CS57300: Homework 4

Long Zhen, lzhen@purdue.edu

Programming assignment

You should implement your solution using Python. As described above, you are welcome to use your previous HW3 code and/or any methods available in the scikit-learn python library. As before, you should submit your typed HW report as a pdf, along with your source code files.

1. *Inspect the clustering results, comparing standard kmeans to spherical kmeans.* (15 pts)

   (a) Evaluate the standard kmeans algorithm in python for cluster sizes $k = [10, 20, 50, 100, 200]$ Plot the cluster score vs. k for both $\mathbf{W}_P$ and $\mathbf{W}_{NP}$. Identify the "best" k.

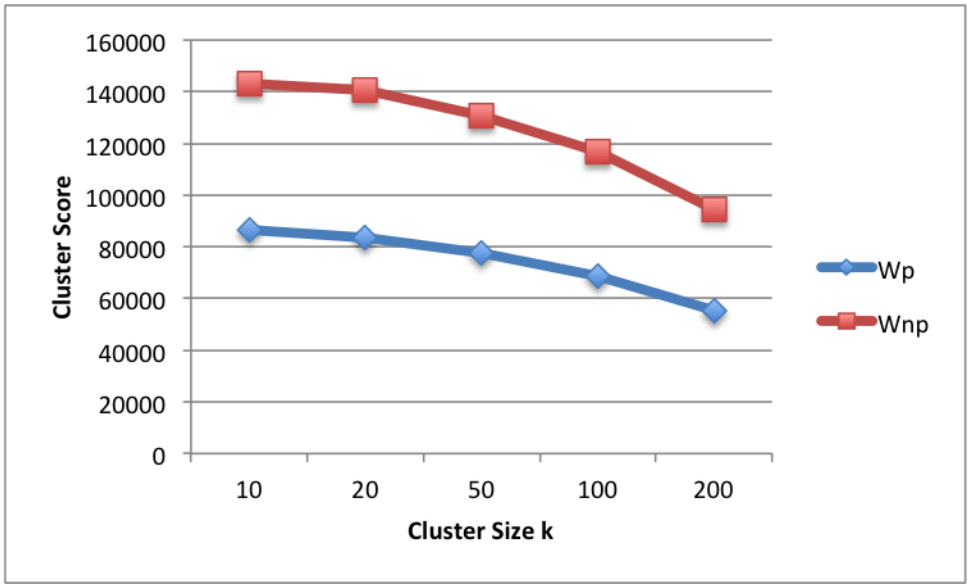

Figure 1: Q1(a) Plot of the cluster score vs. k for both $\mathbf{W}_P$ and $\mathbf{W}_{NP}$

k\Data	W_P	W_{NP}
10	86520.47871	143289.7154
20	83592.56231	140801.3924
50	77645.75359	130700.7133
100	68662.51319	116622.1399
200	55399.53907	94319.66212

Table 1: Q1(a) Table of the cluster score vs. k for both $\mathbf{W}_P$ and $\mathbf{W}_{NP}$

Answer: From the table and plot, we can easily identify that the largest k, 200, is the best one.

(b) For the best choice of k, inspect the words in the clusters and report any noticeable topics that you find.

Answer: For both W_P and W_{NP} clusters, there is always one or two clusters with large numbers of words. These big clusters does not contain too much information because the meanings of most words are not related or even opposite, such as fantastic and terrible may be in one cluster. But some clusters with two or three words are more meaningful. For example,

- american, native → nationality word
- office, dr, doctor → doctor related word
- moved, move, unit, complex, apartment, rent, maintenance, lease → housing related word

(c) Repeat (a-b) using your spherical kmeans algorithm.

