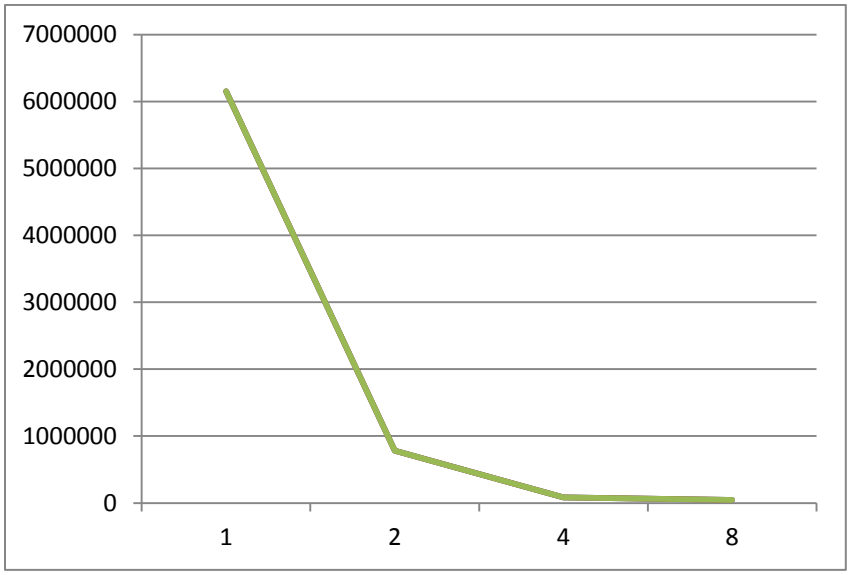
CS57300: Homework 6

Ravi Kiran Rao Bukka

1. a) Continuous Features:

K	WSS
1	6150320.35148
2	784158.105858
4	84329.3514186
8	46500.680818
15	45867.6959084
25	15969.8896897

Below is the graph with K on x-axis and WSS on y-axis:

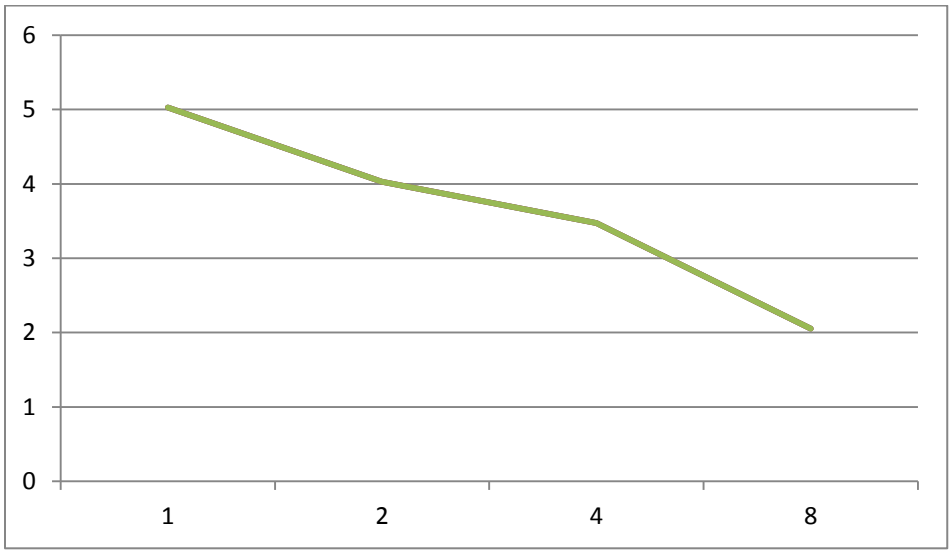


b) Binary Features:

K	WSS
1	5.02777777778
2	4.02777777778
4	3.47222222222
8	2.05555555556
15	1.91666666667

25	0.86555333333

Below is the graph with K on x-axis and WSS on y-axis:



2. a) Continuous Features:

We choose the value of 'k' where the curve is almost flattened. Choosing a value of K= 4 is an appropriate choice. It is purely subjective and there is no mathematical formula which shows the value of 'k' to be chosen. It is based on intuition.

b) Binary Features:

Here as we see the decrease in WSS after K= 8 is very minimal so we choose a value of up to K=8.

