conclusion:

通过不同的machine learning地方法拟合数据，得到每种方法所对应的函数预测数据，并通过计算数据拟合度，对每种方法进行评估。选取拟合度适中的方法，作为预测未知数据的最终方法。

Through using a set of methods intended for regression in which include Support Vector Regression (SVR), Decision Tree Regression, Linear Regression, K Neighbors Regression, Random Forest Regression, Gradient Boosting Regression, we obtained 6 different learning curves. And the target values we use in our machine learning process are collected from the query module. To evaluate the methods, we calculate the fitting score for each of them. After that, we chose a method which is neither over fitting nor under fitting; and use it as our predict method for our data set. Based on this method, we can predict any data in our data set interval.

objective:

1.How to fit the data set in to a function curve by using regression methods in machine learning.

2.How to evaluate the quality of the methods we use?

impact:

Community: The learning curves can help the meteorological department to predict weather more accurate.

Scientific: In the machine learning module, we discuss several regression methods for the weather prediction. Based on the date set of the weather changed per day, this module find out a suitable machine learning methods for the weather liked data set.

methodology:

根据从上一模块获得的数据信息（id，时间信息，温度）

machine learning method:

**Linear Regression:**

Linear regression is a linear approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X.

In our data set fitting, linear regression method is under fitting in the search.

**SVR(Support Vector Regression):**

SVR回归，就是找到一个回归平面，让一个集合的所有数据到该平面的距离最近。

The method of Support Vector Classification can be extended to solve regression problems. This method is called Support Vector Regression. The basic principle of SVR is to find a regression plane which can make all the distance of the data set to the plane is the closest to the plane.

The model produced by support vector classification (as described above) depends only on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin. Analogously, the model produced by Support Vector Regression depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction.

**Nearest Neighbors Regression:**

Neighbors-based regression can be used in cases where the data labels are continuous rather than discrete variables. The label assigned to a query point is computed based the mean of the labels of its nearest neighbors.

K Neighbors Regression is one of the Nearest Neighbors Regression which implements learning based on the k nearest neighbors of each query point, where k is an integer value specified by the user.

**Decision Tree Regression：**

决策树（decision tree）是一种基本的分类与回归方法。

决策树学习通常包括3个步骤：特征选择、决策树的生成、决策树的修剪

Decision tree regression is a basic classification and regression method.

It usually consists of 3 steps:

1. Feature selection: using information gain criterion algorithm select feature；

2. Decision tree generation: using classical ID3 algorithm to generate tree;

3. The pruning of the decision tree: in order to prevent over fitting phenomenon.

The generation of the decision graph corresponds to the local selection of the model, while the pruning of the decision tree takes into account the global minimum selection.

**Random Forest Regression:**

随机森林是对决策树的拓展，用随机的方式建立一个森林，森林里面有很多的决策树组成，随机森林的每一棵决策树之间是没有关联的。预测值为叶节点目标变量的加权均值

Random forests regression is the expanding of the decision tree regression. It establishes a forest with many decision trees in it in random way. There is no correlation between every decision tree in the random forest. The predicted value is the weighted average value of the target variable of the leaf node.

**Gradient Boosting Regression:**

与梯度学习相比，Gradient Boosting 是在迭代的时候选择梯度下降的方向来保证最后的结果最好。

Gradient Boosting Regression is the expanding of the decision tree regression.

Compared with the traditional gradient learning, Gradient Boosting Regression chooses the direction of gradient descent to ensure that the final results are best at the time of iteration.

**Method Evaluation:**

To evaluate these methods, we calculate the fitting score for each of them. Estimator in sklearn has a score method, which provides a default evaluation rule to solve the problem. Each score calculate the residual sum of squares ((y\_true - y\_pred) \*\* 2).sum() and the total sum of squares ((y\_true - y\_true.mean()) \*\* 2).sum().

After that, we chose a method which is neither over fitting nor under fitting; and use it as our predict method for our data set. The reason for under fitting is that the complexity of the model is too low to fit all the data well, and the training error is large. The reason for over fitting is that the model complexity is too high, the training data are too few, the training error is small, and the test error is big.

