

CS1951A: Data Science

Lecture 15: Unsupervised learning

Lorenzo De Stefani Spring 2022

Today

- Supervised vs. Unsupervised Learning
- Clustering with K-Means
- The Expectation-Maximization paradigm
 - Soft K-Means
 - EM for weighting weak experts in mail spam detection

Supervised vs. Unsupervised Learning

- Supervised: Traning data has explicit labels
 - Sentiment analysis—review text -> star ratings
 - Image tagging—image -> caption
- Unsupervised: Traning data does not have explicit labels
 - Clustering—find groups similar customers
 - Dimensionality Reduction—find features that differentiate individuals

Many variations

- Semi Supervised—Combining large amounts of unlabelled with smaller amounts of labelled (pretraining)
- Weakly/Distantly Supervised—using noisy labels or partial labels (bootstrapping, automaticallylabeled data)
- Reinforcement Learning—label on the result of a sequence of actions, but not on each action (games, robotics)

Supervised vs. Unsupervised Learning

- Supervised: Traning data has explicit labels
 - Sentiment analysis—review text -> star ratings
 - Image tagging—image -> caption
- Unsupervised: Traning data does not have explicit labels
 - Clustering—find groups similar customers
 - Dimensionality Reduction—find features that differentiate individuals

Unsupervised Learning

• "Finding structure in data" (vs. predicting labels)

Learning the distribution of the independent variables

- In data science, this is typically for "exploratory analysis".
- Or for preprocessing/featurizing
- In ML, right now, used extensively for "pretraining"
 - (e.g. autoencoding, dimensionality reduction, language modeling*)

Clustering Scenario

- Find groups of customers with similar tastes
 - Find topics within a set of news articles
 - Find genres within a library of music
 - Extrapolating: make predictions about your new setting based on behavior of similar settings
- Think of it as partition/assign costumers to groups in which they fit well

Clustering



tempo

Lorenzo De Stefani - CSCI 1951A Data Science - Spring'22

Clustering



- Each cluster has a center:
 - Points are assigned to the cluster whose center is the closest
 - Ties are broken arbitrarily or by default
- To assign the clusters correctly we essentially need to find good centers

We start by randomly guessing the initial placement of the centers





tempo

For each cluster compute a new center by averaging the coordinates of the points currently allocated to the center



Assign each point to closest center among the new ones



Compute new centers again by averaging the coordinates

3/23/22



Re-assign each point to the cluster with the closest center



Re-assign each point to the cluster with the closest center



- Re-assign each point to the cluster with the closest center
- When the centers and the cluster assignments do not change anymore, we have convergence

```
define parameters: K, max_iter, min_diff
iter = 0
change = inf
means = [random() for _ in range(K)]
while iter < max_iter and change > min_diff:
    update_assignments()
    compute_new_means()
    change = max_i(dist(new_mean_i, old_mean_i))
    iter += 1
```

"Hyperparameters" (i.e. not model parameters)

```
define parameters: K, max iter, min diff
iter = 0
change = inf
means = [random() for in range(K)]
while iter < max iter and change > min diff:
    update assignments()
    compute new means()
    change = max i(dist(new mean i, old mean i))
    iter += 1
```

How many clusters we want to find

```
define parameters: K, max iter, min diff
iter = 0
change = inf
means = [random() for in range(K)]
while iter < max iter and change > min diff:
    update assignments()
    compute new means()
    change = max i(dist(new mean i, old mean i))
    iter += 1
```

When to stop iterating: maximum number of iterations or minimum threshold of improvement

 Useful to prevent situations when convergence would take too long

```
define parameters: K, max_iter, min_diff

iter = 0
change = inf
means = [random() for _ in range(K)]
while iter < max_iter and change > min_diff:
    update_assignments()
    compute_new_means()
    change = max_i(dist(new_mean_i, old_mean_i))
    iter += 1
```

Initialization: Randomly guess what the means are

- Purely random can lead to cold start
- If we have prior/domain knowledge we can use it for a better initial guess

```
define parameters: K, max_iter, min_diff
iter = 0
change = inf
means = [random() for _ in range(K)]
while iter < max_iter and change > min_diff:
    update_assignments()
    compute_new_means()
    change = max_i(dist(new_mean_i, old_mean_i))
    iter += 1
```

