# 实验报告

## 1. 探究分布之间的距离影响

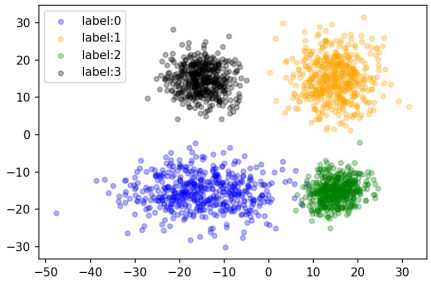
事先取定$Max_k = 10, fold_num = 15$, 数据取为每个$label$ 600个点,1/5划分为test集.

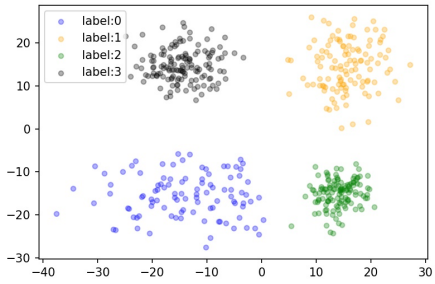
(1). 先采用:

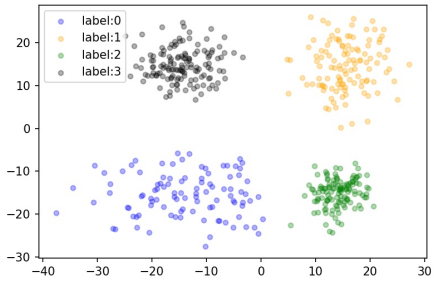
$$\Sigma_0 = \begin{bmatrix} 73 & 0 \\ 0 & 22 \end{bmatrix} \Sigma_1 = \begin{bmatrix} 21.2 & 0 \\ 0 & 32.1 \end{bmatrix} \Sigma_2 = \begin{bmatrix} 10 & 2 \\ 2 & 10 \end{bmatrix} \Sigma_3 = \begin{bmatrix} 15 & 0 \\ 0 & 15 \end{bmatrix}$$

$$\mu_0 = (-15, -15), \mu_1 = (15, 15), \mu_2 = (15, -15), \mu_3 = (-15, 15)$$

$train_data, test_data, predict$的分布分别如下图所示:




程序输出为:

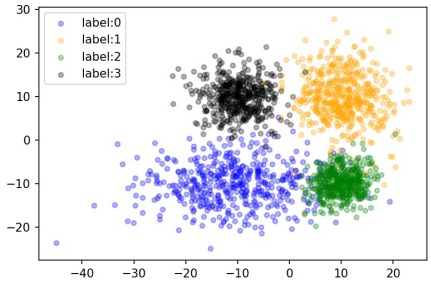
| k | manhattan_distance | euclidean_distance |
|---|---|---|
| 1 | 1.0 | 0.99 |
| 2 | 1.0 | 0.99 |
| 3 | 1.0 | 1.0 |
| 4 | 1.0 | 1.0 |
| 5 | 1.0 | 1.0 |
| 6 | 1.0 | 1.0 |
| 7 | 1.0 | 1.0 |
| 8 | 1.0 | 1.0 |
| 9 | 1.0 | 1.0 |

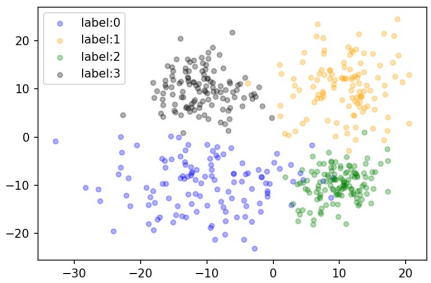
(2).再增加重叠度，将参数改为：

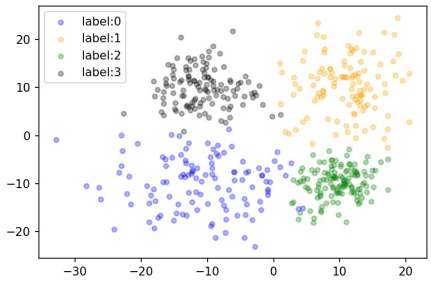
$$\Sigma_0 = \begin{bmatrix} 73 & 0 \\ 0 & 22 \end{bmatrix} \Sigma_1 = \begin{bmatrix} 21.2 & 0 \\ 0 & 32.1 \end{bmatrix} \Sigma_2 = \begin{bmatrix} 10 & 2 \\ 2 & 10 \end{bmatrix} \Sigma_3 = \begin{bmatrix} 15 & 0 \\ 0 & 15 \end{bmatrix}$$

$$\mu_0 = (-10, -10), \mu_1 = (10, 10), \mu_2 = (10, -10), \mu_3 = (-10, 10)$$

$train_data, test_data, predict$的分布分别如下图所示:



