Analyzing COVID-19 Search Trends and Hospitalization

Zilong Wang, Kehan Li, Rebecca Zhang
Comp 551 Applied Machine Learning
Instructor: Siamak Ravanbakhsh

Abstract

In this mini project, we cleaned and merged two COVID-19 related datasets into one and visualized the search trends data. Principal Component Analysis (PCA) was used to reduce the data dimensionality and the data was clustered by k-means method. We also investigated the performance of two supervised learning frameworks, namely k-nearest neighbours (KNN) and decision trees (DT), on predicting COVID-19 hospitalization cases from related symptoms search. Data was split into train and validation sets based on regions and time. Furthermore, a test set and cross-validation were applied on the region-based split data. For creativity, we selected the states that have the most complete coverage of some popular symptoms to predict the hospitalization from these symptoms. Finally, we predicted each regions' hospitalization based on date -split. We found that the search trends of symptoms are related to the regions in the United State. Additionally, the KNN approach achieved slightly better accuracy than decision trees when predicting hospitalization cases on the time-based split data while the performance of KNN and DT on the region-based data varied from run to run since the test set was changed each run. The MSE for time-based data split is significantly lower than region-based data split.

Introduction

I. Background Information

Principal Component Analysis is used to reduce the dimensionality of a dataset consisting of a large number of features, while retaining as much as possible of the variation present in the dataset. K-means clustering is a commonly used data clustering for performing unsupervised learning tasks. This method partitions n observations into k clusters in which each observation belongs to the cluster with the nearest cluster centers. K-nearest neighbor (KNN) and decision trees (DT) are both commonly used supervised learning models which were used to predict COVID-19 hospitalization cases in our project. While KNN assigns a predicted value to a new observation based on the mean of its k "Nearest Neighbors" in the training set, DT divides the feature space into regions using a tree structure and assigns a prediction label to each region.

II.Two Datasets

There are two datasets used in this project, one is Search Trends dataset, which shows the weekly search trends of 422 symptoms in 16 regions of the United States. Another one is COVID hospitalization dataset, which includes time series data for COVID-19 cases, deaths, tests, hospitalizations, discharges among other attributes over the world. We only used 'hospitalized_new', 'open_covid_region_code' and 'date' data in the US for this dataset.

III.Task Description and Important Findings

In order to compare the data across different regions, we re-normalized the symptoms using the mean of symptoms in each region. We cleaned and merged the above two datasets into one and visualized how the distribution of search frequency of each symptom aggregated across different regions changed over time. We used Principal Component Analysis (PCA) to reduce the retained 119 features (i.e. dimensions) after cleaning to only 2 features in order to visualize them. K-means was performed to evaluate possible groups of features for raw and PCA-reduced data. K-nearest neighbours (KNN) and decision trees (DT) were performed to predict COVID-19 hospitalization cases from related symptoms search. The merged dataset was split into train and validation sets using two strategies - based on regions and based on time. Furthermore, a test set and 5-fold cross-validation were applied on the region-based split data. For originality, we selected four states that have the most complete coverage of the 20 most popular symptoms to predict the hospitalization from these 20 symptoms. Each region's hospitalization was also predicted by KNN and DT.

From these tasks, We found that symptoms are somewhat related to the regions, but not so relevant to dates. In comparison with the DT model, the KNN approach achieved slightly better accuracy when predicting hospitalization cases on the time-based split data. However,

when the data was split based on regions, the performance of KNN and DT varied from run to run since the test set was changed each run. Overall, The MSE for time-based data split is significantly lower than region-based data split. The accuracy of KNN and DT didn't get improved using four mostly covered states and twenty most popular symptoms. We eventually predicted the hospitalization of each state and found that the performance of models largely depends on which region we were working with.

Datasets

I. Data Acquire

Search Trends dataset and COVID hospitalization cases dataset are used in the project. **The date of the datasets we ended up using is 2020.10.03**.

-Search Trends dataset shows the daily search trends of different symptoms in different regions of the United States, covering 16 states and 422 symptoms.

-The COVID hospitalization cases dataset includes time series data for COVID-19 cases, deaths, tests, hospitalizations, discharges among other attributes over the world.

II. Data Analysis

Among the 422 symptoms in the Search Trends dataset , 303 symptoms have a search trend of NaN. We deleted these missing features and retained 119 symptoms.