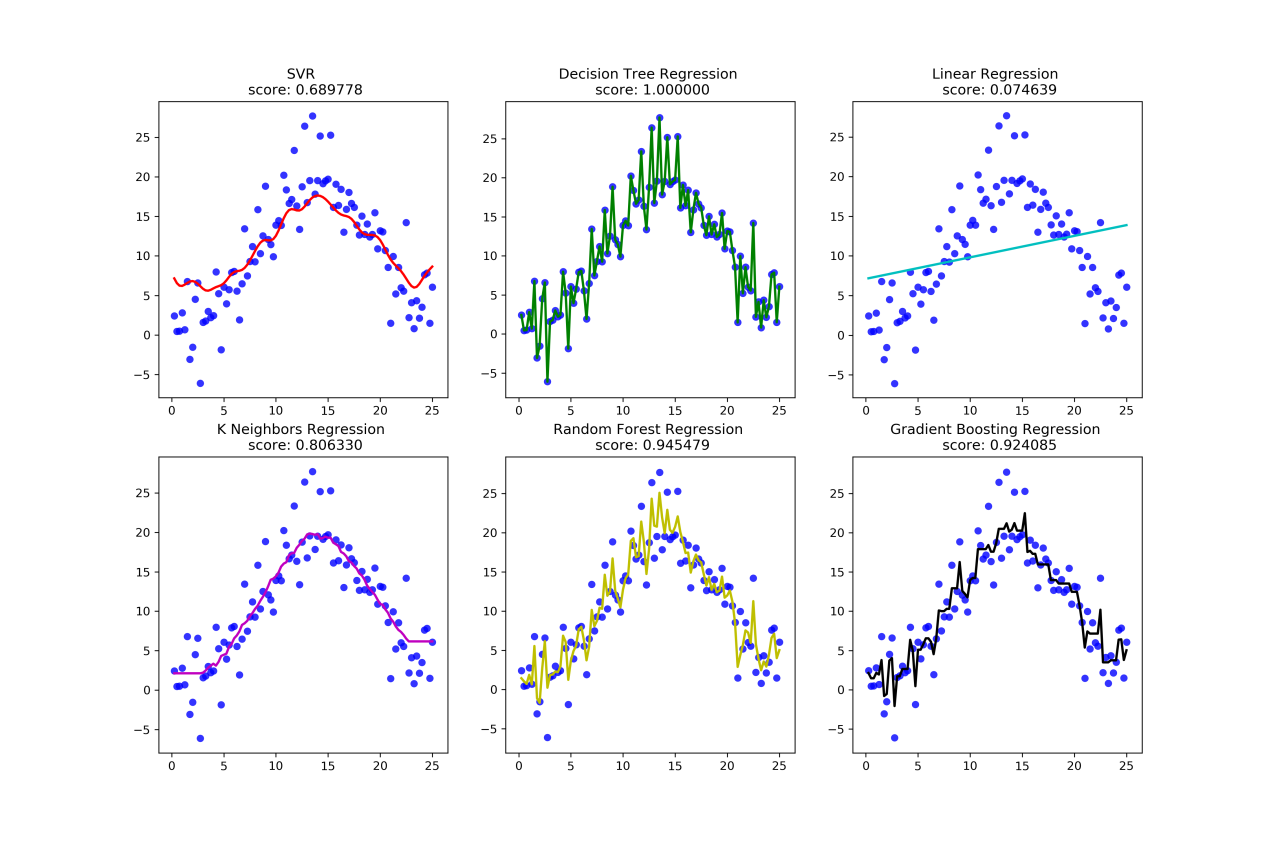
**Result:**

After collect the data, we use those 6 methods to fit the dataset in order to solve the query mission in the KD module.

Query B: temperature in a spell at particular location

In this query we use a set of methods intended for regression in which include SVR, Decision Tree Regression, Linear Regression, K Neighbors Regression, Random Forest Regression, Gradient Boosting Regression. The data set in this query is 2 dimensional(temperature, time).