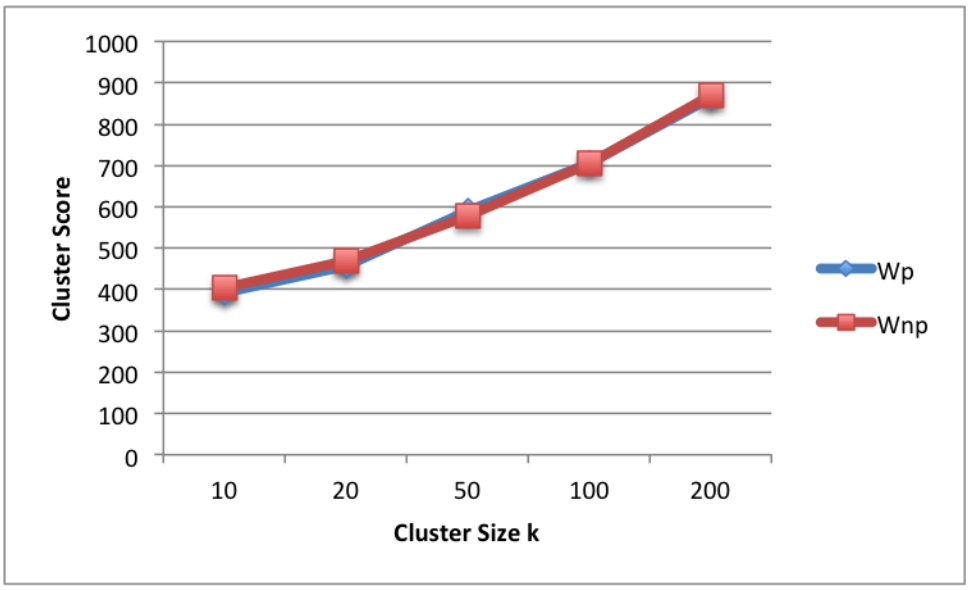

Figure 2: Q1(c) Plot of the cluster score vs. k for both $\mathbf{W}_P$ and $\mathbf{W}_{NP}$

k\Data	W_P	W_{NP}
10	392.2220511	404.4031075
20	458.1323861	469.932151
50	591.3608539	579.2008092
100	705.6703374	704.2789863
200	863.3954781	868.496281

Table 2: Q1(c) Table of the cluster score vs. k for both $\mathbf{W}_P$ and $\mathbf{W}_{NP}$

Answer: Since in spherical $k-means$ cosine similarity is used instead of distance function, the larger score is better. From the table and plot, we can easily identify that the largest k, 200, is still the best one.

While using spherical $k-means$, words are more equally separated among all clusters. Here is some meaningful cluster examples,

- awesome, highly, fantastic, yummy → nice word
- salon, normal, nail, nails, color, tech, cat, rushed, polish → salon related word
- worst, far, horrible, absolute, fin, indoor → negative word

(d) Discuss any noticeable differences in the results from the two algorithms.

Answer: Spherical $k-means$ can classify words more equally into all clusters, while standard $k-means$ always classify most of words into one or two cluster. This is the most obvious difference between two algorithms. It seems that spherical $k-means$ can more equally separate words.

2. *Assess whether the clustering approach improves performance.* (10 pts)
Let approach A be the NBC using binary features from the 100 topics that you identified with standard kmeans (50 topics from $\mathbf{W}_P$ and 50 from $\mathbf{W}_{NP}$).
Let approach B be the NBC using 100 binary features derived from spherical kmeans topics.

(a) Plot the learning curves for A and B including error bars that indicate ± 1 *standard error*, using evaluation based on incremental CV as described above.