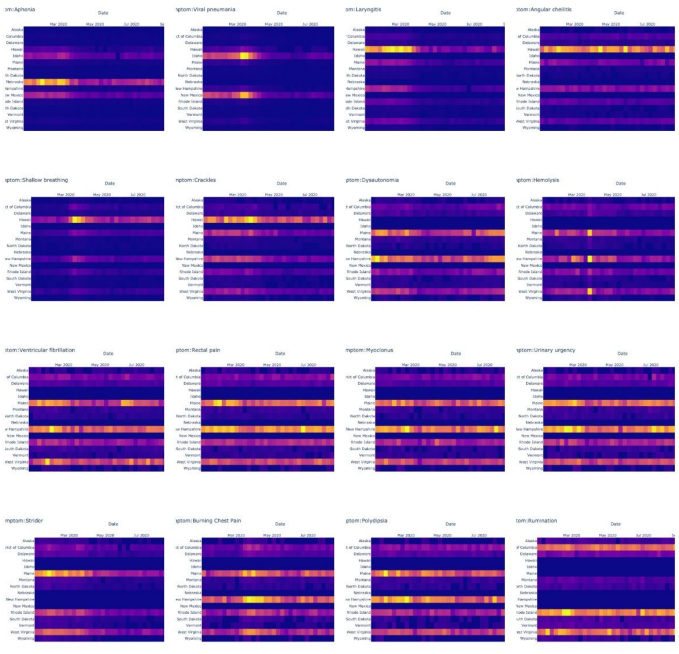
The Search Trends dataset was provided in a way such that there is an unknown normalization constant to divide the raw values of search trends for each region. We need to re-normalize the data in order to compare the values of symptom popularity across regions. Here, our acquired data in each region was divided by the mean of the symptoms in each region since the mean can represent an average value of each region and it cannot be zero so that we didn't need to worry about the infinity issue.

'hospitalized_new', 'open_covid_region_code' and 'date' are the only features we used in the COVID hospitalization cases dataset. In this project, we only use the data whose region is the US, which has a total of 11858 items. The start date of the data in the United States is 2020.03.02, we merge the date according to the last day of the week, and the hospitalized_new value is the sum of 7 days.

When we merged the two datasets, we only kept the 16 states shared by them. The merged data set has a total of 481 data, with dates from 2020.03.02 to 2020.09.21.

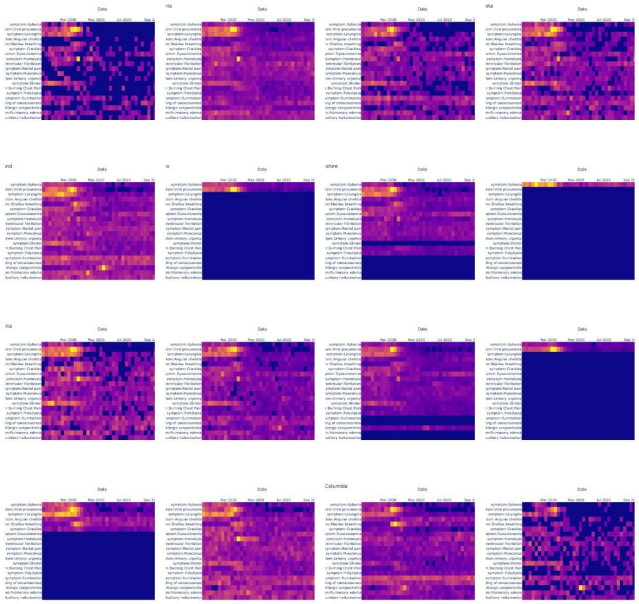
III. Data process

First, we selected the top 20 symptoms with the most data and used a heat map to visualize the popularity of each symptom over time in different regions. From the figure, we find that some symptoms are mainly concentrated in a few states, while others are concentrated in other states. In addition, some symptoms have their search trends concentrated before March 30th, 2020, with a peak value between March 16th and March 30th.



