Introduction of Sklearn

Date: Apr - 2015

I start to learn scikit-learn for the package as my jigsaw of machine learning. Combining a book **<Statistical Learning Methodology>** (author: Dr.Li Hang in Huawei), I begin to step into the world of MachineLearning.

Write first, I will use English & Chinese in the blog, for the reason that I want to improve my poor English.

2015年四月, 我打算看看Sklearn这个著名的Python下的第三方机器学习工具. 另结合了手头上的李航博士的**<统计学习方法>**, 作为今年的学习目标的第一步.

install

The install page provides different ways of installing on PC\Mac\Unix.

For me (OS X), it's simple, just typing this line on Terminal:

```
pip install -U scikit-learn
```

Considering later examples code which also use numpy & matplotlib, my suggestion is that just `pip them all`.

sklearn包的安装很简单， 不管是Windows或者MacLinux环境，都提供了各自的方式。

另外在后续的例子学常用到numpy和Matplotliby也请一起安装.

- import convention

There have many useful sub-libraries in scikit-learn(i.e. sklearn for short). In most of cases, **import *** is absolutely not a good way to do that. Just import those libraries or specified functions to be used.

```
from sklearn import xxx
```

And it is a good way to keep your code in a elegant way not only using sklearn. In my another blog <suggestions to improve your python code> mentioned this.

在sklearn包中, 拥有各式各样的(现在还在不断扩充中)机器学习模型, 因此

如无特别注明， 建议用此形式对sklearn中的包进行引用。

在我以前瞎翻译的一篇文章<改善Python代码的建议>中，也有与此相关的建议.

dataset

When we want to learn machine learning, we need some dataset to analysis. The difficult for new guys (like me) is **We do not know what kind of datasets is match to the model.**

sklearn have prepared some typical datasets –– iris digits diabetes

… … (I will list all datasets to be used here.)

Digits

Pen-Based Recognition of Handwritten Digits Dataset

Properties: 10992 instances, 16 Attributes, without missing values.

This is easy understanding to build a estimator to classify.

Iris

Another famous dataset made up of iris of three related species(Setosa, Virginica and Versicolor)

Properties: 50 samples from each of three species, each observation has 4 features: A-width, A-length, B-width, B-length (A, B is part of flower, does not matter).

I think some students have used it for linear discriminant analysis. For me, I have used it in my SPSS exercise.

```
>>> list(data.target_names)
 ['setosa', 'versicolor', 'virginica']
```

diabetes

The diabetes dataset consists of 10 healthy variable related to diseases(age, sex, weights, blood pressure and so on) to measure more than 400 patients, and recording an indication of disease progression after one year as labelled target value.

Properties: 442 inputs, 10 features (-0.2 <= x <= 0.2),

*sklearn*已经内置了一些经典的数据集， 如用于判别分析的*IRIS,* 常用于*SVM*的*digits*等等。

load data, split data

All datasets can be loaded like this:

```
from sklearn import datasets

d_name = datasets.load_xxxx()
```

and then the data will be splitter into two parts (training set and testing set) in many situations. Take an example like this:

```
import matplotlib.pyplot as plt
from sklearn import svm
from matplotlib import style
style.use('ggplot')

model initilize
clf = svm.SVC(gamma=0.001, C=100)

model fitting
clf.fit(digits.data[:-1], digits.target[:-1])
print('predict result:
{predict_result}'.format(predict_result=clf.predict(
digits.data[-1])))

show the real img
img = digits.data[-1].reshape(8, 8)
plt.imshow(img)
plt.show()
```

- After this, long code will be linked to github for layout.

Machine Learning Introduction

Okay, start from here, we are just stepping across the door of scikit to machine learning.

What is the problem setting of machine learning?

> *In general, a learning problem considers a set of samples of data and then tries to predict the unknown data's properties.*

It sounds like a summary of absolutely right words, and **useless** (in my opinion). Especially for newbies, they just stick to the categories. So, depending on the purpose of proposed problem, we divide these categories:

- supervised learning

Here is the page of methods provided by scikit learn.

Shortly to say, it is a kind of problem to predict.

- classification: When we want to predict the number of unknown dataset's category, it is a classification problem. The target vector is always a finite number of discrete categories. Use case: to identify a client is good or bad for some credit-loan companies.

    - regression: In other cases, we want to predict a precise value, such as the price of house, the salary, and so on.

