NLP 南京大学
自然语言处理研究组
Natural Language Processing Group

南京大学
NANJING UNIVERSITY

有监督学习

黄书剑

- 有监督学习
  - 回归和分类
- 评估方法
- 回归模型
  - 线性回归
- 分类模型
  - 逻辑斯蒂回归
- k近邻

有监督学习

- Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. (监督学习/有指导学习/指导学习)
  - mapping input to output
  - with input-output pairs

- Input x / $\vec{x}$, Output y
- Input output pair (x, y)
- Examples (x_i, y_i)

https://en.wikipedia.org/wiki/Supervised_learning

回归问题（回顾）

- **输出是一个实数数值**

  - 如：波特兰房价问题

Living area (feet^2)	#bedrooms	Price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮

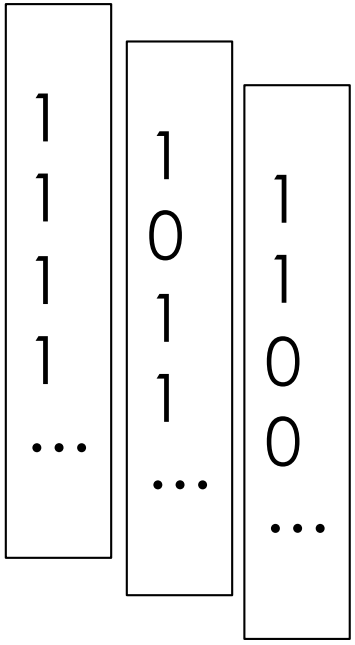
$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ \vdots \\ \vdots \\ x_d \end{bmatrix}$$

--- living size
--- lot size
--- # floors
--- condition
--- zip code
⋮

分类问题（回顾）

- **输出是一个离散值（表示类别）**
  - 输出是0和1（二类分类）
  - 如：西瓜分类问题

	是	否
表面光滑	○	
花纹清晰	○	
纹路明显	○	
底面发黄	○	
…		



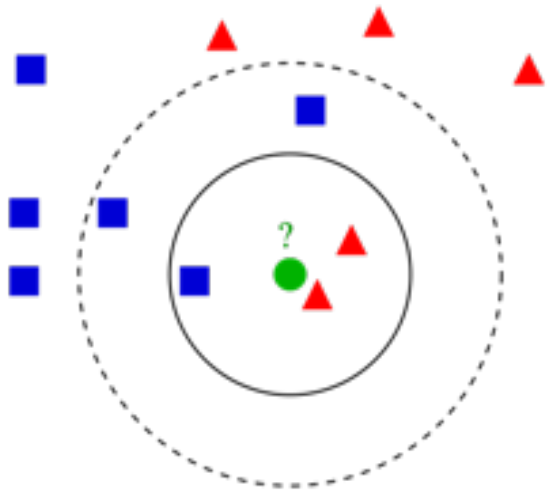
如何得到预测结果?

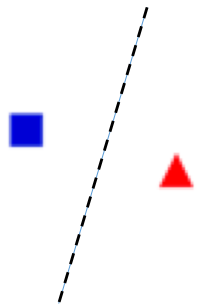
- **通过线性模型综合不同输入变量**
  - 回归模型可以完成输入到实数值(z)的映射
  - 分类模型可以在回归模型基础上再映射到类别

