777
```


Understanding Similarity

- Normalize the matrix by subtracting from each nonblank entry the average value for its user.

```
A_norm = map(lambda x: x-np.mean(A) if x>0 else 0, A)
B_norm = map(lambda x: x-np.mean(B) if x>0 else 0, B)
C_norm = map(lambda x: x-np.mean(C) if x>0 else 0, C)
```

```
U_norm = np.array([A_norm,B_norm,C_norm])
```

```
U_norm
```

```
array([[ 1.5 ,  2.5 ,  0. ,  2.5 , -1.5 ,  0. ,  0.5 , -0.5 ],
       [ 0. ,  1.25,  2.25,  1.25, -0.75,  0.25, -0.75,  0. ],
       [-0.25,  0. , -1.25,  0.75,  0. ,  1.75,  2.75,  0.75]])
```

- Using the normalized matrix from above part compute the cosine distance between each pair of users.

```
print 'A,B:', cosine(A_norm,B_norm)
print 'A,C:', cosine(A_norm,C_norm)
print 'B,C:', cosine(B_norm,C_norm)
```

```
A,B: 0.546504040851
A,C: 0.163408291384
B,C: -0.312561520424
```



Item Item 2 NEAREST CF – Example²

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarichal
Clustering

Recommender
Systems

		users												
		12	11	10	9	8	7	6	5	4	3	2	1	
movies			4		5			5			3		1	1
	3		1	2			4			4	5			2
			5	3	4		3		2	1		4	2	3
			2			4			5		4	2		4
	5		2					2	4	3	4			5
			4			2			3		3		1	6
			- unknown rating					- rating between 1 to 5						

²Example Courtesy:

<http://www.mmids.org/mmids/v2.1/ch09-recsys1.pdf>

Item Item CF – Example

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

		users													
		12	11	10	9	8	7	6	5	4	3	2	1		
movies			4		5			5	?		3		1	1	
	3		1	2			4			4	5			2	
			5	3	4		3		2	1		4	2	3	
			2			4			5		4	2		4	
	5		2					2	4	3	4			5	
			4			2			3		3		1	6	



- estimate rating of movie 1 by user 5

Item Item CF – Example

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarichal
Clustering

Recommender
Systems

movies	users												
	12	11	10	9	8	7	6	5	4	3	2	1	
		4		5			5	?		3		1	1
	3	1	2			4			4	5			2
		5	3	4		3		2	1		4	2	3
		2			4			5		4	2		4
	5	2					2	4	3	4			5
		4			2			3		3		1	6
													sim(1,m)
													1.00
													-0.18
													<u>0.41</u>
													-0.10
													-0.31
													<u>0.59</u>

Neighbor selection:

Identify movies similar to
movie 1, rated by user 5

Here we use Pearson correlation as similarity:

1) Subtract mean rating m_i from each movie i

$$m_1 = (1+3+5+5+4)/5 = 3.6$$

row 1: [-2.6, 0, -0.6, 0, 0, 1.4, 0, 0, 1.4, 0, 0.4, 0]

2) Compute cosine similarities between rows

Item Item CF – Example

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarichal
Clustering

Recommender
Systems

		users													
		12	11	10	9	8	7	6	5	4	3	2	1		
movies			4		5			5	?		3		1	1	sim(1,m)
	3		1	2			4			4	5			2	1.00
			5	3	4		3		2	1		4	2	<u>3</u>	<u>0.41</u>
			2			4			5		4	2		4	-0.10
	5		2					2	4	3	4			5	-0.31
			4			2			3		3		1	<u>6</u>	<u>0.59</u>

Compute similarity weights:

$s_{1,3}=0.41$, $s_{1,6}=0.59$

.....

Item Item CF – Example

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarichal
Clustering

Recommender
Systems

		users											
		12	11	10	9	8	7	6	5	4	3	2	1
movies			4		5			5	2.6		3		1
	3		1	2			4			4	5		
			5	3	4		3		2	1		4	2
			2			4			5		4	2	
	5		2					2	4	3	4		
			4			2			3		3		1

Predict by taking weighted average:

$$r_{15} = (0.41 \cdot 2 + 0.59 \cdot 3) / (0.41 + 0.59) = 2.6$$

(Dis)Advantages of Collaborative Systems

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

item-item filtering outperforms user-user

(Dis)Advantages of Collaborative Systems

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems

item-item filtering outperforms user-user because item-item similarity is more meaningful

Advantages

Works for any kind of item No feature selection needed

Disadvantages

- Need enough users in the system to find a match – coldstart
- The user/ratings matrix is sparse. Hard to find users that have rated the same items
- Cannot recommend an item that has not been previously rated (First rater issue)
- Tends to recommend popular items (Popularity bias)

Yann Lecun's Cake Theory at NIPS 2016



■ "Pure" Reinforcement Learning (cherry)

- ▶ The machine predicts a scalar reward given once in a while.
- ▶ **A few bits for some samples**

■ Supervised Learning (icing)

- ▶ The machine predicts a category or a few numbers for each input
- ▶ Predicting human-supplied data
- ▶ **10→10,000 bits per sample**

■ Unsupervised/Predictive Learning (cake)

- ▶ The machine predicts any part of its input for any observed part.
- ▶ Predicts future frames in videos
- ▶ **Millions of bits per sample**



■ (Yes, I know, this picture is slightly offensive to RL folks. But I'll make it up)

References

Quick Primer
on
Unsupervised
Learning

Kmeans

Hierarchical
Clustering

Recommender
Systems



Distances between Clustering, Hierarchical Clustering 2009. <http://www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pdf>



PSU : Statistical Analysis of Genomics Data
<https://online.stat.psu.edu/stat555/node/85>



Mining of Massive Datasets *Jure Leskovec* et al.
<http://infolab.stanford.edu/~ullman/mmds/bookL.pdf>



Elements of Statistical Learning *Trevor Hastie* et al.
<https://web.stanford.edu/~hastie/Papers/ESLII.pdf>



Pattern Recognition *C.A Murthy* ISI Kolkata NPTEL
<https://nptel.ac.in/courses/106/106/106106046/>