In order to observe the impact of date on symptom popularity, we visualized the search trends of different symptoms in each state from January to September 2020. The figure reflects that in a few areas there are only 1-2 symptoms, which is consistent with our previous predictions. However, we cannot judge from the graph whether date affects the change of symptoms, we can only speculate from the color change that the

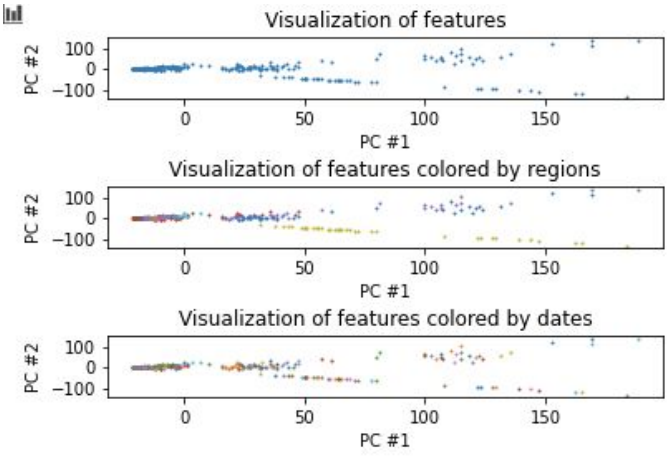
popularity of a few symptoms will slightly decrease over time.



Results

I. **A visualization of the search trends data in lower dimensions**

We used the PCA library from sklearn. By calculating variance and cumulative variance, we found out that the optimal number of principal components is around 4, which can explain 95% cumulative variance.
Since we need to visualize the dataset, we select reduced dimension = 2. In the clustering method, we will still use # PC =4.

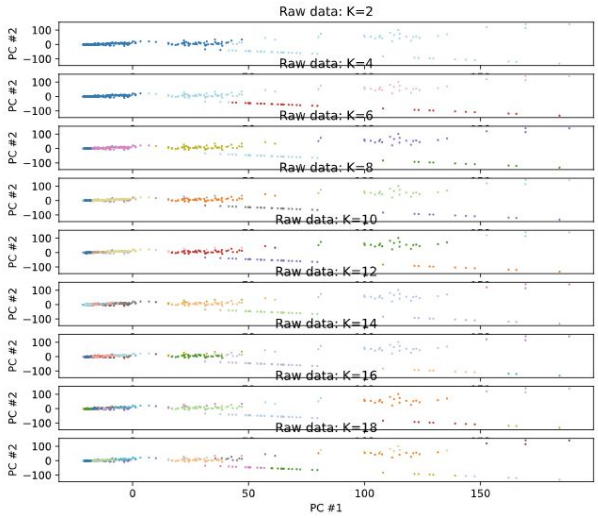


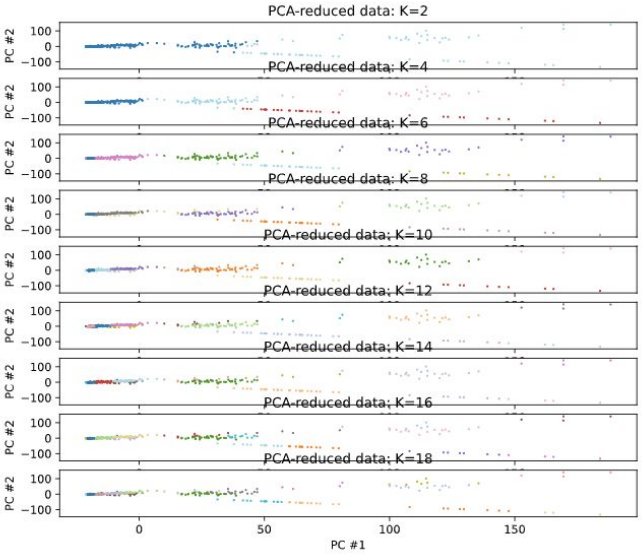
Marking the features with the region, we find that some clusters are of the same color, indicating that features are kind of related to the symptoms.

Then mark the features by date, and the result shows that the date has no obvious aggregation behavior, so we think that date has no effect on symptoms.
Therefore, using PCA, we know that symptoms have obvious clustering phenomenon, region and symptoms are related, but date and symptoms are not.

II. Same plot as above but with cluster labels for each data point to illustrate the clustering results

K-means from sklearn.cluster was used. We selected the optimal reduced dimensionality, which is 4 to cluster. From the plots, we can see that the best K for raw data is 4 and the besk K for PCA-reduced data is also 4.



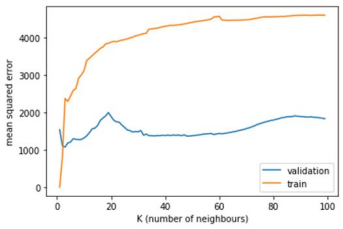


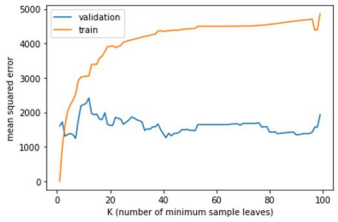
The clusters remained consistent for raw data as well as PCA-reduced data. The clusterings are the same, they just swapped the cluster indices(color).