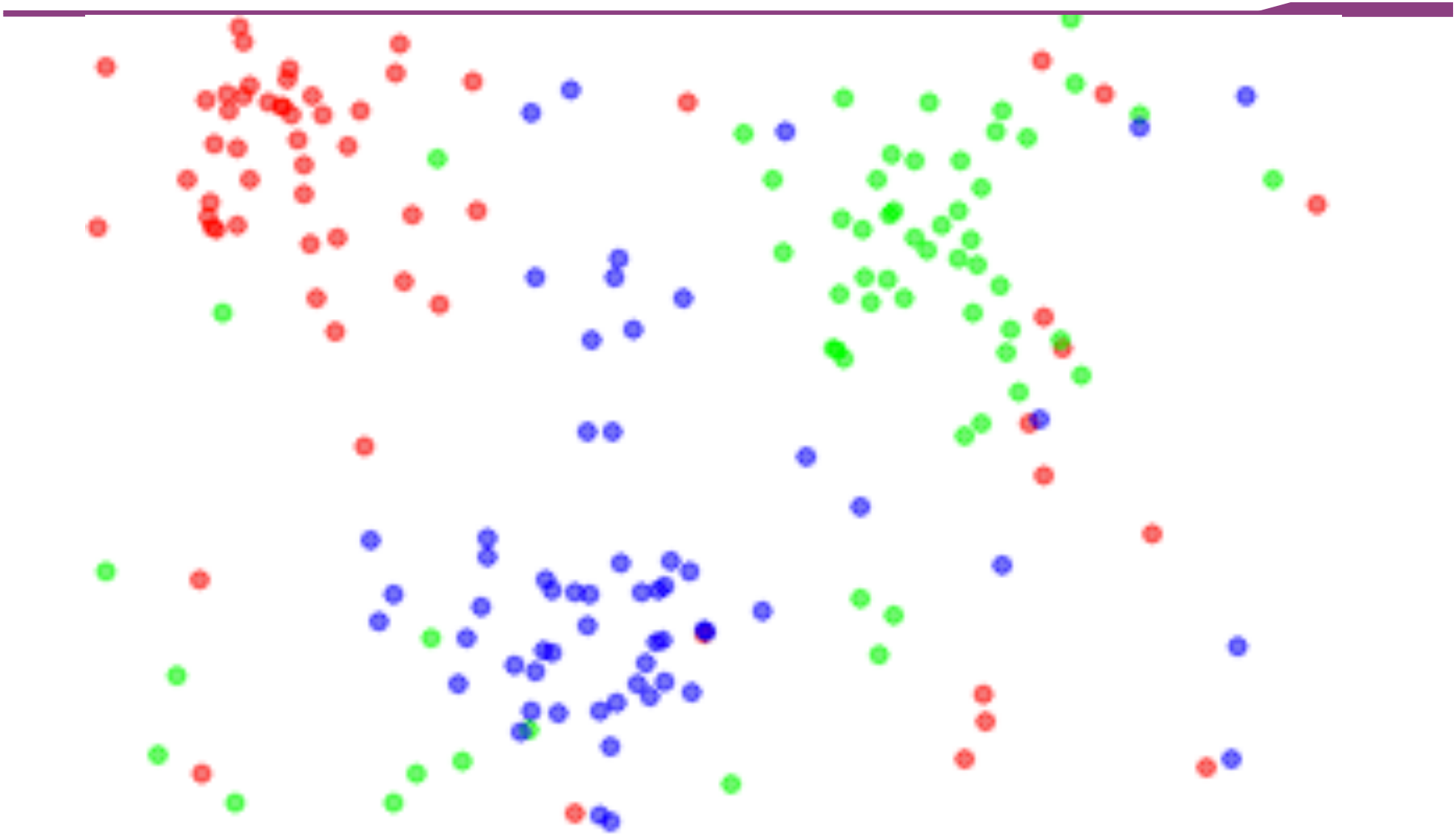
- **通过查询相似样本得到**
  - 相近样本的标记结果也相近（k-近邻）

https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