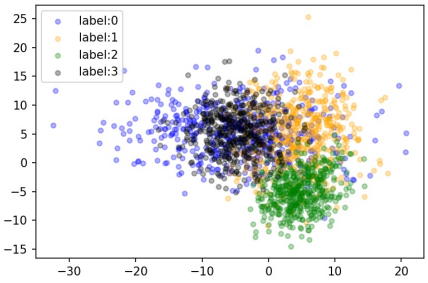

| k | manhattan_distance | euclidean_distance |
|---|---|---|
| 1 | 0.95 | 0.95 |
| 2 | 0.95 | 0.95 |
| 3 | 0.96 | 0.96 |
| 4 | 0.96 | 0.96 |
| 5 | 0.96 | 0.96 |
| 6 | 0.96 | 0.96 |
| 7 | 0.96 | 0.96 |
| 8 | 0.97 | 0.97 |
| 9 | 0.96 | 0.96 |

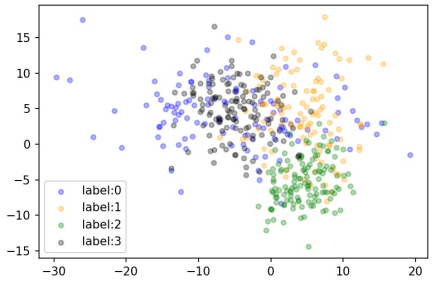
(3).将参数改为：

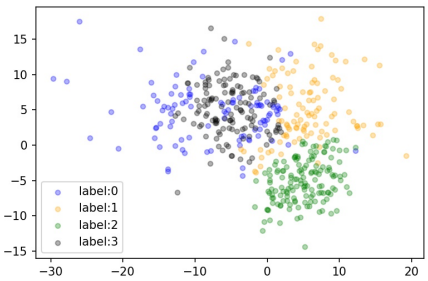
$$\Sigma_0 = \begin{bmatrix} 73 & 0 \\ 0 & 22 \end{bmatrix} \Sigma_1 = \begin{bmatrix} 21.2 & 0 \\ 0 & 32.1 \end{bmatrix} \Sigma_2 = \begin{bmatrix} 10 & 2 \\ 2 & 10 \end{bmatrix} \Sigma_3 = \begin{bmatrix} 15 & 0 \\ 0 & 15 \end{bmatrix}$$

$$\mu_0 = (-5, -5), \mu_1 = (5, 5), \mu_2 = (5, -5), \mu_3 = (-5, 5)$$

$train_data, test_data, predict$的分布分别如下图所示:




| k | manhattan_distance | euclidean_distance |
|---|---|---|
| 1 | 0.60 | 0.60 |
| 2 | 0.60 | 0.60 |
| 3 | 0.61 | 0.61 |
| 4 | 0.61 | 0.62 |
| 5 | 0.62 | 0.63 |
| 6 | 0.64 | 0.63 |
| 7 | 0.64 | 0.64 |
| 8 | 0.65 | 0.64 |
| 9 | 0.64 | 0.65 |

这三组数据的结果对比，可以得到以下结论：

- 随着用来产生数据的$\mu$值越来越接近，也即数据重叠越来越高，KNN算法的准确率显著下降
- Manhattan distance与Euclidean distance之间最终效果差别不大，并且Manhattan distance有时会outperform Euclidean distance
- 同一数据集下，随$k$的增加，准确率有提高的趋势