Repeat/iterate until your hyperparameters say to stop

• Convergence will trigger mindiff criterion

```
define parameters: K, max iter, min diff
iter = 0
change = inf
means = [random() for in range(K)]
while iter < max iter and change > min diff:
    update assignments()
    compute new means()
    change = max i(dist(new mean i, old mean i))
    iter += 1
```

Assign each point to its closest center

```
define parameters: K, max iter, min diff
iter = 0
change = inf
means = [random() for in range(K)]
while iter < max iter and change > min diff:
    update assignments()
    compute new means()
    change = max i(dist(new mean i, old mean i))
    iter += 1
```

Recompute the center to be the mean of the points assigned to each cluster

```
define parameters: K, max iter, min diff
iter = 0
change = inf
means = [random() for in range(K)]
while iter < max iter and change > min diff:
    update assignments()
    compute new means()
    change = max i(dist(new mean i, old mean i))
    iter += 1
```

Question time !

What is the "loss" that we are trying to minimize here?

- (a) Number of clusters
- (b) Distance of points to their respective clusters
- (c) Distance between clusters
- (d) Probability of observed data

Question time !

Is this a good objective?

(a)Yes(b)No(c) Depends on the application

Computing the distances

- Different scales/ranges in the features being represented may skew our evaluations
 - Imagine data points with features (age, income)
 - If we simply compute the Euclidian distance, we will skew heavily towards income
- Two possible solutions:
 - Normalize the ranges of the dimension to a fixed set range
 - Opportunely weight different features according to their range
 - Weights can also be used to emphasize/deemphasize the impact/importance of some of the features







In the extreme: as many clusters as points

- Each points serves as the center of its own cluster
- Each points has zero distance from the center of its cluster
- Is this really a good idea?



- Elbow point # of clusters such that it is possible to assign points with a "small enough" mean distance from the centers
- Represents an accuracy/complexity trade-off
- More Clusters \rightarrow higher accuracy and complexity
- Less Clusters \rightarrow lower accuracy and complexity



Techniques to fix the number of clusters:

- Intuition
- Domain Expertise
- Iteration

Silhouette technique

- A method to evaluate the quality of your clustering based on the consistency/similarity of the points in each cluster
 - Points *a*, *b*, ..., *i*, *j*, ...
 - For each point *i*, let *C_i* be its assigned cluster
- For each point compute

$$a(i)=rac{1}{|C_i|-1}\sum_{j\in C_i, i
eq j}d(i,j)$$

Average distance from points in its own cluster

•
$$a(i) = 0$$
 if $|C_i| = 1$

$$b(i) = \min_{k
eq i} rac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

Average distance from points in closest other cluster

• $b(i) = \infty$ if there is no other cluster

Silhouette technique

The silhouette score of each point is:

$$s(i)=rac{b(i)-a(i)}{\max\{a(i),b(i)\}}$$

- $-1 \le s(i) \le 1$
- The closest s(i) is to 1 the better it fits in its assigned cluster
- For s(i) to be close to 1, we must have b(i) >> a(i)
- If s(i) is close to -1, it would be more appropriate if i was clustered in its closest neighboring cluster.
- The mean of the s(i)'s over all points of a cluster is a measure of how tightly grouped all the points in the cluster are.

Silhouette technique

- The mean of the s(i)'s over all data of the entire dataset, called the silhouette coefficient is a measure of how appropriately the data have been clustered.
- If there are too many or too few clusters used in the clustering algorithm (e.g. k-means), some of the clusters will have silhouettes close to zero
- We considering multiple possible values k for the number of clusters we can pick

$$k^* = \max_{\{k_1, k_2, \dots\}} \frac{1}{n} \sum s(i)$$

- A very general framework for non-parametric learning algorithms
- Two main phases repeated iteratively:
 - Expectation: given the current belief of the correct properties of the data distribution/parameters of the model, estimate the likelihood of the observed data
 - Maximization: update the parameters of believed distribution/model to maximize the probability

