**KNN**

距离：

我们最熟悉的距离公式就是欧拉距离：

还有一种距离计算公式是曼哈顿距离：

我们发现他们都可以写成最后一种形式，

所以我们将这种规律总结出来，写成距离的一般公式（明可夫斯基距离Minkowski distance）：

这样，继k和method以外，我们就获得了另外一个超参数p

我们可以检测在不同距离计算公式下KNN算法的表现如何

**Least Squared basic**

and are targets and data in our datasets.

So, we need to solve a and b to find the minimum loss function

**for b:**

differentiate loss function w.r.t **b**

**for a:**

differentiate loss function w.r.t **a**

given

also,

so,

**Linear Regression:**

Note 1

使用least squares regression 即使用SSR回归。是一种无偏估计

他要求被解释变量y必须服从正态分布。所以我们在拿到数据后，必须先进行正态化处理（如，取对数。。。）

Note 2

linear\_model.linearRegression()的参数：

fit\_intercept

默认为true，表示是否计算模型的截距

normalize

默认为false，表示是否要在回归前将矩阵X进行正态化（减去均值除以方差）。将数据

标准化是为了防止因为X数据单位不同等因素导致部分estimator过大或者过小。

比如： 中，，，

若想要等式成立，则需要是一个比较小的数字，而则需要比较大，这样我们回归出来

的estimator就没有太大的参考意义了

当fit\_intercept=False时忽略此参数。

copy\_x

默认为True，表示在进行线性回归是，是否在矩阵X的副本中进行运算，而不是直接将

原来的X覆盖掉。默认为True表示不在原数据中进行计算，这样可以防止原始数据被修

改。

n\_jobs

表示用多少核心来计算。默认为None（1）。如果输入-1则表示使用全部cpu核心来计

算

Note 3

R­­­­2为什么会是负值呢？

这一项是可以为负数的

Note 4

multicollinearity problem

逆矩阵存在的充分必要条件是矩阵不存在多重共线性（multicollinearity）

逆矩阵存在的条件， ， ，因为他是分母，分母不能为零

经过高斯变换后将原矩阵变成upper triangle matrix.

Then,

**Full Rank**:(determinant is not zero)

If all diagonal terms in a matrix is not 0, it is called “full rank matrix”.

Recall, OLS requires full rank assumption

Thus, a full rank matrix will not have a row that each of its term equals to zero.

This requires that each row should not have a linear relationship.

For example, if the 1st row and the 3rd row are linear related, each term in the 3rd row can be written as , where b is a scalar.

Then, the 3rd row can be easily canceled out through a basic arithmetic transformation.

Then, we will have a “all zero” row.

Thus, the determinant of this matrix is zero. the matrix is not full rank.

Two phenomena for multicollinearity

1 Rows are absolutely linear related, all zero rows occur.

Determinant of the matrix is zero.

2 Rows are highly linear related, the diagonal term(s) are very

close to zero. Determinant of the matrix is close to 0.

For , if close to zero, will close to infinity. You will get an extremely large estimator.

We can also check multicollinearity through **variance inflation factor** (VIF) in python.

Note 5

Ridge regression and Lasso (two solutions of collinearity)

使用岭回归或者Lasso后，模型效果会下降一些，因为我们删除了一部分数据。

岭回归又称Tikhonov regularization，正则化。

OLS通过SSR求最小值进行回归，ridge regression则是在SSR后额外添加一个正则项：

Now, even if is not full rank, is full rank.

Problem solved.

但是如果出现以下情况怎么办呢：

SSR中某些行本来不为0，但是加上后反而为0了。。

解决方法：我们可以换一个的取值。那么永远full rank,可逆

Notice, if too large, it will affect your regression.

Lasso

Lasso无法解决 = 0 的问题。

Model evaluation

**1.MSE (Mean Squared Error)**

In econometrics, the general form of MSE:

In machine learning,

**2.RMSE (Root Mean Squared Error)**

**3.MAE (Mean Absolute Error)**

**4.R2 (R square)**

we can consider:

1. the numerator as the error generated by using our model
2. the denominator as the error generated by using mean for prediction