- unsupervised learning

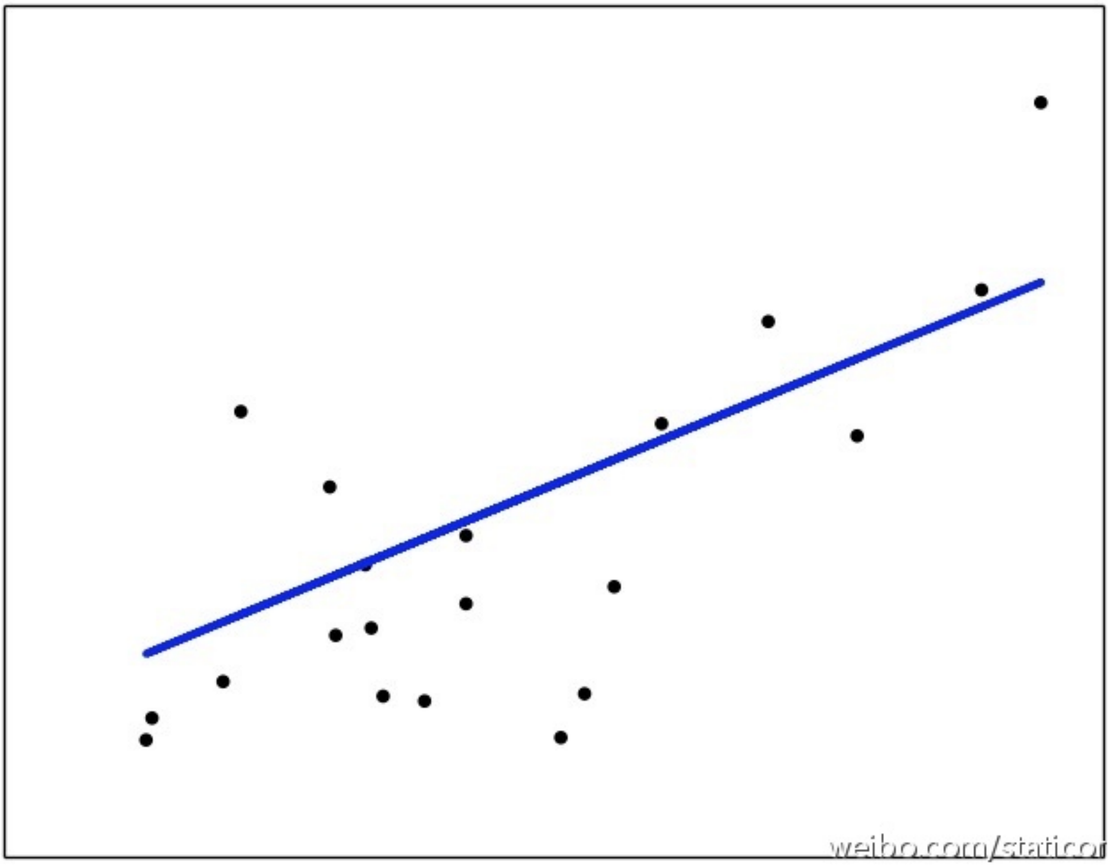
  If your input sets is all x without corresponding target valueslabel, and the goal in such problems maybe to find similar examples within the data, where it is called **clustering**, or to give prediction to the whole input, known as **density estimation**. Data is not labelled. We make labels based their relationship. I can take an example about students classification:

    - Think about we want to group them into 4 levels depending on their SAT scores. We may calculate the distribution of scores and analysis it. Then a new student can be grouped using its position in former distribution.

I think supervised learning is my first step to deep in.

Linear Regression

Let’s start with the simplest model(because almost all teaching books place this as Chapter 2) in supervised learning.



Use a 6X2 numpy array as example, to find the regression line in this case.

code in github, and I will give more details about LinearReg later in the section of Linear Model, including the score of regression equation, the coef_ and intercept_ item, …

这段代码是对简单线性回归OLS模型的操作， 关于LinearRegression的详细讲解, 将会在LinearModel中具体展开.

建模后regr封装了线性回归模型的系数， 模型评分（R Square) 等信息

End of this part

As the first section of this series, I prefer to make it a brief and helpful blog. I will update it for the future, please contact me if you have any suggestions.

结语: 作为起始, 我这里没有作过多的展开, 只是对内容和布局的尝试. 后续还会不断的改进, 任何建议都可与我联系.