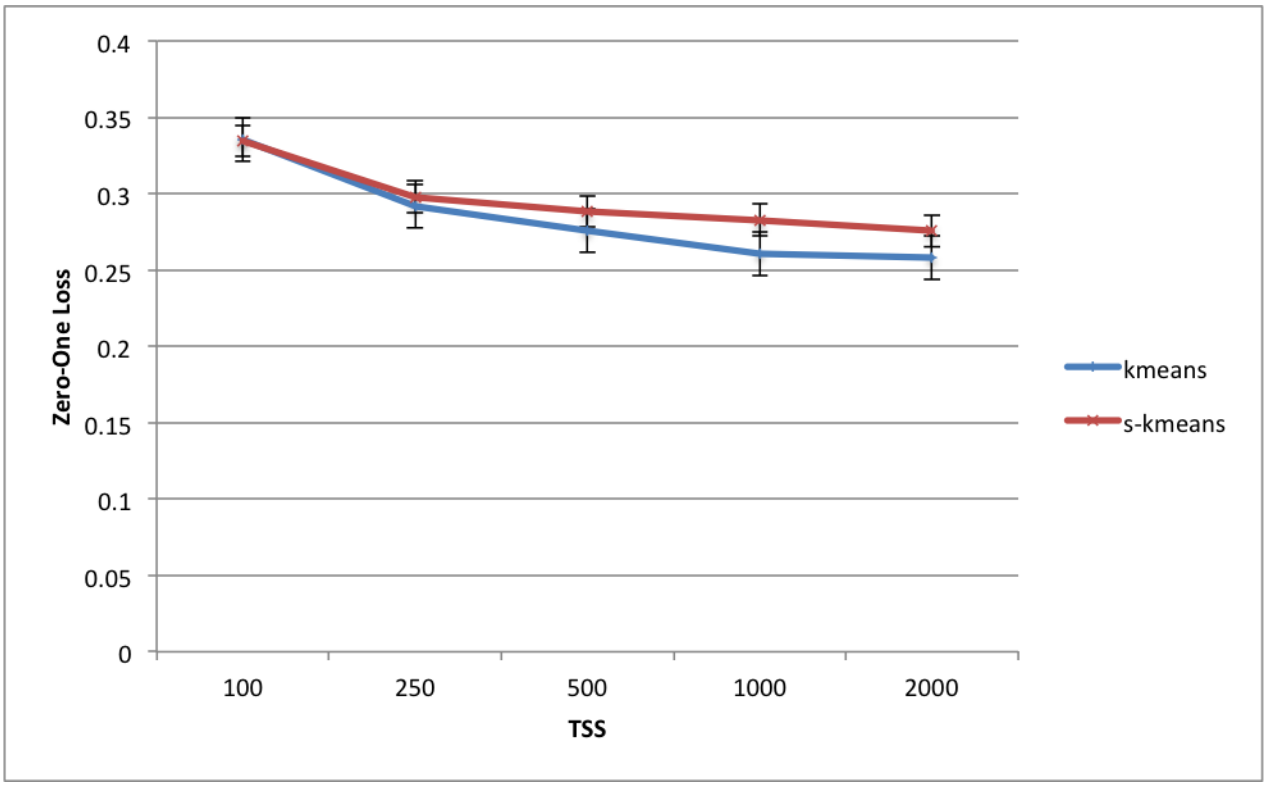

Figure 3: Q2(a) Plot of learning curves for A and B including error bars

(b) Formulate a hypothesis about the performance difference between A and B.

Answer: $k-means$ will perform better than spherical $k-means$ as tss increases.

(c) Discuss whether the observed results support the hypothesis (i.e., are the observed differences significant).

Answer: Yes. My learning curve supports my hypothesis. When tss is relatively small, such as 100 or 250, the points are almost overlapped. As tss increasing, $k-means$ smaller and smaller loss rate than spherical $k-means$. Especially, when tss = 1000, there is totally no overlap considering the error bars. At this point, we may confidently say that $k-means$ performs better than spherical $k-means$.

3. *Assess whether using the topic features improves performance.* (10 pts)
Let approach A be the original NBC with 2000 binary word features (i.e., HW3).
Let approach B be the NBC model using 100 binary features from the kmeans topics (using the best model from part 2).

   (a) Plot the learning curves for A and B including error bars that indicate ± 1 *standard error*, using evaluation based on incremental CV as described above.