The fitting result is showed below:

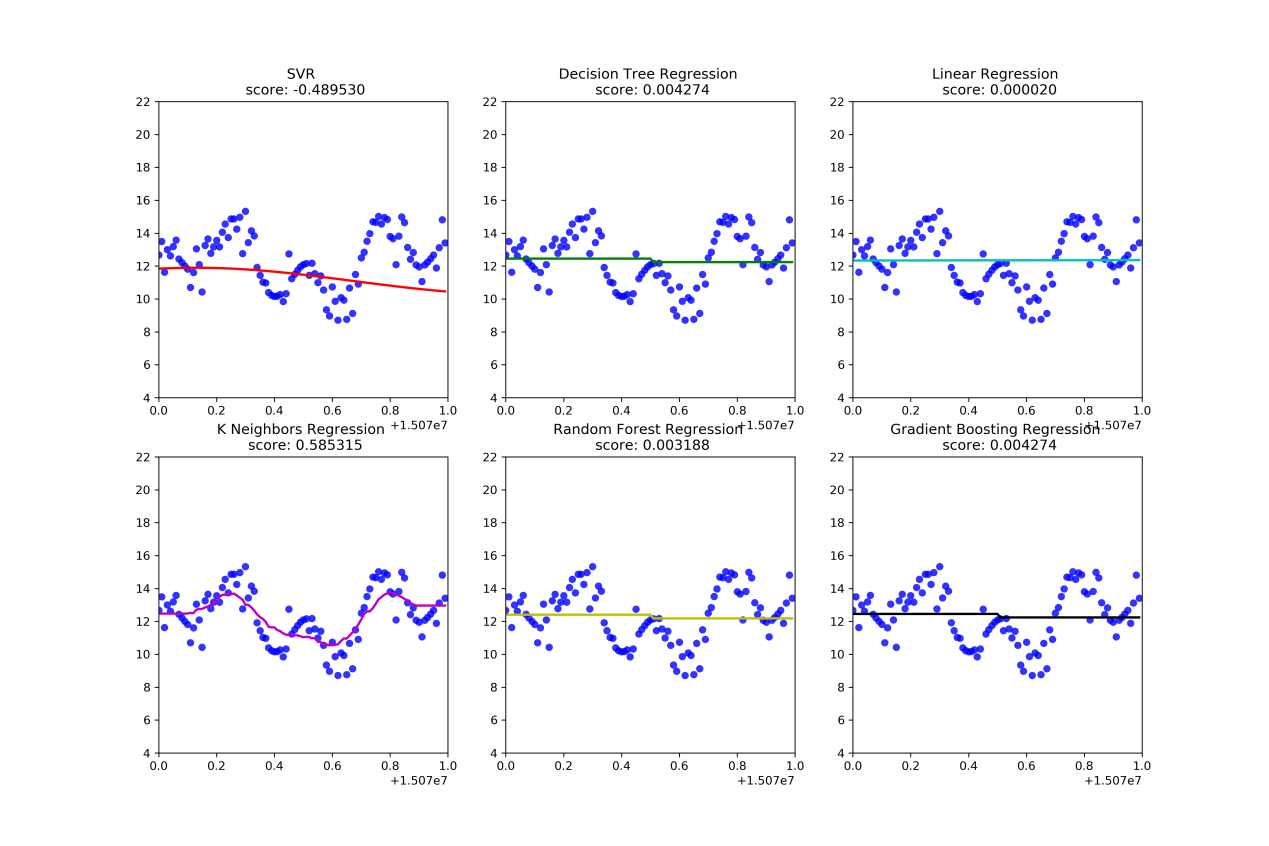


From the picture above, combine with the score we can find that SVR and K Neighbors Regression fit the data set better than other methods.

Query C: temperature in particular time in areas

In this query we use the same methods to fit the data set. The data set we use in this query is 2 dimensional(temperature, location id).

