

COMP9321 Data Services Engineering

Plan of assignment 3

Group name: Macau casino

Group member: Jiachen Li, Patrick Wang, Lesi Huang, Xiyao Li, Chong Zu

Introduction

Aim: In this assignment, we are going to establish a data service to analyze factors which have been effecting Australian GDP during the past ten years, and predict the trend of the GDP according to these factors in next 1-5 years.

Datasets: Kaggle Datasets

(<https://www.kaggle.com/datasets>)

NSW Government Public Datasets

(<https://data.nsw.gov.au/data/dataset>)

Australian Federal Government Public Datasets.

(<https://search.data.gov.au/>)

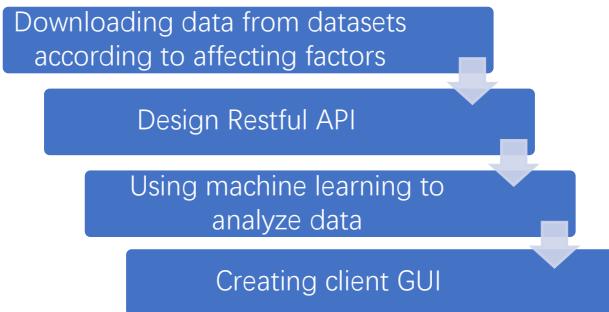
Communication channel: group chat in Wechat.

Code repository: <https://github.com/CocaColaSoap/australia-GDP-analysis>

Task

Members role: Our team would be divided into two sub-teams. The first team (Jiachen Li and Chong Zu) would be responsible for Restful API and GUI Design. The second group (Patrick Wang, Lesi Huang, Xiyao Li) would work on machine learning model to integrate, process and analyze data collection.

Work flow:



Machine learning model selection:

At this stage, we are considering three machine learning models .

1. First model is **linear regression based method**. The basic assumption is that the output variable (a numeric value) can be expressed as a linear combination (weighted sum) of a set of input variable (which is also numeric value).

$$y = w_1x_1 + w_2x_2 + w_3x_3 \dots$$

The strength of Linear model is that it has very high performance in both scoring and learning. The Stochastic gradient descent-based learning algorithm is highly scalable and can handle incremental learning.

The weakness of linear model is linear assumption of input features, which is often false. Therefore, an important feature engineering effort is required to transform each input feature, which usually involved domain expert.

2. We also consider to use **Neural Network model**. Neural Network can be considered as multiple layer of perceptron (each is a logistic regression unit with multiple binary input and one binary output). By having multiple layers, this is equivalent to : $z = \text{logit}(v_1.y_1 + v_2.y_2 + \dots)$, while $y_1 = \text{logit}(w_{11}x_1 + w_{12}x_2 + \dots)$. This multi-layer model enables Neural Network to learn non-linear relationship between input x and output z . The typical learning technique is "backward error propagation" where the error is propagate from the output layer back to the input layer to adjust the weight.

Notice that Neural Network expect binary input which means we need to transform categorical input into multiple binary variable. For numeric input

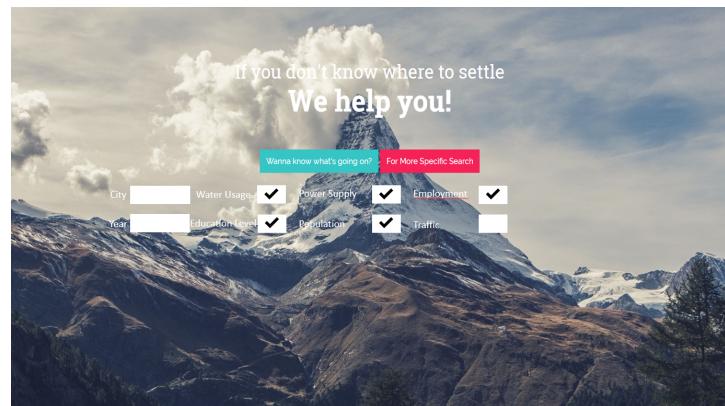
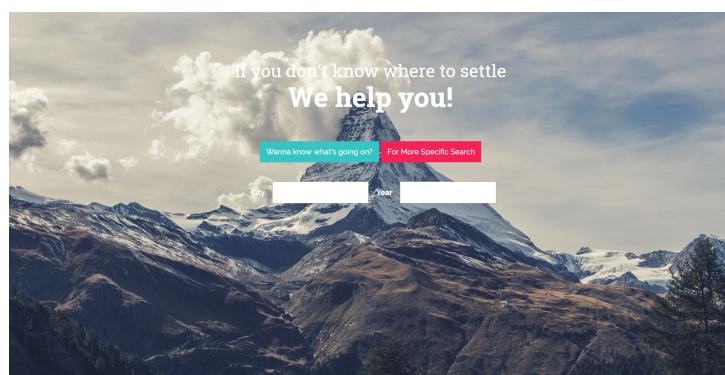
variable, we can transform that into binary encoded 101010 string. Categorical and numeric output can be transformed in a similar way.