3. a) Continuous Features:

Below are the values of average and variance with each 'K'

Average: 6150320.35148 K: 1

Variance: 0.0 K: 1

Average: 784158.105858 K: 2

Variance: 0.0 K: 2

Average: 84329.3514186 K: 4

Variance: 5.50571415715e-21 K: 4

Average: 46500.680818 K: 8

Variance: 1.00585162486e-21 K: 8

Average: 45907.8258836 K: 15

Variance: 296.046130775 K: 15

Average: 38448.9231521 K: 25

Variance: 645569259.725 K: 25

As we can see the variance has different values for different clusters and 0 in only two cases, we can infer that the clustering depends a lot on the initial conditions like initial centroids chosen.

b) Binary Features:

Average: 5.02777777778 K: 1

Variance: 0.0 K: 1

Average: 3.78333333333 K: 2

Variance: 1.93580246914 K: 2

Average: 2.78333333333 K: 4

Variance: 1.92962962963 K: 4

Average: 2.46111111111 K: 8

Variance: 3.24166666667 K: 8

Average: 1.04444444444 K: 15

Variance: 2.20771604938 K: 15

Average: 0.605555555556 K: 25

Variance: 1.85092592593 K: 25

As we can see the variance has different values for different clusters and 0 in only one case, we can infer that the clustering depends a lot on the initial conditions like initial centroids chosen.

4. a) Continuous Features:

Squared Diff: 84329.3514186 nclusters: 4

Interesting Clusters:

The cluster centroids and closest points are as follows:

Closest point: [34.0630451, -118.4468859, 4.5, 2350.0]

Centroid: [34.0630451, -118.4468859, 4.5, 2350.0]

A one point cluster, which definitely shows that its reviews are very high compare to other. It shows that clustering is heavily relying on review count for this case.

Attributes are:

{u'city': u'Los Angeles', u'review_count': **2350**, u'name': u'Diddy Riese Cookies', u'neighborhoods': [u'Westwood'], u'url':, u'latitude': 34.0630451, u'state': u'CA', u'longitude': -118.4468859, u'stars': 4.5, u'schools': [u'University of California - Los Angeles'], u'open': True, u'categories': [u'Food', u'Desserts', u'Bakeries', u'Ice Cream & Frozen Yogurt'], u'photo_url': u'http://s3-media4.px.yelpcdn.com/bphoto/XDoGHAgyPxI1GWOHd_m5ew/ms.jpg'}

Closest point: [34.0646018, -118.4480733, 3.5, 243.0] **Centroid:** [34.06369304181428, -118.44620370635712, 3.5357142857142856, 230.85714285714286]

This cluster mostly has restaurants with an closer number of reviews to the centroid.

Attributes are:

{u'city': u'Westwood', u'review_count': **243**, u'name': u'Calbi Korean BBQ', u'neighborhoods': [u'Westwood'], u'url': u'http://www.yelp.com/biz/calbi-korean-bbq-westwood',u'type': u'business', u'business_id': u'lIPv3-ocdPvkiiE6JaQi9w', u'full_address': u'Gayley Ave\nWestwood\nWestwood, CA 90024', u'latitude': 34.0646018, u'state': u'CA', u'longitude': -118.4480733, u'stars': 3.5, u'schools': [u'University of California - Los Angeles'], u'open': True, u'categories': [u'Asian Fusion', u'Food Stands', u'Restaurants'],u'photo_url': u'http://s3-media1.px.yelpcdn.com/bphoto/FIEd_xG8LlbPpEwp8gosvQ/ms.jpg'}

Closest point: [34.0716414, -118.4523812, 3.5, 16.0] **Centroid:** [34.066816893143695, -118.44525704287379, 3.512135922330097, 9.29611650485437]

This cluster mostly has cafe's with an closer number of reviews to the centroid.

Attributes are:

{u'city': u'Los Angeles', u'review_count': **16**, u'name': u'Bruin Cafe', u'neighborhoods': [u'Westwood'], u'url': u'http://www.yelp.com/biz/bruin-cafe-los-angeles', u'type': u'business', u'business_id': u'rKs40LPiZH9mDGixhfnvQA', u'full_address': u'360 De Neve Drive\nWestwood\nLos Angeles, CA 90024', u'latitude': 34.0716414, u'state': u'CA', u'longitude': -118.4523812, u'stars': 3.5, u'schools': [u'University of California - Los Angeles'], u'open': True, u'categories': [u'Cafes', u'Sandwiches', u'Restaurants']}

b) Binary Features:

Squared Diff: 2.05555555556 nclusters: 8. The general trend we observe is sometimes there are one point clusters and in some cases the nearest point to the cluster is not even a close match. But the good point is in most cases the categories of centroid and the closest point match which indicates some advantage of clustering.