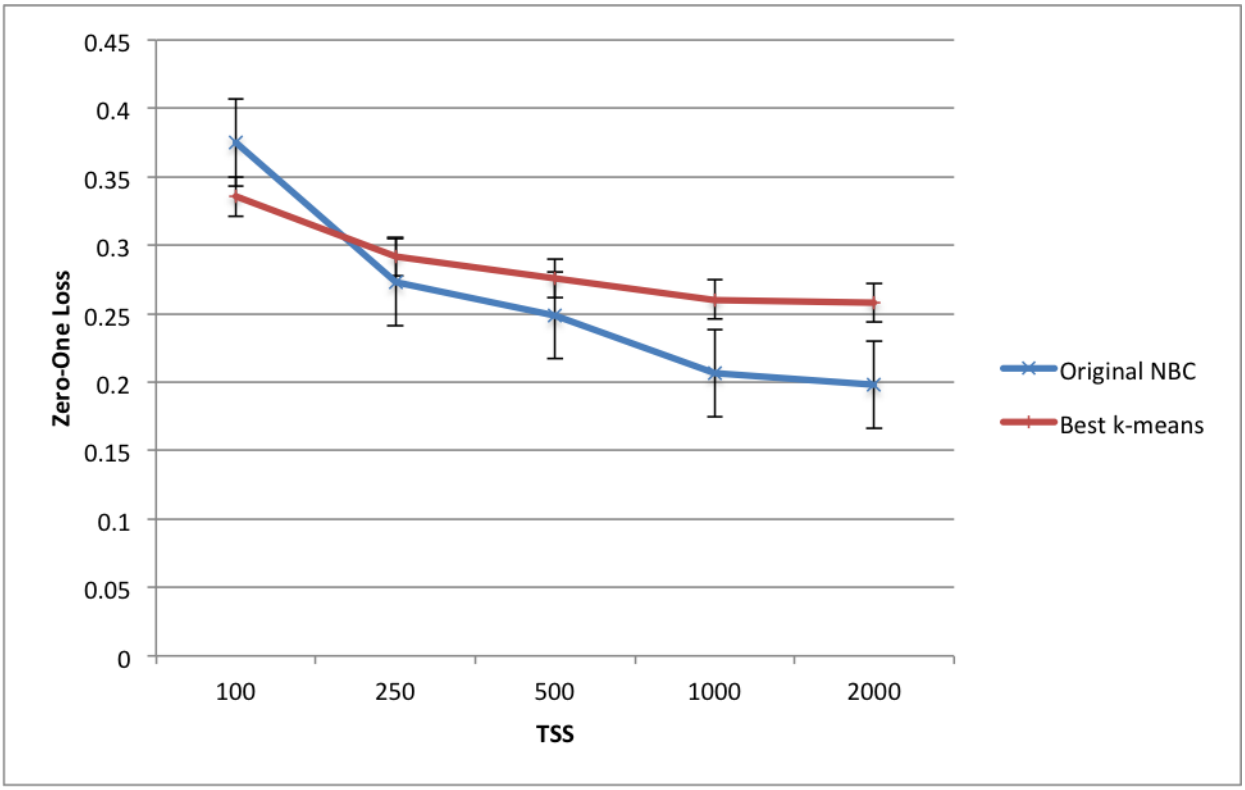

Figure 4: Q3(a) Plot of learning curves for A and B including error bars

   (b) Formulate a hypothesis about the performance difference between A and B.

Answer: $k-means$ should perform better than original NBC. When tss increasing, $k-means$ should perform better and better.

   (c) Discuss whether the observed results support the hypothesis.

Answer: No. Surprisingly, original NBC performs much better than $k-means$ clustered one, especially when tss is large. From the plot, we can see that only when the size of train set is relatively small (100), $k-means$ clustered performs better than original

NBC. As the tss increasing, the original NBC perform better and better. More interestingly, when tss goes above 1000, the original NBC totally beats clustered one considering the error bars. Even top of original NBC error bar is lower than the bottom of clustered error bar. Therefore, I think cluster may only be useful when we do not have enough training data.

4. *Assess whether the number of features improve performance.* (10 pts)
Let approach A be the original NBC with binary word features from HW3, but randomly sample only 100 of the 2000 word features to use in the model.
Let approach B be the NBC model using 100 binary features from the kmeans topics (again using the best model from part 2).
Let approach C be an NBC model that uses the 100 randomly selected word features from A and the 100 binary topic features from B.

   (a) Plot the learning curves for A, B, and C, including error bars that indicate ± 1 standard error, from the evaluation based on incremental CV as described above.

