offsite test for university recruitment

https://github.com/TerryChanHoYin/offsite-test.git

Q3

Question A

How many visits are in the data set?

357912

The following SQL command is used to query the total number of visits.

```
SELECT COUNT(*) FROM test_data
```

How many distinct users are in the data set?

64265

The following SQL command is used to query the number of distinct users.

```
SELECT COUNT(DISTINCT user_id) FROM test_data
```

How many distinct pages are in the data set?

15163

The following SQL command is used to query the number of distinct pages.

```
SELECT COUNT(DISTINCT page_id) FROM test_data
```

Which hour gives the smallest number of visits?

04:00:00-04:59:59

Which hour gives the largest number of visits?

12:00:00-12:59:59

The following SQL command is used to query the number of visits during an hour, in a “brute force approach”.

```
SELECT COUNT(*) FROM test_data WHERE surf_time >= '00:00:00' AND surf_time < ='00:59:59'
```

Which page has the largest number of visits in the data set? What is the corresponding number of visits?

The page with page id 3897, corresponding number of visit is 26625

The following SQL command is used to the page with largest number of visits and the corresponding number of visits.

SELECT page_id, COUNT(page_id) FROM test_data GROUP BY page_id ORDER BY COUNT(page_id) DESC

To perform the analysis in question A, the phpMyAdmin in XAMPP is used to do query in MariaDB, a fork of MySQL.

Question B

Please open the visualization.ipynb by jupyter notebook.

The following SQL command is used to query the number of new visitors in an hour. The surf_time lower bound, here 00:59:59, in the “inner command” is needed to be changed when another particular hour is queried.

SELECT COUNT(DISTINCT user_id) FROM test_data WHERE surf_time >= '01:00:00' AND surf_time < ='01:59:59' AND user_id NOT IN (SELECT DISTINCT user_id FROM test_data WHERE surf_time >= '00:00:00' AND surf_time < ='00:59:59') ORDER BY surf_time ASC

Question C

1) Stratified 5-fold cross validation is used to test the training accuracy. The accuracy is unexpectedly high, when it is compared with my experience on English text classification. According to the returned value, the training accuracy can achieve 99.1%.
2) The parameter of support vector machine, or specifically, C-SVM, is the penalty term and the choice of kernel method. According to previous knowledge, problem is natural language process is usually linearly separable, so linear kernel, or say, no kernel is used.

   When C=0.01, the training accuracy is around 99.1%.

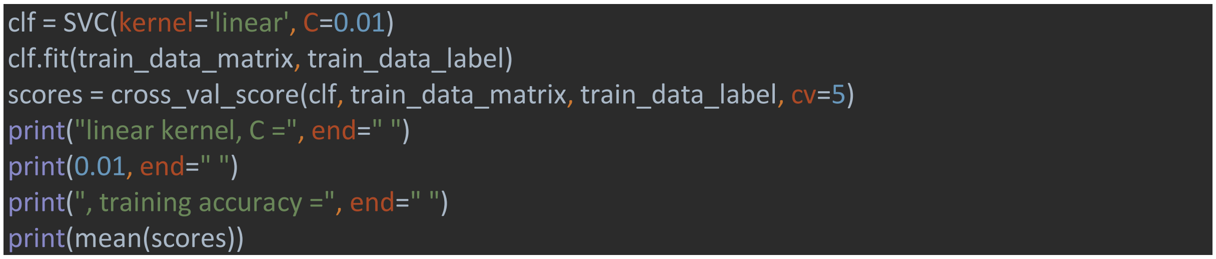
   When C=10, the training accuracy is around 98.9%.

   When C=1, the training accuracy is around 98.9%.

   When C=0.001, the training accuracy is around 98.9%.

   Therefore, C=0.01 is chosen.

   The following code is used to check the training accuracy.



```
clf = SVC(kernel='linear', C=0.01)
clf.fit(train_data_matrix, train_data_label)
scores = cross_val_score(clf, train_data_matrix, train_data_label, cv=5)
print("linear kernel, C =", end=" ")
print(0.01, end=" ")
print(", training accuracy =", end=" ")
print(mean(scores))
```

3)

Please open submission.csv to check the prediction.

Support vector machine, or specifically, c-svm. If we assume the data points are linearly separable in hyperspace, then there should be a hyperplane that can be used to separate them. So, support vector machine classifier, by its nature, is a binary classifier. If the hyperspace has n dimension, then the hyperplane, or called decision boundary, is determined by $\vec{w} \cdot \vec{x} + b = 0$, $\vec{w}$ is the weight vector, with length equals to n and is perpendicular to the decision boundary, $\vec{x}$ is data point, b is biased term.