---

## 2. 探究协方差的影响

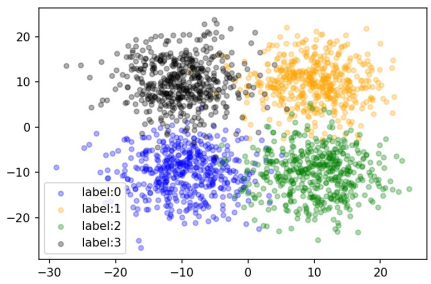
与前类似，事先取定$Max_k = 10, fold_num = 15$, 数据取为每个$label$ 600个点,1/5划分为test集。 为形成对照，我们取定第一组实验中第二次数据的$\mu$值，再改变协方差矩阵来探究。

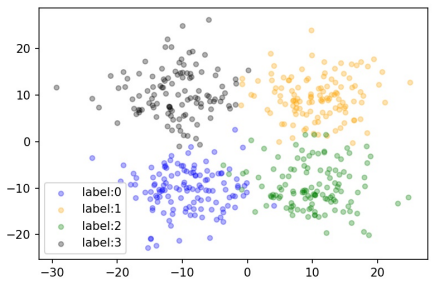
(1).参数改为：

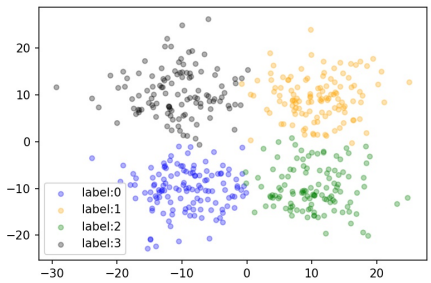
$$\Sigma_0 = \begin{bmatrix} 25 & 0 \\ 0 & 25 \end{bmatrix} \Sigma_1 = \begin{bmatrix} 25 & 0 \\ 0 & 25 \end{bmatrix} \Sigma_2 = \begin{bmatrix} 25 & 0 \\ 0 & 25 \end{bmatrix} \Sigma_3 = \begin{bmatrix} 25 & 0 \\ 0 & 25 \end{bmatrix}$$

$$\mu_0 = (-10, -10), \mu_1 = (10, 10), \mu_2 = (10, -10), \mu_3 = (-10, 10)$$

$train_data, test_data, predict$的分布分别如下图所示:



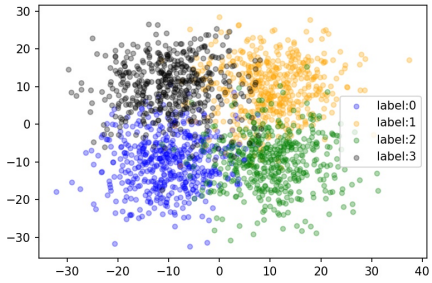

| k | manhattan_distance | euclidean_distance |
|---|---|---|
| 1 | 0.92 | 0.92 |
| 2 | 0.92 | 0.92 |
| 3 | 0.94 | 0.94 |
| 4 | 0.94 | 0.94 |
| 5 | 0.94 | 0.94 |
| 6 | 0.94 | 0.94 |
| 7 | 0.95 | 0.95 |
| 8 | 0.95 | 0.95 |
| 9 | 0.95 | 0.95 |

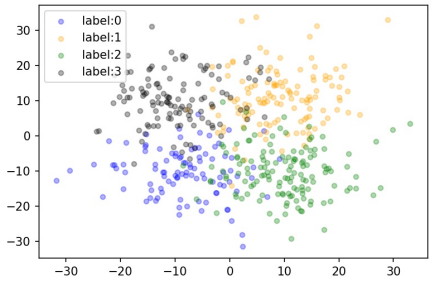
(2).参数改为:

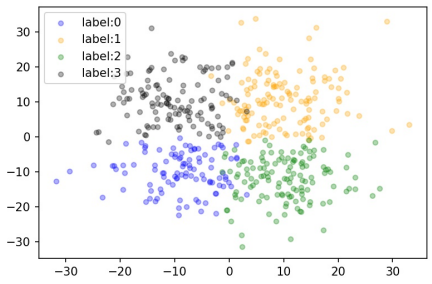
$$\Sigma_0 = \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix} \Sigma_1 = \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix} \Sigma_2 = \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix} \Sigma_3 = \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix}$$

$$\mu_0 = (-10, -10), \mu_1 = (10, 10), \mu_2 = (10, -10), \mu_3 = (-10, 10)$$

$train_data, test_data, predict$的分布分别如下图所示:



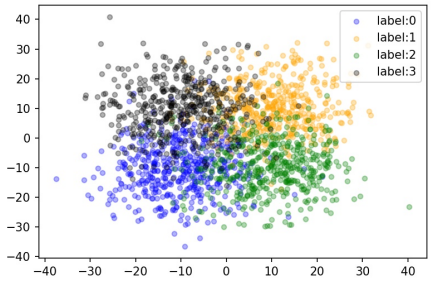

| k | manhattan_distance | euclidean_distance |
|---|---|---|
| 1 | 0.79 | 0.79 |
| 2 | 0.79 | 0.79 |
| 3 | 0.82 | 0.82 |
| 4 | 0.83 | 0.83 |
| 5 | 0.83 | 0.83 |
| 6 | 0.83 | 0.83 |
| 7 | 0.83 | 0.83 |
| 8 | 0.84 | 0.83 |
| 9 | 0.83 | 0.83 |

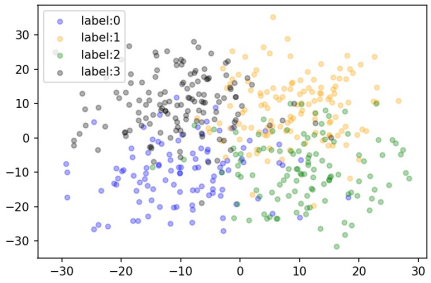
(3).参数改为：

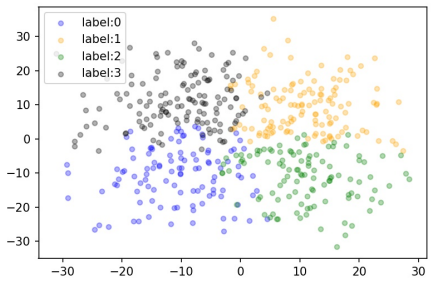
$$\Sigma_0 = \begin{bmatrix} 75 & 0 \\ 0 & 75 \end{bmatrix} \Sigma_1 = \begin{bmatrix} 75 & 0 \\ 0 & 75 \end{bmatrix} \Sigma_2 = \begin{bmatrix} 75 & 0 \\ 0 & 75 \end{bmatrix} \Sigma_3 = \begin{bmatrix} 75 & 0 \\ 0 & 75 \end{bmatrix}$$

$$\mu_0 = (-10, -10), \mu_1 = (10, 10), \mu_2 = (10, -10), \mu_3 = (-10, 10)$$

$train_data, test_data, predict$的分布分别如下图所示:




| k | manhattan_distance | euclidean_distance |
|---|---|---|
| 1 | 0.70 | 0.69 |
| 2 | 0.70 | 0.69 |
| 3 | 0.72 | 0.72 |
| 4 | 0.73 | 0.73 |
| 5 | 0.74 | 0.74 |
| 6 | 0.75 | 0.75 |
| 7 | 0.76 | 0.76 |
| 8 | 0.76 | 0.75 |
| 9 | 0.76 | 0.76 |

这三组数据的结果对比，可得如下结论：

- 随着协方差矩阵对角元的数越来越大，即数据之间相关性越高，KNN算法的准确率逐渐下降
- 与前面利用均值增加数据重叠度相比，这里提高了协方差，因此数据点之间虽然重叠，但离散度也增大，因此准确率没有下降到之前0.6左右

---

## 3.探究数据量的影响

这一步中，为使实验效果明显，我们取定参数为2中(2)的参数，再改变数据量大小，与前类似，事先取定$Max_k = 10, fold_num = 15$, 数据取为每个$label$ 取一定点数,1/5划分为test集。

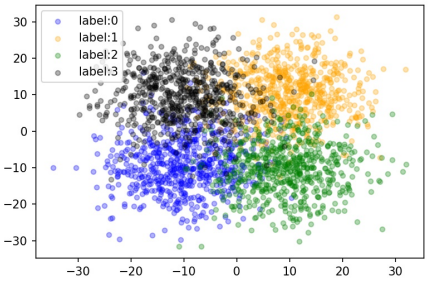
数据参数定为：

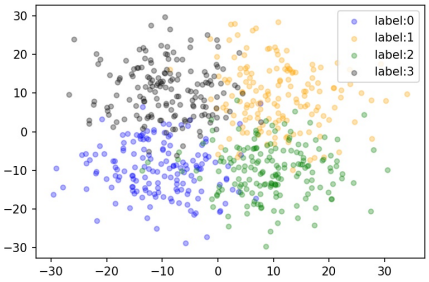
$$\Sigma_0 = \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix} \Sigma_1 = \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix} \Sigma_2 = \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix} \Sigma_3 = \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix}$$