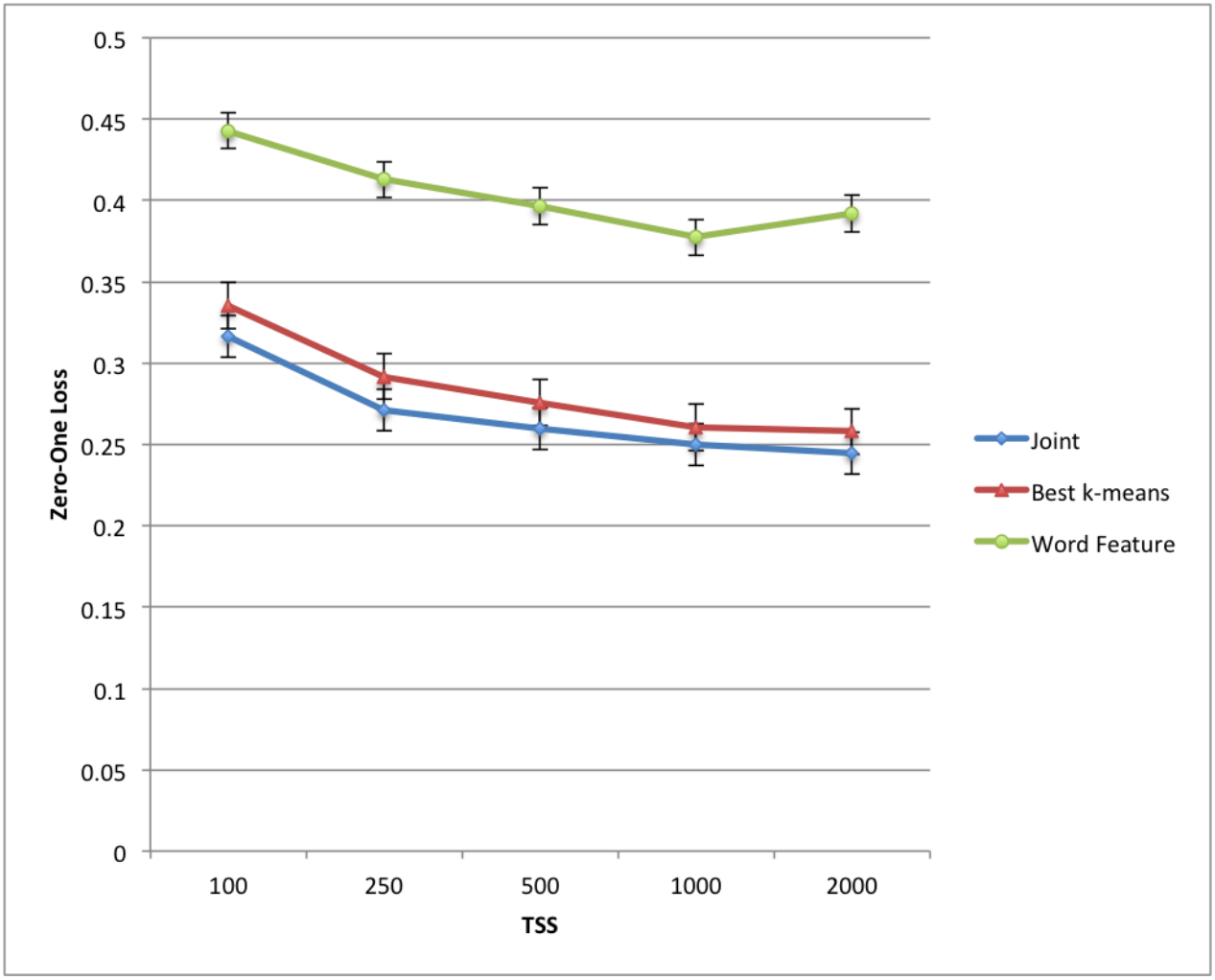

Figure 5: Q3(a) Plot of learning curves for A, B, and C, including error bars

(b) Formulate a hypothesis about the performance difference between A and B. Discuss whether the observed results support the hypothesis.

Hypothesis: No matter how tss changes, clustered NBC should always perform much better than randomly selected word features.

Support: From the plot, we can easily see that there is huge gap between clustered topic features and randomly selected word feature. Even the bottom of randomly selected word error bar is not even close to the top of clustered error bar. It supports my hypothesis.

(c) Formulate a hypothesis about the performance difference between C and A/B. Discuss whether the observed results support the hypothesis.

Hypothesis: C should perform sightly better B and much better than A when tss increases.

Support: The plot supports my hypothesis. Because from the plot, we can see that the average of C is always better than B. But when tss is small, the error bar scopes have large overlap. When the tss increases, this overlap becomes smaller. However, considering the error bar overlap, we are not confident enough.