To construct $\vec{w} \cdot \vec{x} + b = 0$, the following steps is needed. First, consider there is a datum $\vec{x}_+$ which make

$\vec{w} \cdot \vec{x}_+ + b = 1$ (1)

and there is another datum which make

$\vec{w} \cdot \vec{x}_- + b = -1$ (2)

, and perform reduction to get

$\vec{w} \cdot (\vec{x}_+ - \vec{x}_-) = 2$ (3)

$\hat{w} \cdot (\vec{x}_+ - \vec{x}_-) = \frac{2}{\|\vec{w}\|}$ (4)

The physical meaning of $\frac{2}{\|\vec{w}\|}$ is the following: consider the decision boundary, there is a space around the decision boundary that no data point appears and $\frac{2}{\|\vec{w}\|}$ is the width of that space. To find the maximum margin, it is to find the minimum of $\frac{2}{\|\vec{w}\|}$. So, it comes to the second step, use Lagrange multiplier to find the minimum of $\frac{2}{\|\vec{w}\|}$. The related equation is

$L = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{N} \lambda_i((\vec{w} \cdot \vec{x}_i + b)y_i - 1)$ (5)

where λ_i is the Lagrange multiplier, and $(\vec{w} \cdot \vec{x}_i + b)y_i - 1$ is generalized from (1) and (2) and i is the number of support vectors. Then,

$\frac{\partial L}{\partial \vec{w}} = \vec{w} - \sum_{i=1}^{N} \lambda_i y_i \vec{x}_i = 0$ (6)

$\frac{\partial L}{\partial b} = \sum_{i=1}^{N} \lambda_i y_i b = 0$ (7)

Substitute (6) and (7) back into (5), we get

$L = -\frac{1}{2}\sum_{j=1}^{N}\sum_{i=1}^{N} \lambda_i \lambda_j \left((\vec{x}_j \cdot \vec{x}_i) y_i y_j\right) + \sum_{i=1}^{N} \lambda_i$ (8)

(8) is a convex problem and can be solved by quadratic equation. Solving (8) can be used to determine which data points have $\lambda \geq 0$ and then use these data pointers, which is also called support vectors, to determine $\vec{w}$ and b in $(\vec{w} \cdot \vec{x}_i + b)y_i = 1$. Be specific, $b = -\frac{1}{N}\sum_{i=1}^{N} \vec{w} \cdot \vec{x}_i$.

The testing is quite intuitive. Plug in testing data point $\vec{x}_{testing}$ into $sign(\vec{w} \cdot \vec{x}_i + b)$, and see what the result is.

Challenges Encountered

My major challenge in question A is the fear that I cannot finish the SQL query task because SQL and database is fresh new for me. I have never learnt SQL before. After I read the requirements on SQL, my first reaction was to search W3C website and had a quick start on SQL. Then, I went to HKUST computer science department course website and search database course lecture notes. After that, I tried to solve the problem one by one, and each small success encourage me to finish question A.

Another challenge come from conducting query by SQL. I am inspired by the exercise of HKUST database course to solve the problems here.

The challenge of question B is debugging. Just after starting question B, I remember that matplotlib do not provide graph with hovering interaction. So, I search plotly, a Python library on data visualization. I used plotly once in Deepsky. But, debugging the data visualization code consumed more time than I expected. Finally, I slept much less than normal.

In question C, the challenge comes from the understanding of SVM. After several skims on the lecture note, I started to search the lecture video on youtube. Lecture videos from MIT and Caltech help me a lot to understand SVM. Since online lecture video allows pauses, I find that many previous misunderstood concepts may be from careless listening. The time used on choosing parameter is less because previous knowledge tells me text classification is usually linearly separable and best penalty term is usually the minimum in a convex training error.

Data set

I deduce the reason to investigate the total and new view numbers of different pages is to study how Hong Kong people read newspaper, because there is comparison between total and new view number. And, the view numbers are all integers. Since a special and less frequent event, e.g. typhoon, may affect how we read newspaper significantly on a single day, it is recommended that use an averaged view number from a period to instead.

Question set

I supposed some implementation details of classifiers will be asked, e.g. the major part of computation in sklearn SVM use C/C++.