k近邻（kNN)

作者：刘伟杰 日期：2015-11-27
参考：
[1]《统计学习方法》 李航 2012年3月第一版
[2]《机器学习实战》 Peter Harrington

1. k-nn描述:

给定一个训练数据集，对于新输入的实例，在训练集中找到与该实例最近的k个实例，统计这k个实例中多数的类别，就把该类别作为新输入实例的类别。

2. 参数:

1. 距离度量：

   Lp距离（欧式距离、曼哈顿距离等）、皮尔逊距离、夹脚余弦距离。。。

   $$L_p\left(x_i, x_j\right)=\left(\sum_{l=1}^{n}\left|x_i^{(l)}-x_j^{(l)}\right|^p\right)^{\frac{1}{p}}$$

2. k的选择:

   如果k较小，相当于模型过于复杂，容易过拟合；

   k过大，相当于模型变得简单，容易欠拟合。

   在应用中，k值一般取一个比较小的数值。通常采用交叉验证法来选取最优的k值。

3. python代码:

1. 我的实现:

   https://github.com/autoliuweijie/MachineLearning/tree/master/kNN (https://github.com/autoliuweijie/MachineLearning/tree/master/kNN)

2. sickit – learn:

   python:

```
#Import Library
from sklearn.neighbors import KNeighborsClassifier
#Assumed you have, X (predictor) and Y (target) for training data set and x_test(p
# Create KNeighbors classifier object model
KNeighborsClassifier(n_neighbors=6) # default value for n_neighbors is 5
# Train the model using the training sets and check score
model.fit(X, y)
#Predict Output
predicted= model.predict(x_test)
```

4. 扩展:

稍加改变，例如对k个实例加权，可以用于回归。