Cold start: randomly pick a starting point

- Values of parameters of the model
- Initial naïve belief on the model/distribution

randomly initialize params

```
while not converged:
    data = estimate likelihood(params)
```

```
params = maximize likelihood(data)
```



```
Maximization Step: adjust the the parameters to maximize the expectation of the data
```

```
randomly initialize params
while not converged:
    data = estimate_likelihood(params)
    params = maximize_likelihood(data)
```

EM and clustering

We can modify K-means clustering to include properties of the EM design (Soft K-Means)

• Start with a random guess of the centers



tempo

Expectation (E) step:

- We associate to each point a discrete distribution of belonging in each cluster
 - For example, a point belong in a cluster with probability inversely proportional to its distance to the centers
- We assign points to clusters randomly according to such probabilities

EM and clustering



Maximization (M) step:

- Compute the new centers of the clusters as the means of the points currently assigned to each of them
- These would be the centers that would maximize the probability of the observed outcome

Iterate E and M step until convergence is reached or stopping conditions are met

EM K Means

The critical difference will be the way we update the assignments (the E-steps)

```
define parameters: K, max iter, min diff
iter = 0
change = inf
means = [random() for in range(K)]
while iter < max iter and change > min diff:
    update assignments()
    compute new means()
    change = max i(dist(new mean i, old mean i))
    iter += 1
```

EM for confidence in user reviews

Goal: Find "true" labels despite noisy annotations from workers...

	worker1	worker2	worker3	worker4	worker5
email1					
email2					
email3					
email4					
email5					

EM for confidence in user reviews



	w1	w2	w3	w4	w5	
email1	spam	not	not	not	spam	
email2	spam	spam	spam	spam	spam	
email3	not	spam	not	not	spam	
email4	spam	spam	spam	spam	not	
email5	spam	not	not	not	spam	

	spam	not
email1	?	?
email2	?	?
email3	?	?
email4	?	?
email5	?	?

w1	spam	not		
is spam	?	?		
is not	?	?		
w2	spam	not		
is spam	?	?		

w2	spam	not
is spam	?	?
is not	?	?

w3	spam	not
is spam	?	?
is not	?	?

w4	spam	not		
is spam	?	?		
is not	?	?		

w5	spam	not
is spam	?	?
is not	?	?

		w1	w2	w3	w4		w5
	email1	spam not		not	not	t	spam
	email2	spam	spam	spam	spa	m	spam
	email3	not	not spam		not	t	spam
	email4	spam	spam	spam	spa	m	not
	email5	spam	not	not	not	t	spam
P(em	m)	spam		not			
		email1		?		?	
		email2	2	?		?	
	email3		}	?		?	
	email4		?		?		
		email5	5	?		?	

w1	spam	not
is spam	?	?
is not	?	?
	1	
w2	spam	not
is spam	?	?
is not	?	?
w3	spam	not
is spam	?	?
is not	?	?
w4	spam	not
is spam	?	?
is not	?	?
w5	spam	not
is spam	?	?

?

is not

?

	w1	wa	w2	<u>м</u> /Л		w5		w1	spam	not
omail1	cnam	P(w	1 says	sp	an	n no	ot spam)	is spam	?	?
emair	span	hot		1100	,	span				•
email2	spam	spam	spam	spa	m	spam		is not	ڊ ا	?
email3	not	spam	not	not	;	spam		w2	spam	not
email4	spam	spam	spam	spa	m	not		is spam	?	?
email5	spam	not	not	not	;	spam		is not	?	?
					not			w3	spam	not
	L		spam		ΠΟΙ			is spam	?	?
	email1		?		?			is not	?	?
	email2	<u>)</u>	?		?			w4	spam	not
	<u> </u>							is spam	?	?
	email3	3	?		?			is not	?	?
	email4	L	?		?			w5	spam	not
	email5	5	?		?			is spam	?	?
		Lorenz	o De Stefan	i - CS	<u>CI 19</u>	51 A Data S	cience - Spring'22	is not	?	?