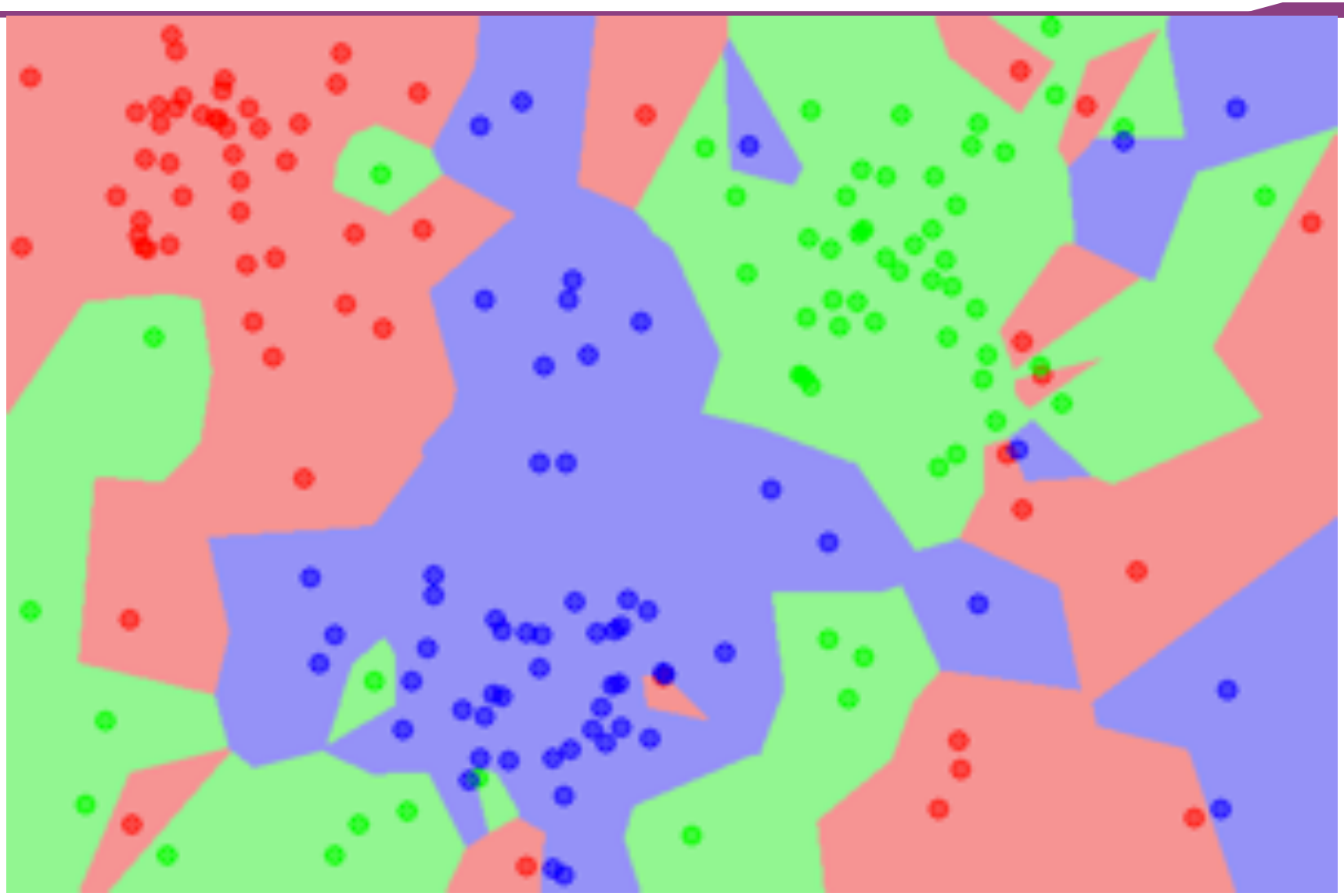
- k近邻（k-Nearest Neighbors，k-NN）
  - 查询与待预测样本最相近的k个训练样本
  - 使用k个样本的投票决策决定分类结果
  - 使用k个样本的均值决定回归结果