Compared to the plot above, some of the clusters for PCA-reduced and raw data were similar to the groups in the visualization of features colored by regions. We inferred that the cluster labels were somewhat related to the regions.

III. A comparison of regression performance (mean squared error or mean absolute error) *between KNN and decision trees on the aforementioned cross-validation schemes*

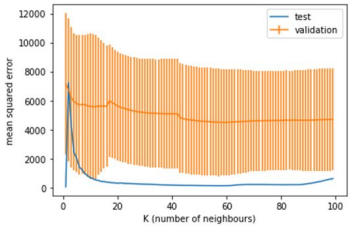
In the case of data split based on Date, the best performance of KNN with number of neighbors=3 and MSE=1071.5046296296296 is slightly better than DecisionTree with number of minimum sample leaves= 7 and MSE=1261.2578812814268

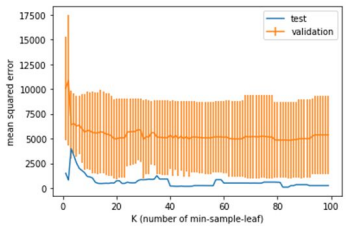




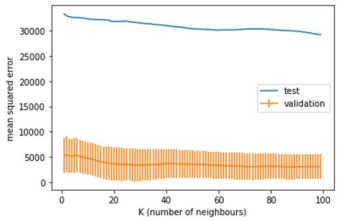
In the case of data split based on Region, in addition to train/validation split, we first shuffled data based on regions and randomly selected one state's data as our test set. Therefore, everytime we run the code, a different training/validation scheme is generated, and the performance varys from run to run, for example,

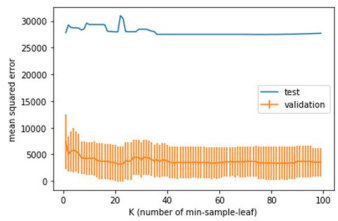
When US-AK is opt out as the test set, best KNN K= 59, Best KNN MSE=4487.780974380721, Best DT K= 85, Best DT MSE=4839.378466360133



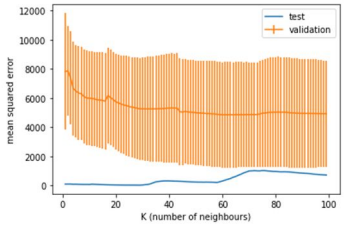


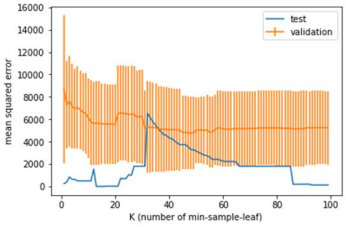
When US-NE is opt out as the test set, best KNN K= 99
, Best KNN MSE=2881.8046051068277. Best DT K= 35
, Best DT MSE=3106.74901411999





When US-DC is opt out as the test set, best KNN K= 64
, Best KNN MSE=4845.224147612371 Best DT K= 48
, Best DT MSE=4767.442161298257



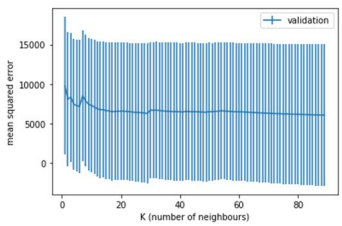


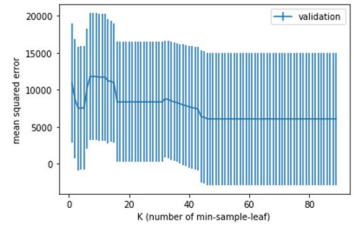
Note that the performance of the model on the test set also varies from run to run. We used all the regions except the test region as our training data and predicted test set. This is because each test set may have a very different data distribution from others. While some may have a very good performance at as low as 2-digit MSE, some may have a very poor performance.

Overall, the MSE for date-based data split is significantly lower than region-based data split.

This makes us rethink the impact of different regions on the performance. Based on the visualization in Task 2, we found that four states, namely Rhode Island, Delaware, DC and West Virginia, have the most complete coverage of the 20 most popular symptoms. Therefore we decided to model these regions' hospitalization with the 20 most popular symptoms using 4-fold cross validation. However, the MSE for both KNN and DT didn't improve.