	w1	w2	w3	w4		w5		w1	spam	not
email1	spam	not	not	not		spam		is spam	1	0
email2	spam	spam	spam	Initializatio All workers perfect dec spam not			on:	is not	0	1
email3	not	spam	not				s make	w2	spam	not
email4	spam	spam	spam					is spam	1	0
email5	spam	not	not	not	;	spam		is not	0	1
		·			nat			w3	spam	not
			spam	bam				is spam	1	0
	email1	L	?		?			is not	0	1
	email2	<u>)</u>	?		?			w4	spam	not
	<u> </u>							is spam	1	0
	email3	}	?		?			is not	0	1
	email ²	L I	?					w5	spam	not
	email5	5	?		?			is spam	1	0
		Lorenz	o De Stefan	i - CS	<u>-</u> CI 19	51A Data S	cience - Spring'22	is not	0	1

		-			-		-						
		w1	\	w2	w3	w4	w5						
	email1	spam		not	not	not	spam						
email2		spam		spam	spam	spam	spa	m					
	email3	not spam		not		spam	not	not	spa	am			
	email4			spam	spam	spam	not						
	email5	spam	r	not	not	not	spam						
Compute probability of						spam		no	t				
labels b	y y							email	1	?	?		
averagır distribu [.]	ng ove tion of	r F		email	2	?		?	?				
votes				email	3	?		?					
				email	4	?		?					
3/23/22				email Lorenzo	5 5 De <u>Stefan</u>	? - CSCI <u>1</u> 9	51A E	? ata S	cience - Spr				

w1	spam	not
is spam	1	0
is not	0	1
w2	spam	not

w2	spam	not
is spam	1	0
is not	0	1

w3	spam	not
is spam	1	0
is not	0	1

w4	spam	not		
is spam	1	0		
is not	0	1		

w5	spam	not		
is spam	1	0		
is not	0	1		

		w1	w2	w3	w4	w5		w1	spam	not				
	email1	spam	not	not	not	spam		is spam	1	0				
	email2		spam	spam spam spam			is not	0	1					
	email3	not	spam	not	not	spam		w2	spam	not				
	email4	spam	spam	spam	spam	not		is spam	1	0				
	email5	spam	not	not	not	spam		is not	0	1				
Compute probability of						cnom not		w3	spam	not				
					spam		ΠΟΙ	is spam	1	0				
labels b	y y		email	email1 email2		email1		0+0+1	0+1+1+1+0 /5	is not	0	1		
averagin	ng ove	r c	email				?	w4	spam	not				
aistribu	tion of							is spam	1	0				
voles		ema		email3		email3 ?		email3 ?			?	is not	0	1
			email4		?		?	w5	spam	not				
							is spam	1	0					
3/23/22			email Lorenzo	ס De Stefan	f i - CSCI 19	051A Data	f Science - Spring'22	is not	0	1				

		w1	w2	2	w3	w	4	w5				w1	spam	not		
	email1	spam	am not not not		ot	spam				is spam	1	0				
	email2	spam	spa	am	spam	sp	bam	spam				is not	0	1		
	email3	not	spa	am	not	no	ot	spam				w2	spam	not		
	email4	spam	spa	am	spam	sp	bam	not				is spam	1	0		
	email5	spam	no	t	not	no	ot	spam				is not	0	1		
Comput	te						cnan	0	not			w3	spam	not		
probabi	probability of						span		not			is spam	1	0		
labels b	y ,			ema	email1		ail1		. 0.4		0.6		is not	is not	0	1
averagii	ng ove	r c		email2		il2 1		1			0	0		w4	spam	not
aistribu	tion of										i	is spam	1	0		
voles				ema	ail3		0.4		0.6		Ì	is not	0	1		
					email4		0.8		0.2	0.2		w5	spam	not		
					omailE				06			is spam	1	0		
3/23/22			L	orenzo	De Stefar	ni - (U.4 CSCI 19	51 A Data	Science -	Spring'22		is not	0	1		

		w1	v	v2	w3	w4	w5		
	email1	spam	n	ot	not	not	spar	n	
	email2	spam	S	pam	spam	spam	spar	n	
	email3	not	s	pam	not	not	spar	n	
	email4	spam	s	pam	spam	spam	not		
	email5	spam	n	ot	not	not	spar	n	
Compute						spam		nc	ot
labels b	inty of iy			email1		0.4	0.4 ?		6
averagi distribu	ng ove Ition of	r f	emai		12	?)
votes					13	?	?		
					4	?	?		
				emai	15	?	???		