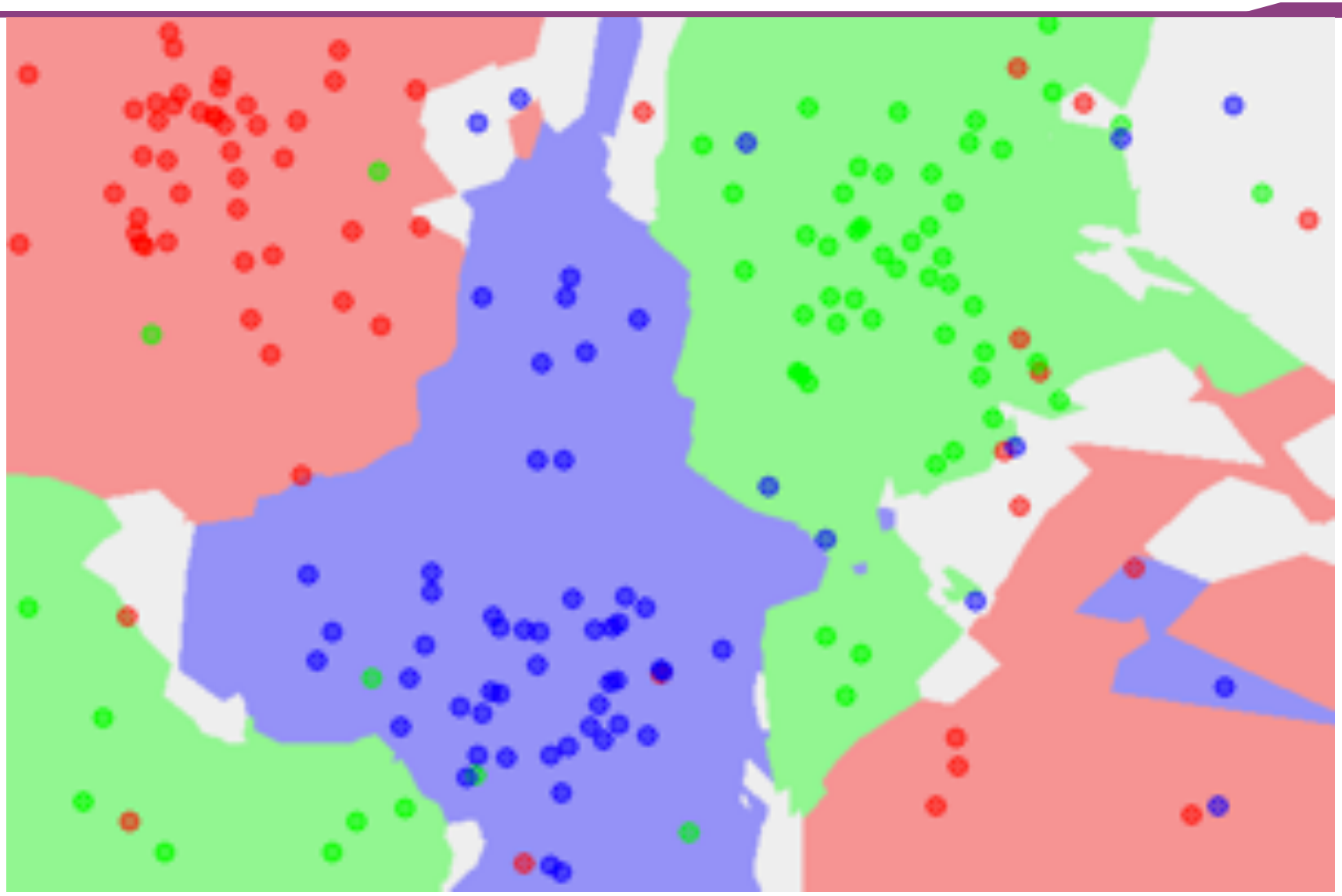




训练数据分布



1-NN的决策情况



5-NN的决策情况

k的选择

- k较小时
  - 决策更为接近数据本身的分布
  - 较容易收到噪音和异常数据的影响
- k较大时
  - 决策边界更为平滑
  - 对数据变化的敏感程度下降
- k为整个训练集时?

关于k-NN的一些简单讨论

- k的选择
- 没有显式学习过程
  - Lazy Learning
  - 实际运行代价
- 相似判断

什么是相似?