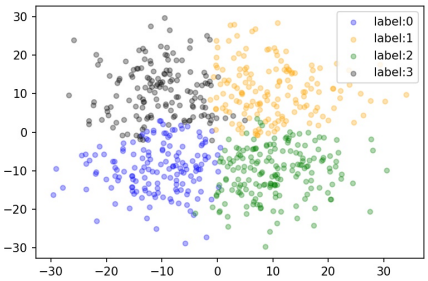
$$\mu_0 = (-10, -10), \mu_1 = (10, 10), \mu_2 = (10, -10), \mu_3 = (-10, 10)$$

(1).每个$label$取800个点

$train_data, test_data, predict$的分布分别如下图所示:



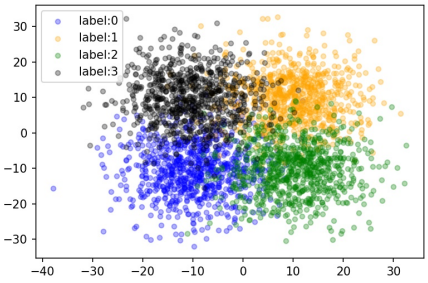



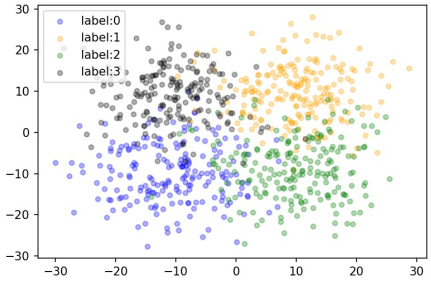



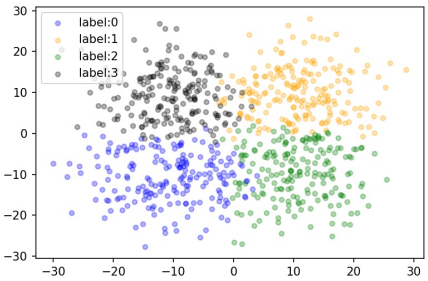

| k | manhattan_distance | euclidean_distance |
|---|---|---|
| 1 | 0.77 | 0.77 |
| 2 | 0.77 | 0.77 |
| 3 | 0.80 | 0.81 |
| 4 | 0.80 | 0.81 |
| 5 | 0.81 | 0.82 |
| 6 | 0.81 | 0.81 |
| 7 | 0.82 | 0.82 |
| 8 | 0.82 | 0.82 |
| 9 | 0.83 | 0.83 |

(2).每个$label$取1000个点

$train_data, test_data, predict$的分布分别如下图所示:




| k | manhattan_distance | euclidean_distance |
|---|---|---|
| 1 | 0.78 | 0.77 |
| 2 | 0.78 | 0.77 |
| 3 | 0.82 | 0.82 |
| 4 | 0.82 | 0.82 |
| 5 | 0.83 | 0.83 |
| 6 | 0.83 | 0.83 |
| 7 | 0.83 | 0.83 |
| 8 | 0.83 | 0.83 |
| 9 | 0.84 | 0.84 |

上面两组实验与2.（2）中的实验对照，有如下结论：

- 增加数据量虽然增大了数据的重叠度，但最终KNN算法的准确性并未受到较大影响
- 数据量较大时，采取较高的k值，反倒可能取得较高的准确率

---

# 4.实验总结

1. 数据分布和数据离散度会显著影响KNN算法的准确率，数据集的重叠度越高，KNN算法准确率越低
2. 提高k值能够提高准确率，但也受数据分布影响，提升空间的大小基本与数据的分布无关
3. 数据量能够提高KNN算法的准确率，但影响不大且提升较小，增加数据量也需要更多计算时间
4. 以上所有实验中均存在可能的偶然误差，应单独多次实验取平均值，从而得到更可信的结论