The fitting result is showed below:

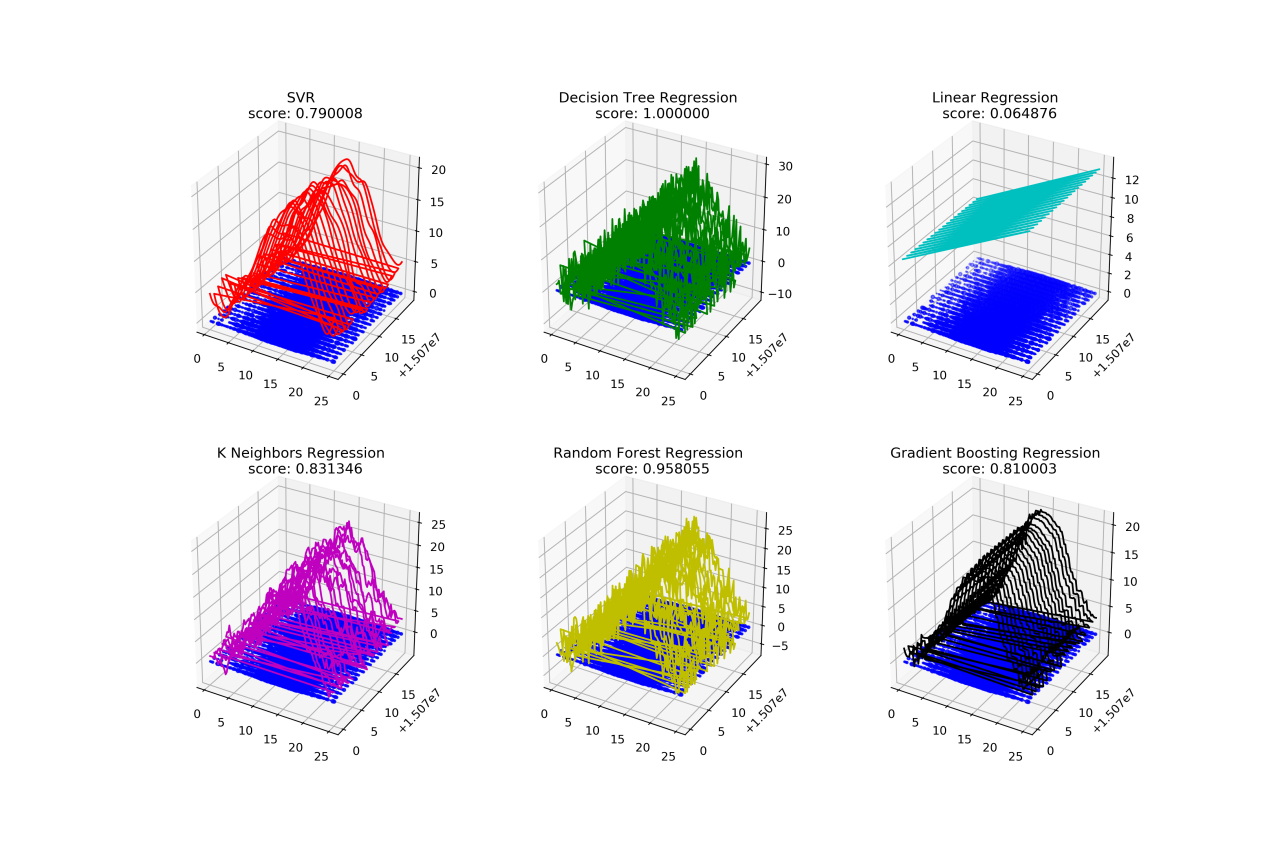


From the picture above, combine with the score we can find that K Neighbors Regression fit the data set better than other methods.