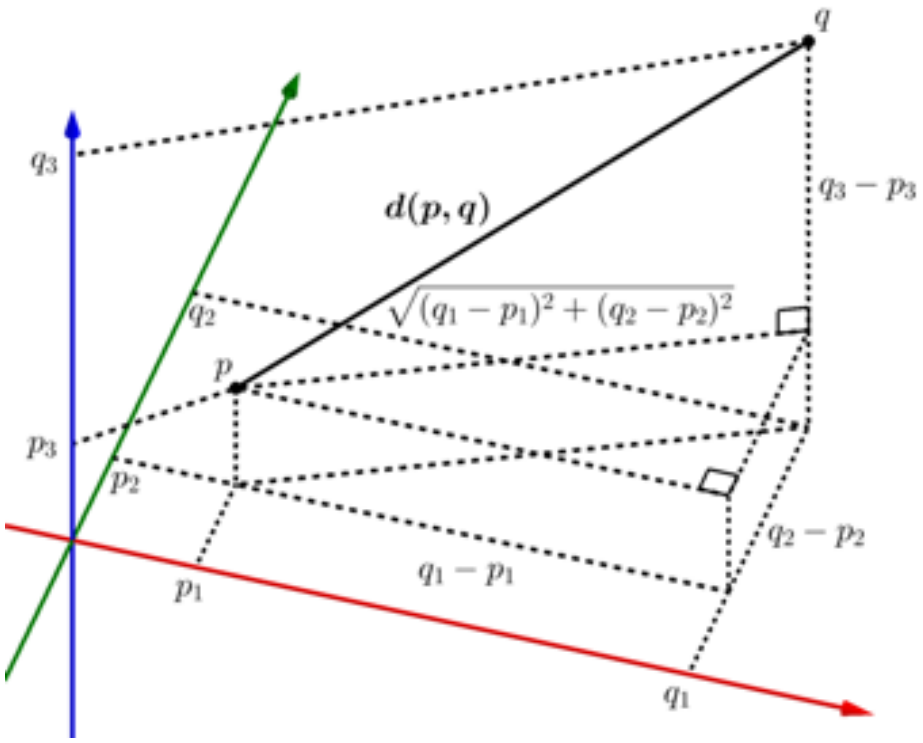
- Input Output Pair (x, y)
  - 样本的相似性表现为输入的相似性

- 对于描述性的特征（类别型categorical：nominal v.s. ordinal）
  - 颜色：红色、黄色、蓝色
  - 大小：很大、一般、较小
  - 纹理：清晰、正常、模糊
- 对于数值型特征
  - 重量、长度