Closest point: [0.0, 0.0, 0.0, 0.0, 0.0, 0.0]

Centroid: [0, 0, 0, 0, 0, 0]

{u'city': u'Los Angeles', u'review_count': 231, u'name': u'W Los Angeles', u'neighborhoods': [u'Westwood'], u'url': u'http://www.yelp.com/biz/w-los-angeles-los-angeles',u'type': u'business', u'business_id': u'9ta6U2F5ma2UauHBS-i1dg', u'full_address': u'930 Hilgard Ave\nWestwood\nLos Angeles, CA 90024', u'latitude': 34.063201, u'state': u'CA', u'longitude': -118.441029, u'stars': 4.0, u'schools': [u'University of California - Los Angeles'], u'open': True, u'categories': [u'Hotels & Travel', u'Event Planning & Services', u'Hotels'], u'photo_url': u'http://s3-media4.px.yelpcdn.com/bphoto/QLpMmWpIVSjIbhuXOjykEw/ms.jpg'}

Closest point: [0.0, 0.0, 0.0, 0.0, 0.0, 0.0]

Centroid: [1, 1, 1, 1, 1, 1]

{u'city': u'Los Angeles', u'review_count': 231, u'name': u'W Los Angeles', u'neighborhoods': [u'Westwood'], u'url': u'http://www.yelp.com/biz/w-los-angeles-los-angeles',u'type': u'business', u'business_id': u'9ta6U2F5ma2UauHBS-i1dg', u'full_address': u'930 Hilgard Ave\nWestwood\nLos Angeles, CA 90024', u'latitude': 34.063201, u'state': u'CA', u'longitude': -118.441029, u'stars': 4.0, u'schools': [u'University of California - Los Angeles'], u'open': True,

u'categories': [u'Hotels & Travel', u'Event Planning & Services', u'Hotels'], u'photo_url': u'http://s3-media4.px.yelpcdn.com/bphoto/QLpMmWpIVSjIbhuXOjykEw/ms.jpg'}

Closest point: [1.0, 0.0, 0.0, 0.0, 0.0, 0.0]

Centroid: [1, 0, 0, 0, 0, 0]

{u'city': u'Los Angeles', u'review_count': 4, u'name': u'Bruin Buzz', u'neighborhoods': [u'Westwood'], u'url': u'http://www.yelp.com/biz/bruin-buzz-los-angeles', u'type': u'business', u'business_id': u'x6M5SjgjDjxWWBkxYvtLCA', u'full_address': u'308 Westwood Plaza\nAckerman 1st Fl\nWestwood\nLos Angeles, CA 90095', u'latitude': 34.0730451, u'state': u'CA', u'longitude': -118.4448434, u'stars': 3.5, u'schools': [u'University of California - Los Angeles'], u'open': True, u'categories': [u'Food', u'Coffee &Tea'], u'photo_url': u'http://s3-media4.px.yelpcdn.com/bphoto/1vDElFBCXEjL6XiNUa519g/ms.jpg'}

Closest point: [0.0, 1.0, 0.0, 1.0, 0.0, 0.0]

Centroid: [0, 1, 0, 1, 0, 1]

{u'city': u'Los Angeles', u'review_count': 9, u'name': u'Westwood Arcade', u'neighborhoods': [u'Westwood'], u'url': u'http://www.yelp.com/biz/westwood-arcade-los-angeles', u'type': u'business', u'business_id': u'wu8DoG-OyXMHZIlJcOmYIA', u'full_address': u'10965 Weyburn Ave\nWestwood\nLos Angeles, CA 90024', u'latitude': 34.062373, u'state': u'CA', u'longitude': -118.447842, u'stars': 4.5, u'schools': [u'University of California - Los Angeles'], u'open': False, u'categories': [u'Hobby Shops', u'Arts & Entertainment', u'Shopping', u'Arcades'], u'photo_url': u'http://media1.px.yelpcdn.com/static/20101216322333644l/img/gfx/blank_biz_medium.gif'}