Query D: temperatures in a spell in areas

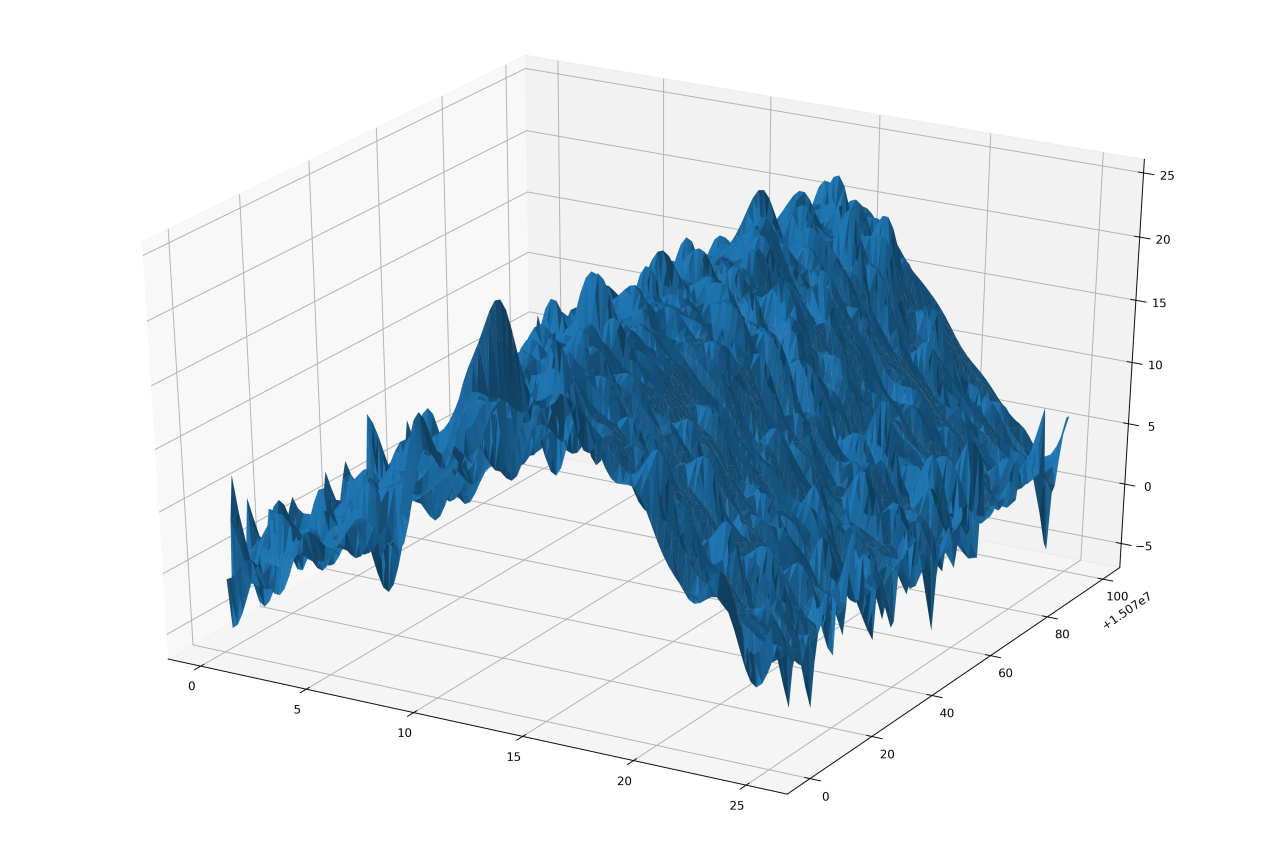
In this query we use the same methods to fit the data set. But the difference of the data set in this query is 3 dimensional(temperature, location id, time).

The fitting result is showed below:



From the picture above, combine with the score we can find that SVR fit the data set better than other methods.

But this method is still fitting in linear level; the result of the 3 dimensional data fitting isn't continuously and can't present the overall trend of data. So we fit the curve we get from the SVR. As a result, we get a hyper plane based on the data set and our prediction is continuously.



So that, we can predict any time's and id's temperature now.

key problem:

How to fit the discrete data to a continuously curve or a hyper plane.

reference:

李航 《统计学习方法》

1.“A Tutorial on Support Vector Regression”, Alex J. Smola, Bernhard Schölkopf - Statistics and Computing archive Volume 14 Issue 3, August 2004, p. 199-222.

2.M. Dumont et al, Fast multi-class image annotation with random subwindows and multiple output randomized trees, International Conference on Computer Vision Theory and Applications 2009