3. The last model we want to try is **Nearest Neighbor model**.

The idea is to find K similar data point from the training set and use them to interpolate the output value, which is either the majority value for categorical output, or average (or weighted average) for numeric output. K is a tunable parameter which needs to be cross-validated to pick the best value.

The strength of K nearest neighbor is its simplicity as no model needs to be trained. Incremental learning is automatic when more data arrives (and old data can be deleted as well). Data, however, needs to be organized in a distance-aware tree such that finding the nearest neighbor is $O(\log N)$ rather than $O(N)$. On the other hand, the weakness of KNN is it doesn't handle high number of dimensions well. Also, the weighting of different factors needs to be hand tuned (by cross-validation on different weighting combination) and can be a very tedious process.

Interface design:



At this stage, we design an interface which is very convenient for users to input different affecting factors in order to get the statistics of GDP in past ten years or get the prediction of GDP in next 1-5 years.

For the most important filter conditions, two searching blanks are designed which are “City” and “Year” respectively. Each searching blank provides many options or different selection interval.

As the output part, the statistics during the past period or predicted in the future are expected to be shown as graphs such as bar chart and line chart, which are straightforward for trends demonstration.

Task progress:

In Week 11, the designing of the task goal is achieved and a specific plan about work follow and separated tasks’ distribution is used to guide team members to work efficiently and orderly.

At the end of Week 12, several target tasks are finished.

The first finished task is data collection, which is the base and the beginning point of the whole project. During the process, except the datasets from assignment description, ABS (Australian Bureau of Statistics) has also been used for data selection. It is useful when finding some specific factors which can influence the fluctuation of GDP of a city, such as population, mark for high shopping level accessibility and GDP per capita and so on. In our selected data, the statistics of 7 different factors of 8 capital cities in Australia are recorded. After this, a data collection in mlab is established for data storage.

As for learning model selection, we have tried three kinds of learning model which are Linear Regression model, Neural Network Regression model and Nearest Neighbor Regression model respectively.

here are two types of Nearest Neighbor Regression model: K-Neighbors regression, which is based on k nearest neighbors of each query point, and Fixed-Radius Neighbors Regression, which is based on the neighbors within a fixed radius of the query point. So now we have four models to choose from.

Criterion:

There are a number of indicators to show the goodness-of-fit of various models. Here we choose R square. R square is defined as

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

It is interpreted as the ratio between model-explained error and total actual errors. The value of R square is between 0 and 1. In general, the higher the R square value, the better the model fits the data.

Result:

We use the four models respectively to fit the several cities' GDP data, and compare their R square values:

City	R Square			
	Linear Regression	Neural Network	K-Neighbors	Fixed-Radius Neighbors
Sydney	0.9601	0.9201	0.8181	0.7909
Perth	0.9567	0.9548	0.9342	0.9758
Canberra	0.8259	0.7932	0.7117	0.8136
Brisbane	0.8197	0.6677	0.7534	0.8566

We found that Linear Regression Model has an generally better performance than others. So we decide to use Linear Regression Model to build our data service.