BEST K for KNN REGION split: 88
Best MSE for KNN REGION split: 6044.563990921865
BEST K for DecisionTree REGION split: 46
Best MSE for DecisionTree REGION split: 6074.441564703759

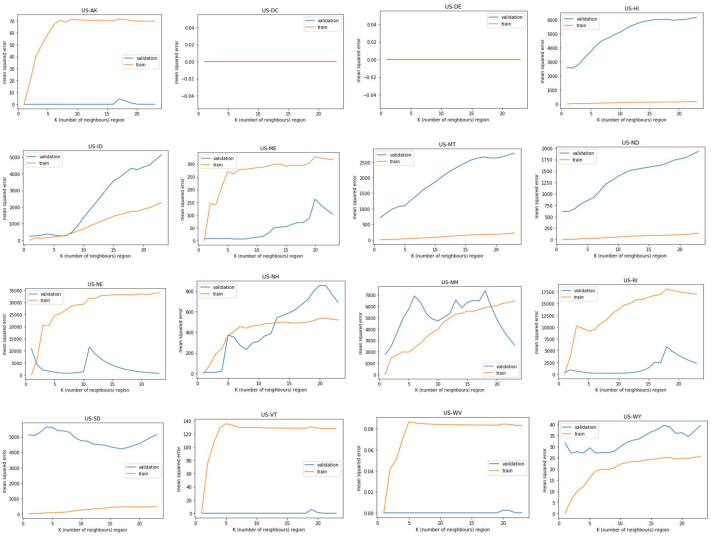




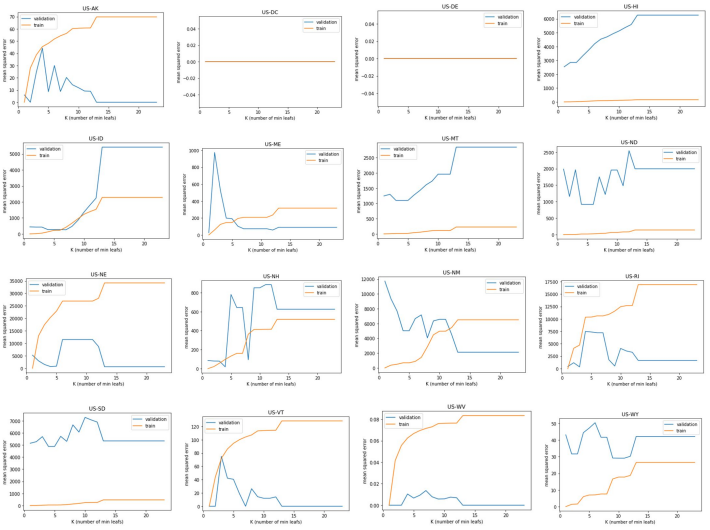
Finally, we predict each region's hospitalization based on date split using both KNN and DecisionTree, which

showed a highly divergent distribution of MSE across all states and there is no better model in terms of performance. In the same region, both models got a similar best MSE for each state. Moreover, for most states, the MSE is significantly reduced compared to date-based data split and region-based data split.

KNN prediction for each region



DecisionTree prediction for each region



Discussion and Conclusion

In this project we found that it is necessary to separately predict the hospitalization of different states with their search trend data, since each region has its own best model. However, date-based data split will have a less impact on the prediction than region-based split, which makes sense because in the same area the search trend is relatively consistent across different times.

In the future, since we've deduced that region-based split data had more impact on the prediction of hospitalization, we wonder if it is the weather condition of different states leading to the difference. We may investigate the performance of the supervised learning models based on weather-split data.

Statement of Contributions

Zilong processed the hospitalization dataset, brought it to weekly resolution and merged two datasets. Zilong also completed the main part of task 3, including data split, KNN model-building for both date and region based data split and DecisionTree model-building for region based data split, model performance comparison and predictions of four most complete regions and each region. Kehan was responsible for datasets downloading, loading to Pandas dataframes, cleaning and visualizing the evolution of popularity of various symptoms across different regions over time. Rebecca re-normalized the symptoms in each region and performed PCA to reduce data dimensionality, visualized the features and explored K-means clustering.

References

[1] (2002) Introduction. In: Principal Component Analysis. Springer Series in Statistics. Springer, New York, NY. https://doi-org.proxy3.library.mcgill.ca/10.1007/0-387-22440-8_1

P.S: Please run our code in Jupyter Notebook as part of the code doesn't work on Google Collab.