CSCI 4022 Fall 2021
EM Wrapup; Itemsets

EM Wrapup Step 0: Initialize clusters and their means, variances, equal proportions.

Step 1: *Expectation*. For each data point x_i and for each each component m

1. $\tilde{p}_{mi} = \phi(x_i | \hat{\mu}_m, \hat{\Sigma}_m) \hat{w}_m$ and then consolidate into the probabilities:
2. $\hat{p}_{mi} = \dfrac{\tilde{p}_{mi}}{\sum_m \tilde{p}_{mi}}$

Step 2: *Maximization*. For each component m,

1. $\hat{w}_m = \dfrac{\hat{n_m}}{N} = \dfrac{\sum_{i=1}^N \hat{p}_{mi}}{N}$
2. $\hat{\mu}_m = \dfrac{1}{\hat{n_m}} \displaystyle\sum_{i=1}^N \hat{p}_{mi} \cdot x_i$
3. $\hat{\Sigma}_m = \dfrac{1}{\hat{n_m}} \displaystyle\sum_{i=1}^N \hat{p}_{mi} \cdot (x_i - \hat{\mu}_m)(x_i - \hat{\mu}_m)^T$

Step 3: Convergence check! Are things changing?

Announcements and To-Dos

Announcements:

1. HW 3 due **Wednesday**; missing code block:

Code addition

```
from sklearn.metrics.cluster import adjusted_rand_score
print(adjusted_rand_score([1, 0, 1], [0,1,0]))
# example that's actually the same assignments!
print(adjusted_rand_score([1, 0, 1], [0,0,1]))
# example: Rand score is negative if very different
```

2. HW 4 due next Monday.
3. Example code for covariance posted
4. Zach no OH tomorrow, 5p-6p tonight instead.
5. Min form comment: Go to DJ's office hours! Or e-mail him.

The EM Algorithm: More than just Gauss

You may note here that we could have fit *any* probability density into components and soft-clusters here, not just normals/Gaussians! We would have to replace the parameters μ, Σ of the normal to the alternative parameters Θ of the underlying distributions. We'd still fit M components and weights $\hat{w}_m$.

How would this change the EM algorithm?

The EM Algorithm: More than just Gauss

You may note here that we could have fit *any* probability density into components and soft-clusters here, not just normals/Gaussians! We would have to replace the parameters μ, Σ of the normal to the alternative parameters Θ of the underlying distributions. We'd still fit M components and weights $\hat{w}_m$.

How would this change the EM algorithm? **Solution:** Not that much!

Step 1: *Expectation*. For each data point x_i and for each each component m

1. $\tilde{p}_{mi} = \underbrace{f(x_i|\Theta)}_{any\, pdf} \hat{w}_m$ and then consolidate into the probabilities:
2. $\hat{p}_{mi} = \dfrac{\tilde{p}_{mi}}{\sum_m \tilde{p}_{mi}}$

Step 2: *Maximization*. For each component m,

1. $\hat{w}_m = \dfrac{\hat{n_m}}{N} = \dfrac{\sum_{i=1}^{N} \hat{p}_{mi}}{N}$
2. Estimate whatever is inside Θ... somehow (MLEs? OPTIM? Bootstrapping?)

Itemsets

That's it for EM and concludes our weeks on clustering! We'll implement EM in the notebook on Friday.

Now we move to Market Basket Analysis

Motivating Tale:

1. Fact: People who buy hot dogs are more likely to also buy ketchup.
2. Result: Stores can run a sale on hot dogs, but hike up the price of ketchup.

So what? That's nice, actionable info, but not particularly *insightful*. We want to use Data Science to find non-obvious insights!

Itemsets

That's it for EM and concludes our weeks on clustering! We'll implement EM in the notebook on Friday.

Now we move to Market Basket Analysis

Motivating Tale:

1. Fact: People who buy hot dogs are more likely to also buy ketchup.
2. Result: Stores can run a sale on hot dogs, but hike up the price of ketchup.

So what? That's nice, actionable info, but not particularly *insightful*. We want to use Data Science to find non-obvious insights!

Researchers have found that people who buy diapers are more likely to also buy beer.

Itemsets

...why?

Researchers have found that people who buy diapers are more likely to also buy beer.

Itemsets

Researchers have found that people who buy diapers are more likely to also buy beer.