Lorenzo De Stefani - CSCI 1951A Data Science - Spring'22

w1	spam	not		
s spam	1	0		
s not	0	1		
w2	snam	not		
	spann	ΠΟΙ		
is spam	1	0		

w3	spam	not
is spam	1	0
is not	0	1

w4	spam	not		
is spam	1	0		
is not	0	1		

w5	spam	not
is spam	1	0
is not	0	1

		w1	w2		w3	w	4	w5			w1	spam	not
	email1 spam		no	t	not	not spam		is spam					
	email2	spam	spa	am	spam	sp	bam	spam			is not		
	email3	not	spa	am	not	n	ot	spam			w2	spam	not
	email4	spam	spa	am	spam	sp	bam	not			is spam		
	email5	spam	no	t	not	n	ot	spam			is not		
Assume	the						cnan	0	n	ot	w3	spam	not
correct	labels	are					span	1			is spam		
those id	entifie	ed		ema	ail1		0.4		0	.6	is not		
by the n	najorit	у У		ema	ail2		1		0		w4	spam	not
and upo	ale in	e									is spam		
decision	I Matri	ces		ema	ail3		0.4		0	.6	is not		
				ema	ail4		0.8		0	.2	w5	spam	not
				ema	ail5		0.4		0	.6	is spam		
3/23/22			L	orenzo) De Stefar	ni - (CSCI 19	51A Data	Sc	tence - Spring'22	is not		

			w1	w2	w3	w4	w5		w1
		email1 spam		not	not	not	spam		is spar
		email2	spam	spam	spam	spam	spam		is not
		email3	not	spam	not	not	spam		w2
		email4	spam	spam	spam	spam	not		spam
		email5	spam	not	not	not	spam	Click	ker C
	Assume	the			spam	n	ot	(a)0.	.4, 0
	correct those id	labels Ientifie	are •d	1	0.4	0.	.6	(b)0	.6, 0
	by the n	najorit	:y	2	1 0			(c)0.8,	.8, 0
	and upc decisior	late th n matri	e ices		1	0		(d)1	.0, 0
				β	0.4	0.	.6		not
	emai		email	email4 email5		0.8 0.3		w5	
			email			0.	.6	spam	
	3/23/22			Lorenz	o De Stefa	ni - CSCI	1951A Data S	cience - Spring'22	not

spam_not ? not spam

Question!).6).4).2).0

w5	spam	not
spam		
not		

w1w2w3w4w5email1spamnotnotnotspamemail2spamspamspamspamspamspamemail3notspamspamspamspamspamemail4spamnotnotnotnotspamemail5spamnotnotnotspamspamemail5spamnotnotnotspamspamemail5spamnotnotnotspamemail5spamnotnotspamspamemail5spamnotnotnotemail5spamnotnotspamemail5spamspamspamnotemail5spamnotnotnotemail5spamspamspamnotemail5spamspamspamnotemail5spamspamspamnotemail5spamspamspamspamemail5spamspamspamspamemail5spamspamspamspamemail5spamspamspamspamemail5spamspamspamspamemail5spamspamspamspamemail5spamspamspamspamemail5spamspamspamspamemail5spamspamspamemail6s							
email1spamnotnotnotspamemail2spamspamspamspamspamspamemail3notspamnotnotnotspamemail4spamspamspamspamspamspamemail5spamnotnotnotnotspamemail5spamnotnotnotnotspamemail5spamnotnotnotspamspamemail5spamnotspamnotnotspamemail5spamnotnotnotspamemail20.40.40.40.4.email30.40.40.40.5.		w1	w2	w3	w4		w5
email2spamspamspamspamspamspamspamemail3notspamspamnotnotnotnotemail4spamnotnotnotnotspamnotemail5spamnotnotnotnotspamemail5spamnotnotnotspamemail1spamnotnotnotspamemail1oldspamnotoldemail2oldoldoldspamemail2oldoldoldemail3oldoldold	email1	spam	not	not	not	t	spam
email3notspamnot not spamspamemail4spamspam $spam$ $spam$ not not not email5spamnotnot not not $spam$ email15email12 0.44 0.6 ccc email2 0.44 0.6 ccc email3 0.44 0.6 ccc email3 0.44 0.6 $cccc$ email3 0.44 0.6 $ccccc$	email2	spam	spam	spam	spa	am	spam
email4spamspamspamspamnotemail5spamnotnotnotnotspamImage: Image: Im	email3	not	spam	not	not	t	spam
email5spamnotnotnotspam $[]_{-1}$	email4	spam	spam	spam	spa	am	not
spamnotemail10.40.6email210email30.40.6email40.80.2	email5	spam	not	not	not	t	spam
email10.40.6email210email30.40.6email40.80.2				spam		not	
email2 1 0 email3 0.4 0.6 email4 0.8 0.2		email1	L	0.4		0.6	
email3 0.4 0.6 email4 0.8 0.2		email2	2	1		0	
email4 0.8 0.2		email3	3	0.4		0.6	
		email4	ļ	0.8		0.2	
email5 Lorenzo De Stefani - CSCI 1951A Data S		email5	Lorenz	o De Stefar	1i - CS	0.6	51A Data S

w1	spam	not
is spam	1	0
is not		
w2	spam	not
spam	1	0
not	0	1
w3	spam	not
spam	1	0
not	0	1
м4	spam	not
spam	1	0
not	0	1
w5	spam	not

spam

	w1	w2	w3	w4		w5
email1	spam	not	not	not	t	spam
email2	spam	spam	spam	spam		spam
email3	not	spam	not	not		spam
email4	spam	spam	spam	spa	am	not
email5	spam	not	not	not	t	spam
			spam		not	
	email1		0.4		0.6	
	email2	2	1	0		
	email3		0.4		0.6	
	email4		0.8		0.2	
	email5	Lorenz	o De Stefan	<u>1 - CS</u>	0.6 CL 19	51A Data S

w1	spa m	not
is spam	1	0
is not	0.67	0.33
w2	spam	not
spam	1	0
not	0	1
w3	spam	not
spam	1	0
not	0	1
w4	spam	not
spam	1	0
not	0	1
w5	spam	not
spam	1	0

not

0

	w1	w2	w3	w4		w5		w1
email1	spam	not	not	not	t	spam		is spam
email2	spam	spam	spam	spa	am	spam		is not
email3	not	spam	not	ot not		spam		w2
email4	spam	spam	spam	spa	am	not		is spam
email5	spam	not	not	not	t	spam		is not
								w3
			spam		not			is spam
	email	L	0.4		0.6			is not
	email2	2	1	L 0				w4
	<u> </u>		0.4					is spam
	email	3			0.4		0.6	
	email4	1	0.8		0.2			w5
	email	_			0.6			is spam
	Cinalls	Lorenz	o De Stefar	ni - CS	CI 19	51A Data S	cience - Spring'22	is not

w1	spam	not
is spam	1	0
is not	0.67	0.33
w2	spam	not
is spam	1	0
is not	0.33	0.67
	-	
w3	spam	not
w3 is spam	spam 1	not 0
w3 is spam is not	spam 1 0	not 0 1
w3 is spam is not w4	spam 1 0 spam	not 0 1 not
w3 is spam is not w4 is spam	spam 1 0 spam 1	not 0 1 not 0

w5	spam	not
is spam	0.5	0.5
is not	1	0

			_								-
		w1	w	2	w3	w4	w5		w1	spam	not
	email1	spam	n	ot	not	not	spam		is spam	1	0
	email2	spam	sp	bam	spam	spam	spam		is not	0.67	0.33
	email3	not	sp	bam	not	not	spam		w2	spam	not
	email4	spam	S	oam	spam	spam	not		is spam	1	0
	email5	spam	n	ot	not	not	spam		is not	0.33	0.67
								_	w3	spam	not
Re-assign labels						spam		not	is spam	1	0
using we	eighteo	b		ema	il1	1.5		4.34	is not	0	1
• Each	worke	r is		ema	il2				w4	spam	not
weigh	nted			<u> </u>					is spam	1	0
accor	ding to	C		ema	il3				is not	0	1
their of bei	probal	bility rrect		ema	il4				w5	spam	not
	0			ema	il5				is spam	0.5	0.5
3/23/22				Lorenzo	n De Stefar	ni - CSCI 19)51A Dat	a Science - Spring'22	is not	1	0

	w1	w2		w3	W	4	w5			w1	spam	not
email1	spam	not	t	not	no	ot	spam			is spam	1	0
email2	spam	spa	m	spam	sp	am	spam			is not	0.67	0.33
email3	not	spa	am	not	no	ot	spam				spam	not
email4	spam	spa	m	spam	sp	am	not			is spam	1	0
email5	spam	not	t	not	no	ot	spam			is not	0.33	0.67
						snan	n	not		w3	spam	not
						Spann		not		is spam	1	0
Normalize!			email1			0.26		0.74		is not	0	1
			email2							w4	spam	not
										is spam	1	0
				email3						is not	0	1
		email4							w5	spam	not	
			email5							is spam	0.5	0.5
3/22 I				bronzo De Stofoni - Ci			51 A Data	Science Spring'	22	is not	1	0

	w1	w2		w3	W	4	w5			w1	spam	not
email	1 spam	not	t	not	no	ot	spam			is spam	1	0
email	2 spam	spa	am	spam	sp	am	spam			is not	0.67	0.33
email	3 not	spa	am	not	not		spam			w2	spam	not
email	4 spam	spa	am	spam	spam		not			is spam	1	0
email	5 spam	not	t	not	no	ot	spam			is not	0.33	0.67
						snar	n	not		w3	spam	not
						Spann				is spam	1	0
			em	nail1		0.26).74		is not	0	1
Compute for all			em	ail2		0.69		0.31		w4	spam	not
emails								0.71		is spam	1	0
			em	ail3	il3 il4					is not	0	1
			em	ail4				0.18	w5	w5	spam	not
			em	nail5		0.26		0.74	Ì	is spam	0.5	0.5
/23/22	L	oronz) De Stefani - (SCI 19	51 A Doto	Science Spring'22		is not	1	0	

	w1	w2	w3	w4	w5			w1	spam	not
email1	spam	not	not	not	spar	n		is spam	1	
email2	spam	spam	spam	spam	spar	n		is not		
email3	not	spam	not	not	spar	n		w2	spam	not
email4	spam	spam	spam	spam	not			is spam	1	0
email5	spam	not	not	not	spar	n		is not	0.33	0.67
								w3	spam	not
				spam		no	t	is spam	1	0
		emai	email1			0.7	74	is not	0	1
Iterate the		emai	email2		0.31		21	w4	spam	not
process until							is spam	1	0	
convergence	email3		0.29		0.7	71	is not	0	1	
		emai	email4			0.18		w5	spam	not
								is spam	0.5	0.5
3/23/22		emai Lorenz	15 o De Stefa	0.26 - CSCI 19)51A D	0.74 LA Data Science - Spring'22		is not	1	0

	w1	w2	2 w3 w4		4	4 w5			w1	spam	not	
email1	spam	no	t not r		n	ot	spam			is spam	1	0
email2	spam	spa	spam spa		spam		spam			is not	0.67	0.33
email3	not	spa	pam not		not		spam			w2	spam	not
email4	spam	spa	am spam s		sp	bam	not			is spam	1	0
email5	spam	no	t	not n		ot	spam			is not	0.33	0.67
							at	w3	spam	not		
In this examp	٦				span		n		is spam	1	0	
have convergence				email1			0.26		.74	is not	0	1
in 1 round				email2			0.69		31	w4	spam	not
						0.05				is spam	1	0
			email3 email4			0.29		0	0.71	is not	0	1
								0	0.18	w5	spam	not
									74	is spam	0.5	0.5
3/23/22	Drenzo De Stefani - CSCI 1				0.74 051A Data Science - Spring		. / 4 cience - Spring'22	is not	1	0		