Closest point: [0.0, 1.0, 0.0, 1.0, 0.0, 0.0]

Centroid: [1, 1, 1, 1, 1, 1]

{u'city': u'Los Angeles', u'review_count': 9, u'name': u'Westwood Arcade', u'neighborhoods': [u'Westwood'], u'url': u'http://www.yelp.com/biz/westwood-arcade-los-angeles', u'type': u'business', u'business_id': u'wu8DoG-OyXMHZIlJcOmYIA', u'full_address': u'10965 Weyburn Ave\nWestwood\nLos Angeles, CA 90024', u'latitude': 34.062373, u'state': u'CA', u'longitude': -118.447842, u'stars': 4.5, u'schools': [u'University of California - Los Angeles'], u'open': False, u'categories': [u'Hobby Shops', u'Arts & Entertainment', u'Shopping', u'Arcades'], u'photo_url': u'http://media1.px.yelpcdn.com/static/20101216322333644l/img/gfx/blank_biz_medium.gif'}

Closest point: [0.0, 1.0, 0.0, 1.0, 0.0, 0.0]

Centroid: [1, 1, 1, 1, 1, 1]

{u'city': u'Los Angeles', u'review_count': 9, u'name': u'Westwood Arcade', u'neighborhoods': [u'Westwood'], u'url': u'http://www.yelp.com/biz/westwood-arcade-los-angeles', u'type': u'business', u'business_id': u'wu8DoG-OyXMHZIlJcOmYIA', u'full_address': u'10965 Weyburn Ave\nWestwood\nLos Angeles, CA 90024', u'latitude': 34.062373, u'state': u'CA', u'longitude': -118.447842, u'stars': 4.5, u'schools': [u'University of California - Los Angeles'], u'open': False, u'categories': [u'Hobby Shops', u'Arts & Entertainment', u'Shopping', u'Arcades'], u'photo_url': u'http://media1.px.yelpcdn.com/static/201012163223336441/img/gfx/blank_biz_medium.gif'}

Closest point: [0.0, 1.0, 0.0, 1.0, 0.0, 0.0]

Centroid: [1, 1, 1, 1, 1, 1]

{u'city': u'Los Angeles', u'review_count': 9, u'name': u'Westwood Arcade', u'neighborhoods': [u'Westwood'], u'url': u'http://www.yelp.com/biz/westwood-arcade-los-angeles', u'type': u'business', u'business_id': u'wu8DoG-OyXMHZIlJcOmYIA', u'full_address': u'10965 Weyburn Ave\nWestwood\nLos Angeles, CA 90024', u'latitude': 34.062373, u'state': u'CA', u'longitude': -118.447842, u'stars': 4.5, u'schools': [u'University of California - Los Angeles'], u'open': False, u'categories': [u'Hobby Shops', u'Arts & Entertainment', u'Shopping', u'Arcades'], u'photo_url': u'http://media1.px.yelpcdn.com/static/201012163223336441/img/gfx/blank_biz_medium.gif'}

Closest point: [0.0, 1.0, 0.0, 1.0, 0.0, 0.0]

Centroid: [1, 1, 1, 1, 1, 1]

{u'city': u'Los Angeles', u'review_count': 9, u'name': u'Westwood Arcade', u'neighborhoods': [u'Westwood'], u'url': u'http://www.yelp.com/biz/westwood-arcade-los-angeles', u'type': u'business', u'business_id': u'wu8DoG-OyXMHZIlJcOmYIA', u'full_address': u'10965 Weyburn Ave\nWestwood\nLos Angeles, CA 90024', u'latitude': 34.062373, u'state': u'CA', u'longitude': -118.447842, u'stars': 4.5, u'schools': [u'University of California - Los Angeles'], u'open': False, u'categories': [u'Hobby Shops', u'Arts & Entertainment', u'Shopping', u'Arcades'], u'photo_url': u'http://media1.px.yelpcdn.com/static/201012163223336441/img/gfx/blank_biz_medium.gif'}

5. a) Continuous Features:

The normalized mutual information gain.value for the continuous features varies between 0.2 and 0.3.

b) Binary Features:

The normalized mutual information gain value for the binary features varies between 0.8 and 0.9.

This shows that the binary features are much better for clustering.