距离度量

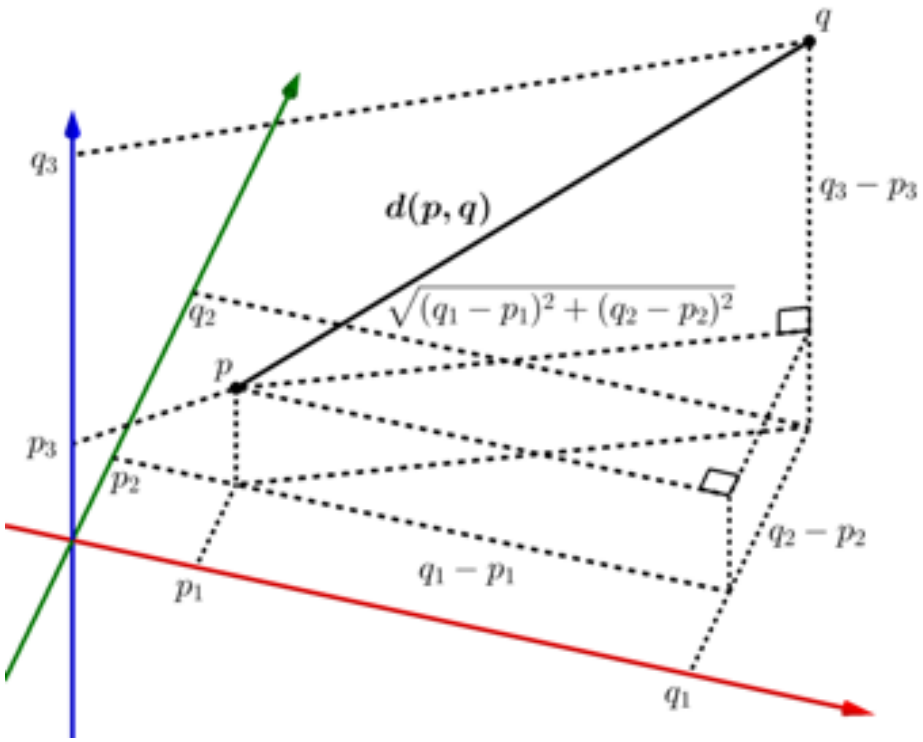
- 欧氏距离（Euclidean distance）

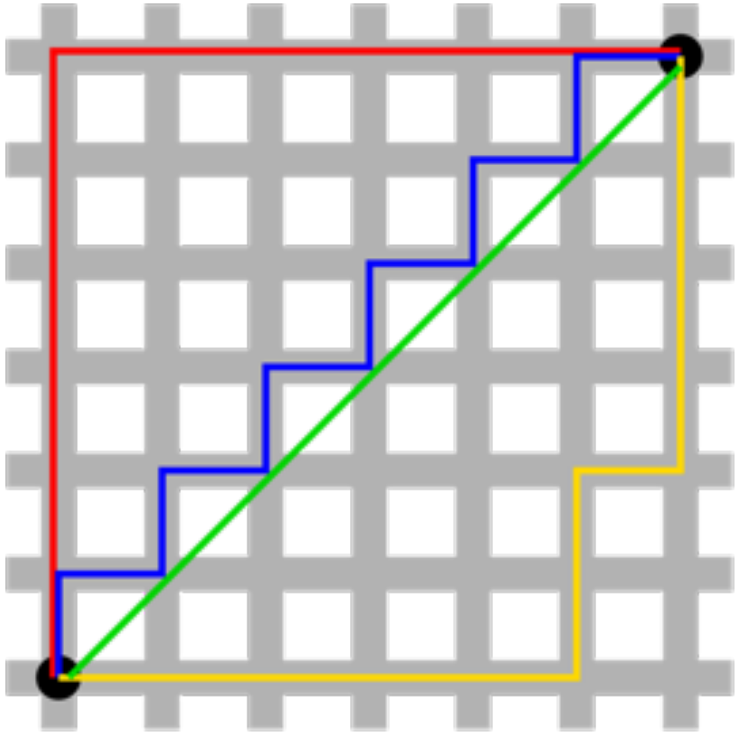
$$d(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{i=1}^{n} (x_i - x_i')^2}$$



- 曼哈顿距离（Manhattan Distance）

$$d(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{n} |x_i - x'_i|$$





- 闵可夫斯基距离（Minkowski Distance）

$$L_p(\mathbf{x},\ \mathbf{x}') = \left(\sum_{i=1}^{n} \left| x_i - x_i' \right|^p \right)^{1/p}$$

  - p = 1 时为曼哈顿距离
  - p = 2 时为欧氏距离
  - p -> inf 时

$$\lim_{p \to \infty} \left(\sum_{i=1}^{n} \left| x_i - x_i' \right|^p \right)^{\frac{1}{p}} = \max_{i=1}^{n} \left| x_i - x_i' \right|$$

$$\lim_{p \to -\infty} \left(\sum_{i=1}^{n} \left| x_i - x_i' \right|^p \right)^{\frac{1}{p}} = \min_{i=1}^{n} \left| x_i - x_i' \right|$$