Why? People who buy diapers are:

Itemsets

Researchers have found that people who buy diapers are more likely to also buy beer.

Why? People who buy diapers are:

1. Probably over 21

Itemsets

Researchers have found that people who buy diapers are more likely to also buy beer.

Why? People who buy diapers are:

1. Probably over 21
2. Probably have a baby at home, so prefer to bring beer home to drink instead of use bars

Itemsets

Researchers have found that people who buy diapers are more likely to also buy beer.

Why? People who buy diapers are:

1. Probably over 21
2. Probably have a baby at home, so prefer to bring beer home to drink instead of use bars
3. Probably stressed out of their minds, so want to drink

Itemsets

Researchers have found that people who buy diapers are more likely to also buy beer.

Why? People who buy diapers are:

1. Probably over 21
2. Probably have a baby at home, so prefer to bring beer home to drink instead of use bars
3. Probably stressed out of their minds, so want to drink

So the store can run a sale on diapers, and raise beer prices a little bit to get similar amounts of $ per customer... but maybe more customers!

Question: Would it work the other way, too? Run a sale on beer and raise diaper prices?

Itemsets

Question: Would it work the other way, too? Run a sale on beer and raise diaper prices?

Itemsets

Question: Would it work the other way, too? Run a sale on beer and raise diaper prices?

Probably not!

1. Nothing about buying beer makes someone more likely to need diapers, beyond maybe an age group overlap
2. Upshot: be careful! Relationships can be asymmetric, and causality and correlation can be hard to parse!

Items for Breakfast

Another **example:**

Here's an example from a Wegman's grocery store.

- What's the *association rule*?
- Is there a clear *direction* of association?

Dr. Tony Wong

Market Basket Analysis

Definition: the *Market basket model* describes a many-many relationship between two types of objects:

1. *Items* (or objects)
2. *Baskets*, which contain a set or count of *items* and an *itemset*.

Baskets don't always have to physically contain the items, but in the original use of MBA it was supermarkets and actual baskets/items.

Data scale: typically, the number of items in a basket is assumed to be small, and certainly much smaller than the number of baskets.

Market Basket Analysis

Definition: the *Market basket model* describes a many-many relationship between two types of objects:

1. *Items* (or objects)
2. *Baskets*, which contain a set or count of *items* and an *itemset*.

Baskets don't always have to physically contain the items, but in the original use of MBA it was supermarkets and actual baskets/items.

Goal: identify sets of items that are frequently bought together. Two descriptives: *frequent itemsets* and *association rules*

Data scale: typically, the number of items in a basket is assumed to be small, and certainly much smaller than the number of baskets.

Market Basket Analysis

Example: At a Walmart, each customer's purchase is a basket, and the products are items

- Customer 1: {bacon, chew toy, eggs}
- Customer 2: {avocados, bacon}
- Customer 3: {avocados, chew toy, tortilla chips, eggs}
- Customer 4: {avocados, bacon, tortilla chips, eggs}
- Customer 5: {chew toy, tortilla chips eggs}

Market Basket Analysis

Example: At a Walmart, each customer's purchase is a basket, and the products are items

- Customer 1: {bacon, chew toy, eggs}
- Customer 2: {avocados, bacon}
- Customer 3: {avocados, chew toy, tortilla chips, eggs}
- Customer 4: {avocados, bacon, tortilla chips, eggs}
- Customer 5: {chew toy, tortilla chips eggs}

Examples: "frequently bought together" or "people who bought X also bought Y."

Market Basket Analysis: Applications

baskets= sentences/ideas; **items**= documents containing those sentences.

- Items that appear frequently together might be another form of plagiarism
- Item's aren't "in" baskets, but this is consistent with our notions of document similarity

baskets= patients; **items**= drugs and their side-effect combinations

- Can detech *combinations* of drugs that lead to undesirable side-effects.
- Need to also represent the *absence* of an item, e.g. Patient 1 had *these* drugs and also *no* side effects

Frequent Itemsets

Goal: Find the sets of items that occur *frequently* together.

Definition: The *support* for itemset I is the number of baskets that contain all items in I. Often, support is expressed as a fraction of the total number of baskets.

Definition: Given a *support threshold* s, the sets of items that appeat in at least s baskets are called *frequent itemsets*.

s is proportion:

10,000

~~s=300~~

s=3%

.03

Frequent Itemsets

Example: Back to Walmart.

Definition: The *support* for itemset I is the number of baskets that contain all items in I. Often, support is expressed as a fraction of the total number of baskets.

Definition: Given a *support threshold* s, the sets of items that appeat in at least s baskets are called *frequent itemsets*.

Example Data:
{bacon, chew toy, eggs}
{avocados, bacon}
{avocados, chew toy, tortilla chips, eggs}
{avocados, bacon, tortilla chips, eggs}
{chew toy, tortilla chips eggs}

Ex: Support of {bacon}? 3 = 3/5

Ex: Support of {bacon, eggs}? 2 = 2/5

Frequent Itemsets

Example: Back to Walmart.

Definition: The *support* for itemset I is the number of baskets that contain all items in I. Often, support is expressed as a fraction of the total number of baskets.

Definition: Given a *support threshold* s, the sets of items that appeat in at least s baskets are called *frequent itemsets*.

Frequent Itemsets

m c ~~p~~ b j

5 4 ~~2~~ 6 4

Example: Suppose our **items** are {milk, coke, pepsi, beer, juice} and we want a support threshold of $s = 3$ baskets.

$B_1 =$	{m,c,b}	$B_5 =$	{m,p,j}
$B_2 =$	{m,b}	$B_6 =$	{c,j}
$B_3 =$	{m,p,b}	$B_7 =$	{m,c,b,j}
$B_4 =$	{c,b,j}	$B_8 =$	{b,c}

What are all of the frequent itemsets?

size 1: {m, c, b, j}

size 2: {mc, mb, cb, cj}

mc 2

~~mp~~

mb 4

~~mj 2~~

~~cp~~

cb 3

cj 3

~~pb~~

~~pj~~

~~bj 1~~

mbc: 2

size 3:

$\binom{5}{3} = 10$

Frequent Itemsets

$$\binom{10{,}000}{5} \approx \frac{10^4\cdot 10^4\cdot 10^4\cdot 10^4\cdot 10^4}{5\cdot 4\cdot 3\cdot 2\cdot 1}$$

Example: Suppose our **items** are {milk, coke, pepsi, beer, juice} and we want a support threshold of $s = 3$ baskets.

$\approx 10^{18} =$

$B_1=$	{m,c,b}	$B_5=$	{m,p,j}
$B_2=$	{m,b}	$B_6=$	{c,j}
$B_3=$	{m,p,b}	$B_7=$	{m,c,b,j}
$B_4=$	{c,b,j}	$B_8=$	{b,c}

What are all of the frequent itemsets?

Solution:

1. Size 1: {{m},{c},{b},{j}}. Not {p}.
2. Size 2: 10 possibilities! (Since that's 5 choose 2). In this case, {m,b},{c,j} and {c,b} are both frequent.
3. Size 3: None... could we have known this already?

Association Rules

Definition: An *association rule* is an *if-then* rule about the contents of baskets. We denote

$$\{i_1, i_2, \ldots i_k\} \to i_j$$

to represent "if a basket contains all of $i_1, i_2, \ldots i_k$, it is *likely* to contain j as well."

In practice, there will of course be lots of rules, so we want to find the most significant ones. This requires a notion of *confidence.*

It also turns out that some rules, like $X \to milk$ will inevitably have high confidence for many itemsets X simply because milk is popular. Having many high-confidence associations isn't necessarily actionable, so we also want a measure of *interest*.

Association Rules

Definition: The *confidence* of the association rule $I \to J$ is the ratio of the support for $I \cup \{j\}$ to the support for I.

$$conf(I \to J) = \frac{support(I \cup \{j\})}{support(I)}$$

Association Rules

Definition: The *confidence* of the association rule $I \to J$ is the ratio of the support for $I \cup \{j\}$ to the support for I.

$$conf(I \to J) = \frac{support(I \cup \{j\})}{support(I)}$$

Definition: The *interest* of the association rule $I \to J$ is the difference between its confidence and the fraction of baskets that contain j

$$interest(I \to J) = conf(I \to J) - P(j)$$

Association Rules

Definition: The *confidence* of the association rule $I \rightarrow J$ is the ratio of the support for $I \cup \{j\}$ to the support for I.

$$conf(I \rightarrow J) = \frac{support(I \cup \{j\})}{support(I)}$$

This is a lot like *conditional probability*. Recall that $P(A|B) = \frac{P(both)}{P(B)} = \frac{P(A \cap B)}{P(B)}$. Then $support(I \cup \{j\})$ is the count of all the baskets that have *both* all of I and j as the numerator: so it's really like an intersection of those two sets! $conf(I \rightarrow J)$ behaves like probability of J **given** I.

Definition: The *interest* of the association rule $I \rightarrow J$ is the difference between its confidence and the fraction of baskets that contain j

$$interest(I \rightarrow J) = conf(I \rightarrow J) - P(j)$$

Association Rules

Definition: The *confidence* of the association rule $I \rightarrow J$ is the ratio of the support for $I \cup \{j\}$ to the support for I.

$$conf(I \rightarrow J) = \frac{support(I \cup \{j\})}{support(I)}$$

This is a lot like *conditional probability*. Recall that $P(A|B) = \frac{P(both)}{P(B)} = \frac{P(A \cap B)}{P(B)}$. Then $support(I \cup \{j\})$ is the count of all the baskets that have *both* all of I and j as the numerator: so it's really like an intersection of those two sets! $conf(I \rightarrow J)$ behaves like probability of J **given** I.

Definition: The *interest* of the association rule $I \rightarrow J$ is the difference between its confidence and the fraction of baskets that contain j

$$interest(I \rightarrow J) = conf(I \rightarrow J) - P(j)$$

In the probability sense, this is a bit like $P(J|I) - P(J)$

Frequent Itemsets

Example: We can't escape Walmart. Consider the association rule {m,b}$\rightarrow$ c. What are the confidence and interest?

$B_1=$	{m,c,b}	$B_5=$	{m,p,j}
$B_2=$	{m,b}	$B_6=$	{c,j}
$B_3=$	{m,p,b}	$B_7=$	{m,c,b,j}
$B_4=$	{c,b,j}	$B_8=$	{b,c}


FEBRUARY 4, 2013

Police Arrest Florida Man For Drunken Joyride On Motorized Scooter At Walmart

Frequent Itemsets

Example: We can't escape Walmart. Consider the association rule {m,b}$\rightarrow$ c. What are the confidence and interest?

$B_1=$	{m,c,b}	$B_5=$	{m,p,j}
$B_2=$	{m,b}	$B_6=$	{c,j}
$B_3=$	{m,p,b}	$B_7=$	{m,c,b,j}
$B_4=$	{c,b,j}	$B_8=$	{b,c}


FEBRUARY 4, 2013

Police Arrest Florida Man For Drunken Joyride On Motorized Scooter At Walmart

Solution:

1. Support {m,b}=4; Support {m,b,c}=2. $conf(\{m,b\} \rightarrow c) = \frac{2}{4}$
2. Interest: $interest(\{m,b\} \rightarrow c) = \frac{2}{4} - \frac{5}{8} = -0.125$

Frequent Itemsets

Example: We can't escape Walmart. Consider the association rule {m,b}$\rightarrow$ c. What are the confidence and interest?

$B_1=$	{m,c,b}	$B_5=$	{m,p,j}
$B_2=$	{m,b}	$B_6=$	{c,j}
$B_3=$	{m,p,b}	$B_7=$	{m,c,b,j}
$B_4=$	{c,b,j}	$B_8=$	{b,c}


FEBRUARY 4, 2013

Police Arrest Florida Man For Drunken Joyride On Motorized Scooter At Walmart

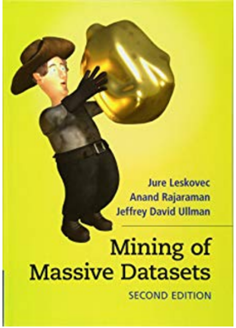
Solution:

1. Support {m,b}=4; Support {m,b,c}=2. $conf(\{m,b\} \rightarrow c) = \frac{2}{4}$
2. Interest: $interest(\{m,b\} \rightarrow c) = \frac{2}{4} - \frac{5}{8} = -0.125$

Coke is pretty popular, since it's in 5/8 baskets. So the rule that half of the milk+beer baskets also have coke is very uninteresting!

Acknowledgments

Some material is adapted/adopted from Mining of Massive Data Sets, by Jure Leskovec, Anand Rajaraman, Jeff Ullman (Stanford University) `http://www.mmds.org`



Special thanks to Tony Wong for sharing his original adaptation and adoption of slide material.