更复杂的距离度量

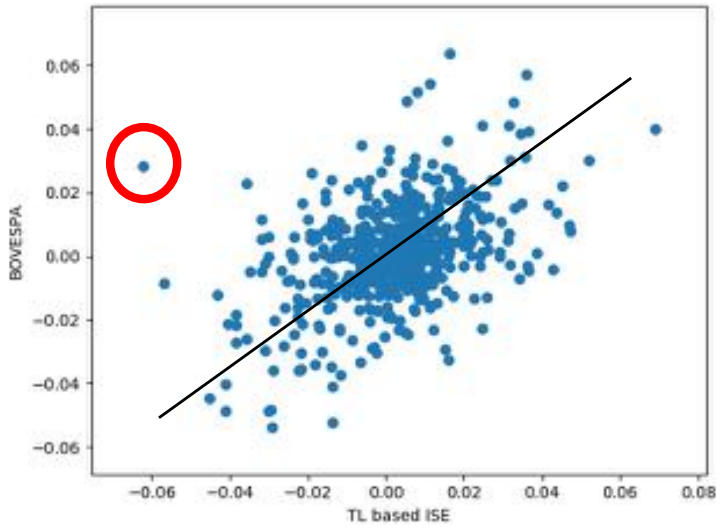
- 考虑到不同维度的重要性

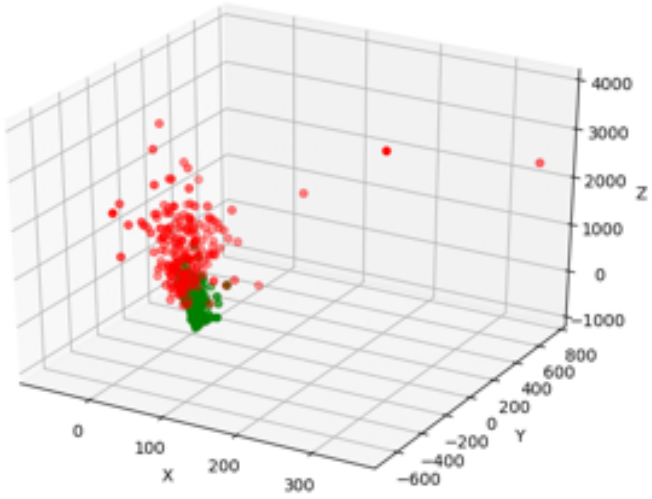
$$d(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{i=1}^{n} w_i \left(x_i - x_i'\right)^2}$$

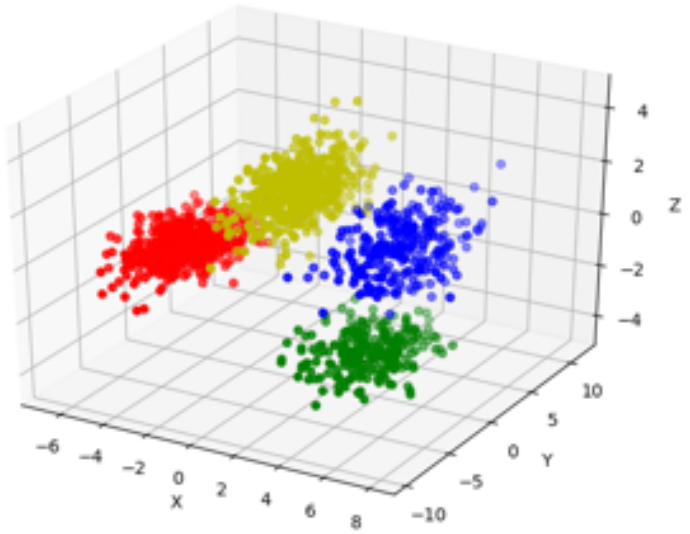
- 降维/维度约简
- 度量学习

有监督学习的应用

- 根据应用目标确定学习任务
  - 分类、回归
- 对数据进行处理
  - 数值转换、降维
- 观察数据，选择合适的模型
- 分析结果，对现有方案进行改进







练习三：

- 实现一个kNN算法用于分类或者回归
- 跟其他的算法的分类或者回归结果进行比较
- 尝试思考在你的问题中出现的分类错误、回归错误

参考资料

- https://en.wikipedia.org/wiki/Minkowski_distance
- https://en.wikipedia.org/wiki/Euclidean_distance
- https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm