Applied Machine Learning

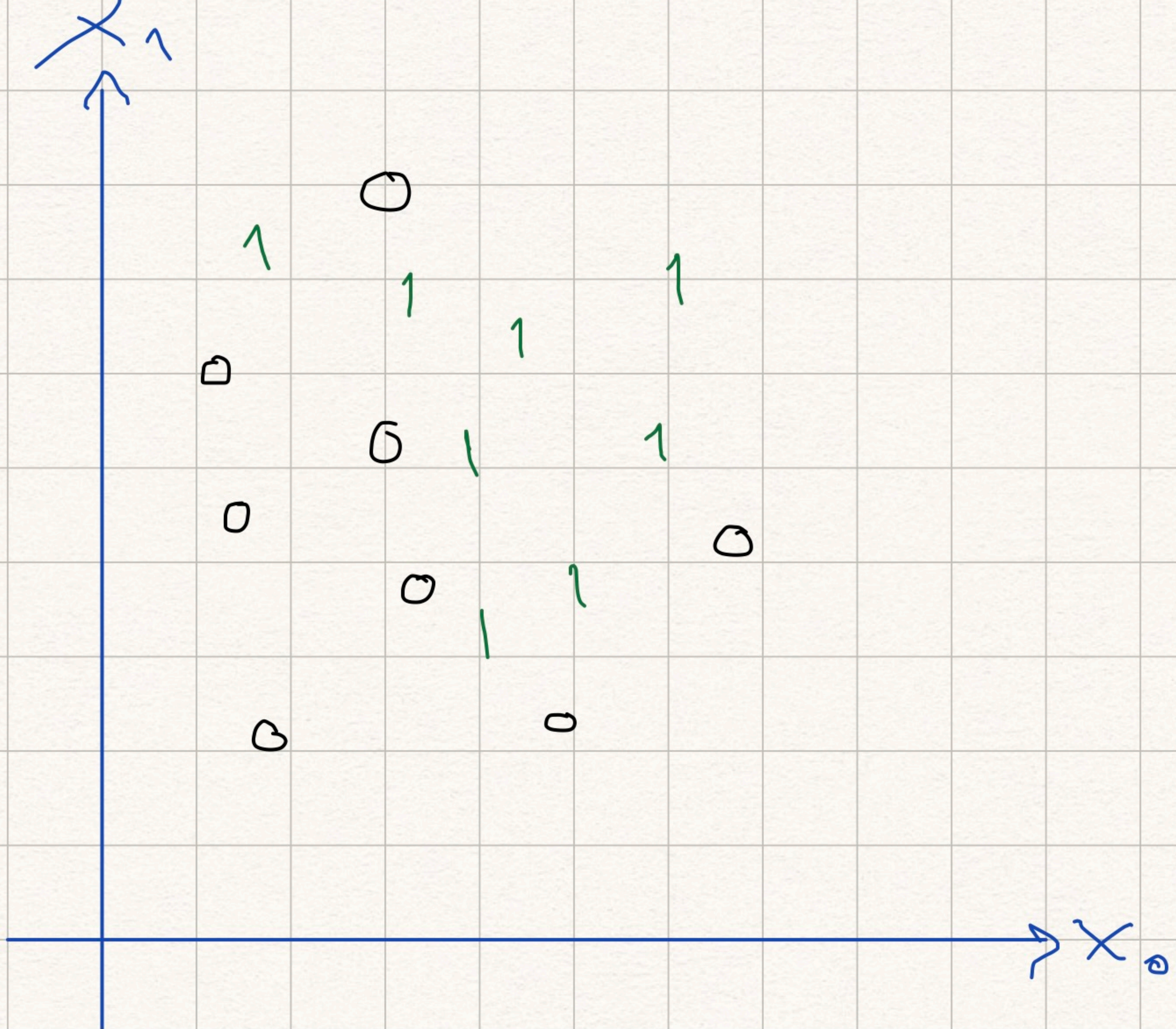
Classification - Nearest Neighbor Learning

Nearest Neighbor Learning

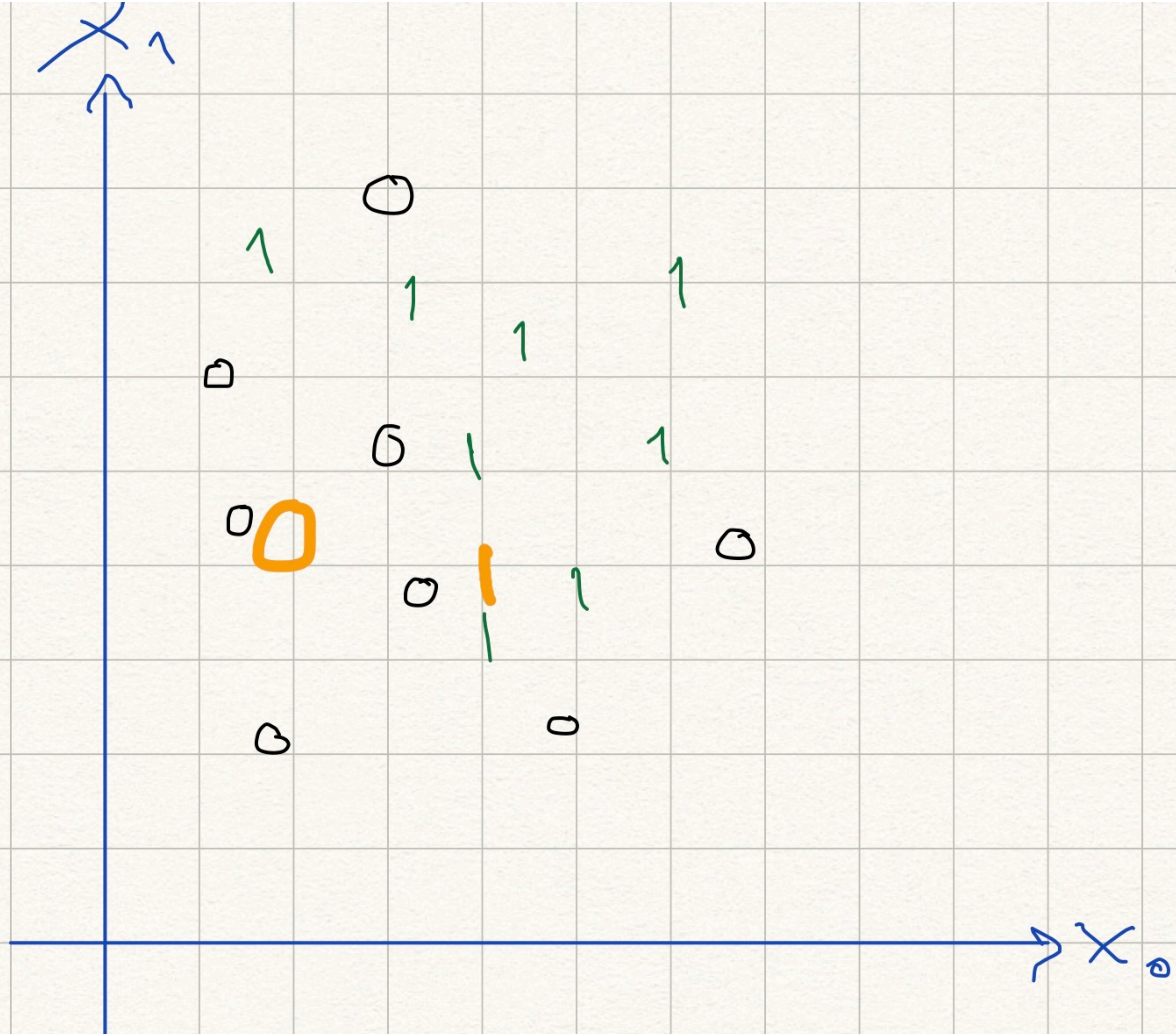
- Nearest Neighbor Learning
- k-Nearest Neighbor Learning
- How to address some issues of the NN algorithms

Nearest Neighbor Overview

- During training: store the Training Set of feature vectors $(\boldsymbol{x}, y)$
    - defer processing until a new data point must be classified
    - Instance-Based Learning
- During classification of item $\boldsymbol{x_c}$
    - compute distance between $\boldsymbol{x_c}$ and stored date
    - report label y of closest stored item

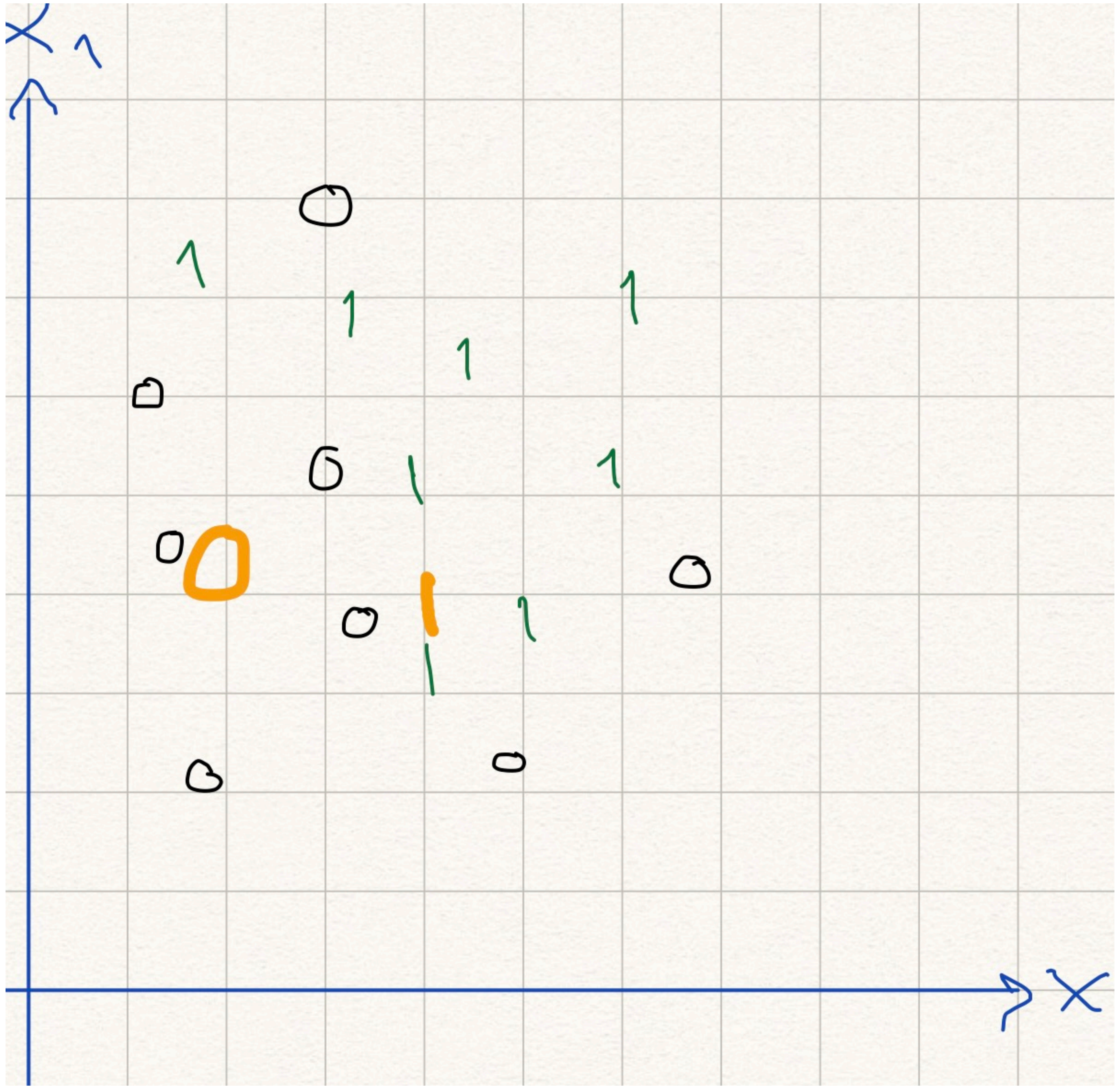


x_1
x_0
0
1


x_1
x_0
0
1
1
1
1
0
0
1
1
0
0
0
1
0
1
1
0
0

Nearest Neighbor Classification

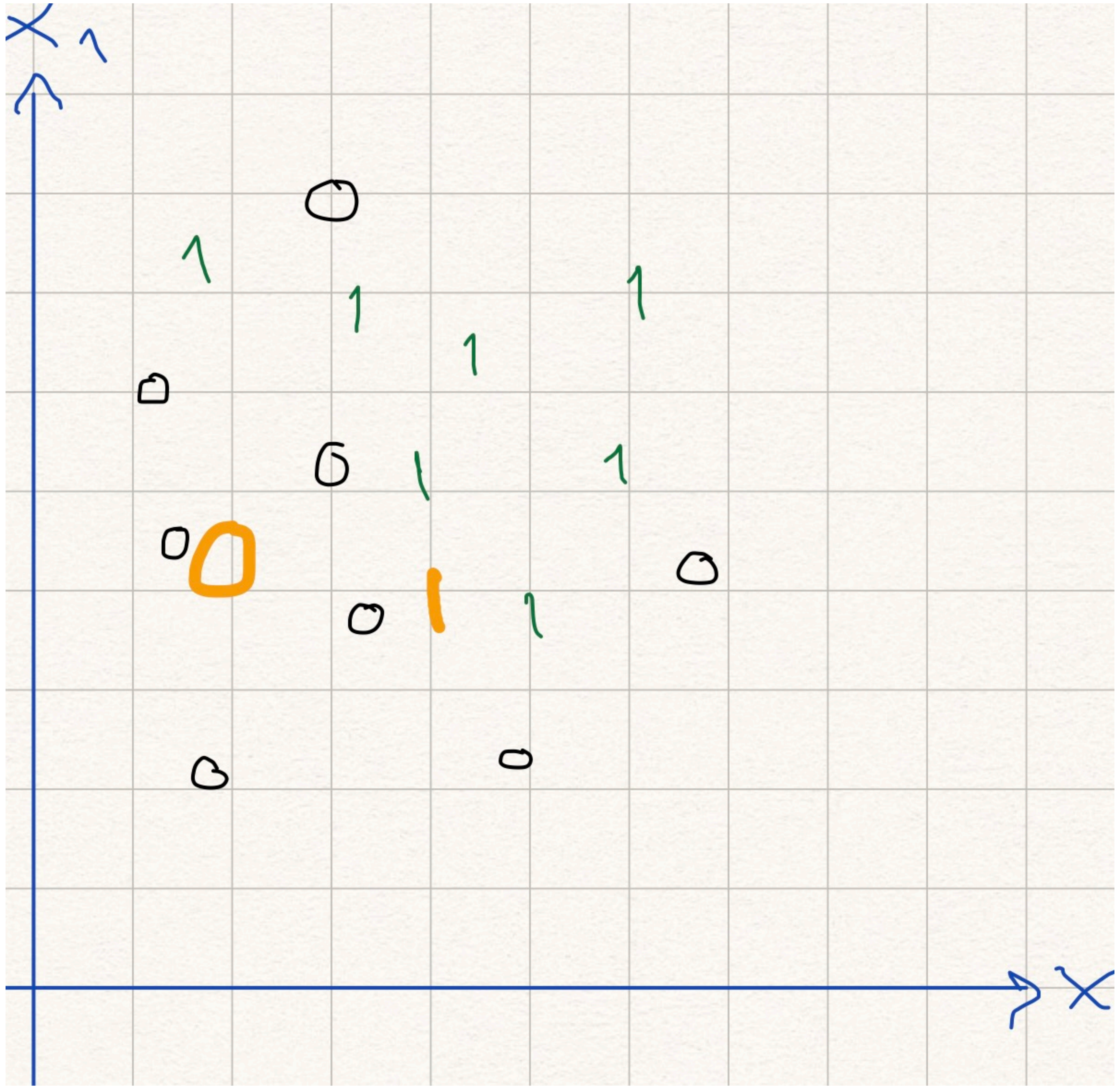
- Training:
  - Store N feature vectors of fixed size in dataset: pairs $(\boldsymbol{x}_i, y_i)$
- Classification
  - Distance metrics: $dist(\cdot, \cdot)$
  - Input: Query vector $\boldsymbol{x}$
  - Algorithm
    - Find the example vector $(\boldsymbol{x}_c, y_c)$ in dataset so that $dist(\boldsymbol{x}, \boldsymbol{x}_c)$ is the smallest
    - Report label $\boldsymbol{y_c}$



Nearest Neighbor Classification

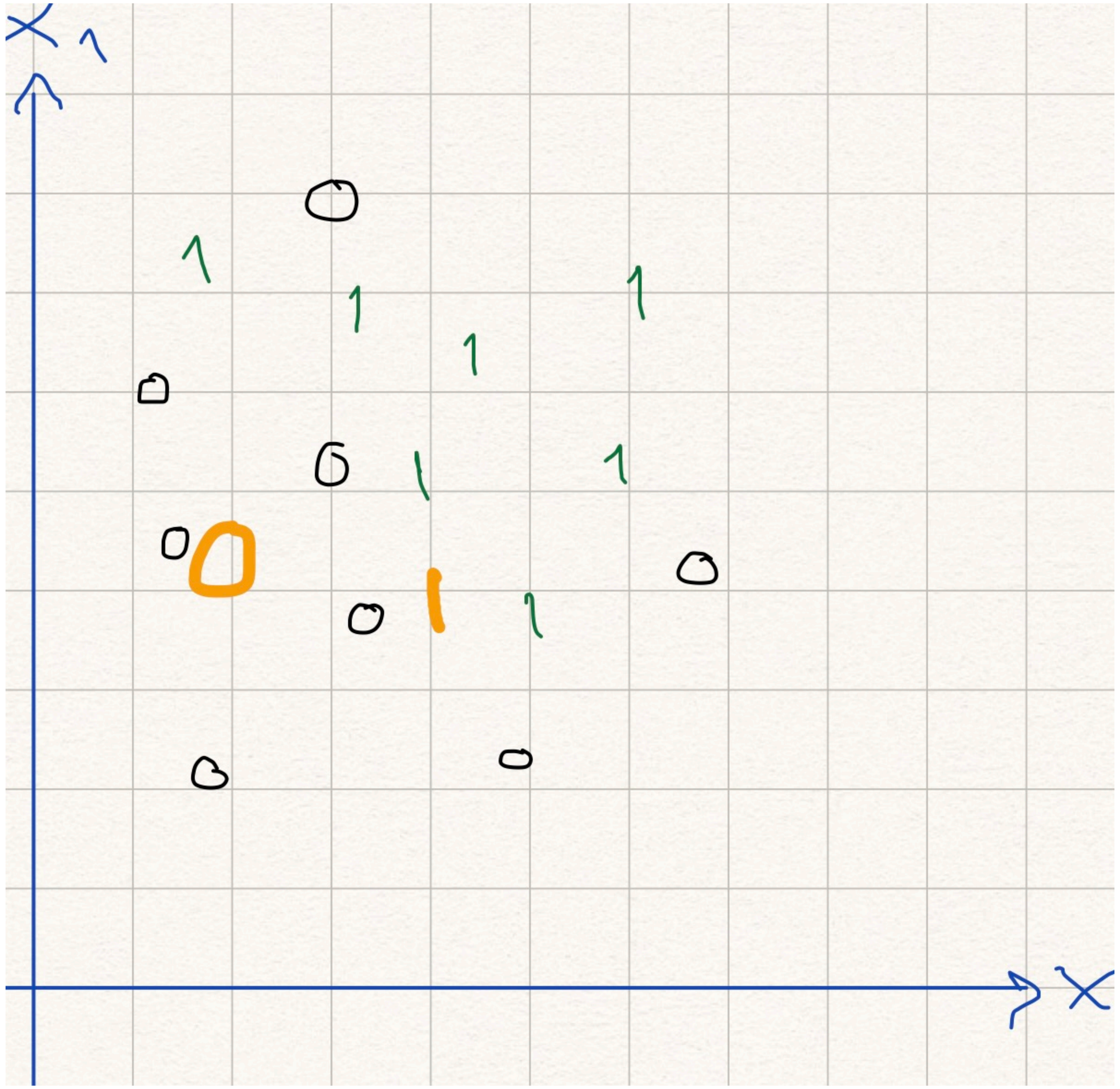
- If the dataset had one less item

- The classification changes



k-Nearest Neighbor Classification

- Training:
  - Store N feature vectors of fixed size in dataset: pairs $(\boldsymbol{x}_i, y_i)$
- Classification
  - Distance metrics: $dist(\cdot\,,\cdot\,)$
  - Input: Query vector $\boldsymbol{x}$
  - Algorithm
    - Find the set of k example vectors $k_{\text{closest}} = \{(\boldsymbol{x}_{c1}, y_{c1}), (\boldsymbol{x}_{c2}, y_{c2}), \ldots, (\boldsymbol{x}_{ck}, y_{ck})\}$ in dataset so that their distances to the query vector $dist(\boldsymbol{x}, \boldsymbol{x}_{ci})$ are the k smallest
    - Report label $\boldsymbol{y}_c \in k_{\text{closest}}$ that is repeated the most



Key Component: Distance Metrics

- Euclidean distance metrics
- Feature normalization
  - recenter the feature around 0
  - rescale so that the variance is 1

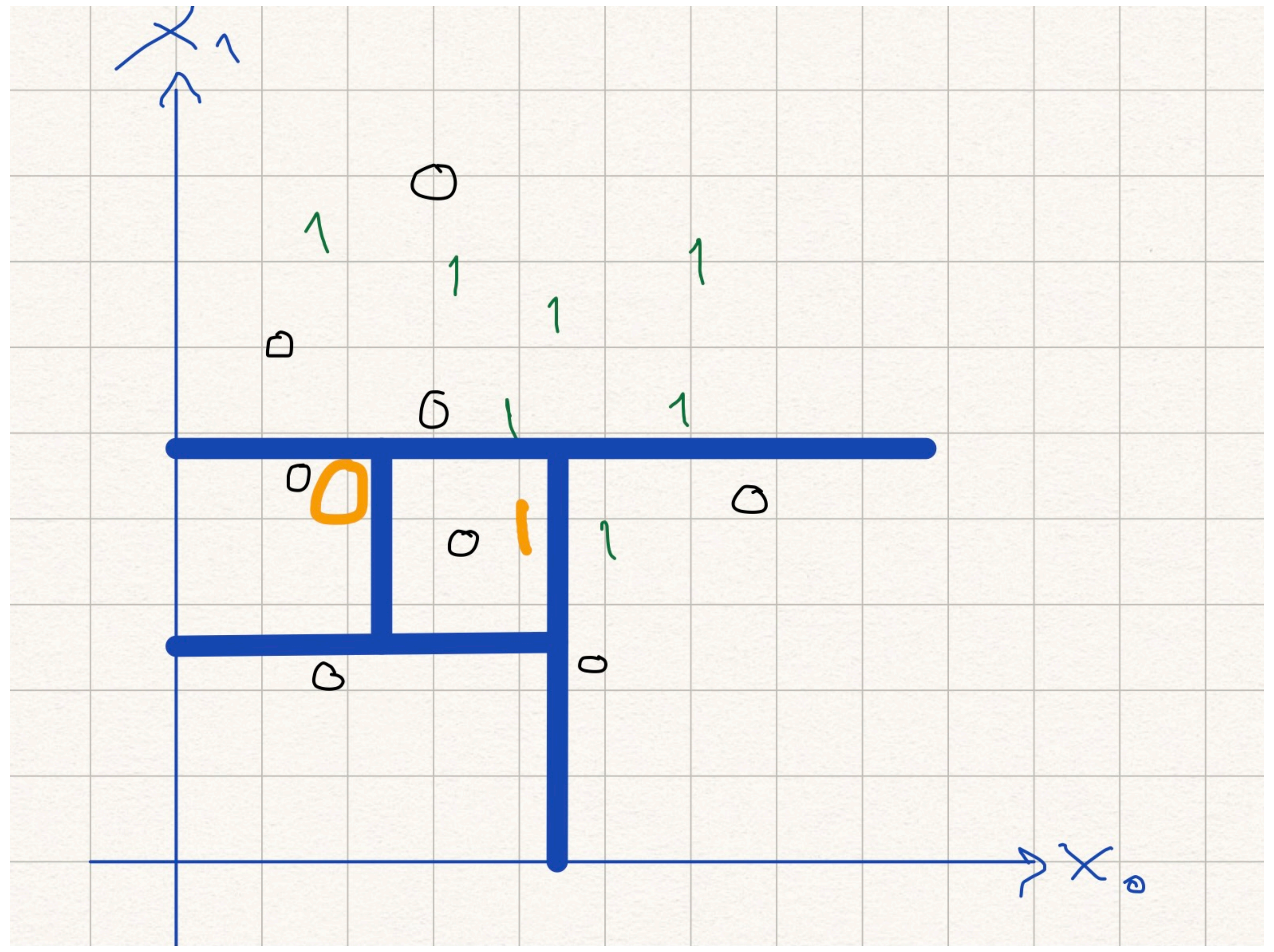
Key Component: Nearest Neighbors Computation

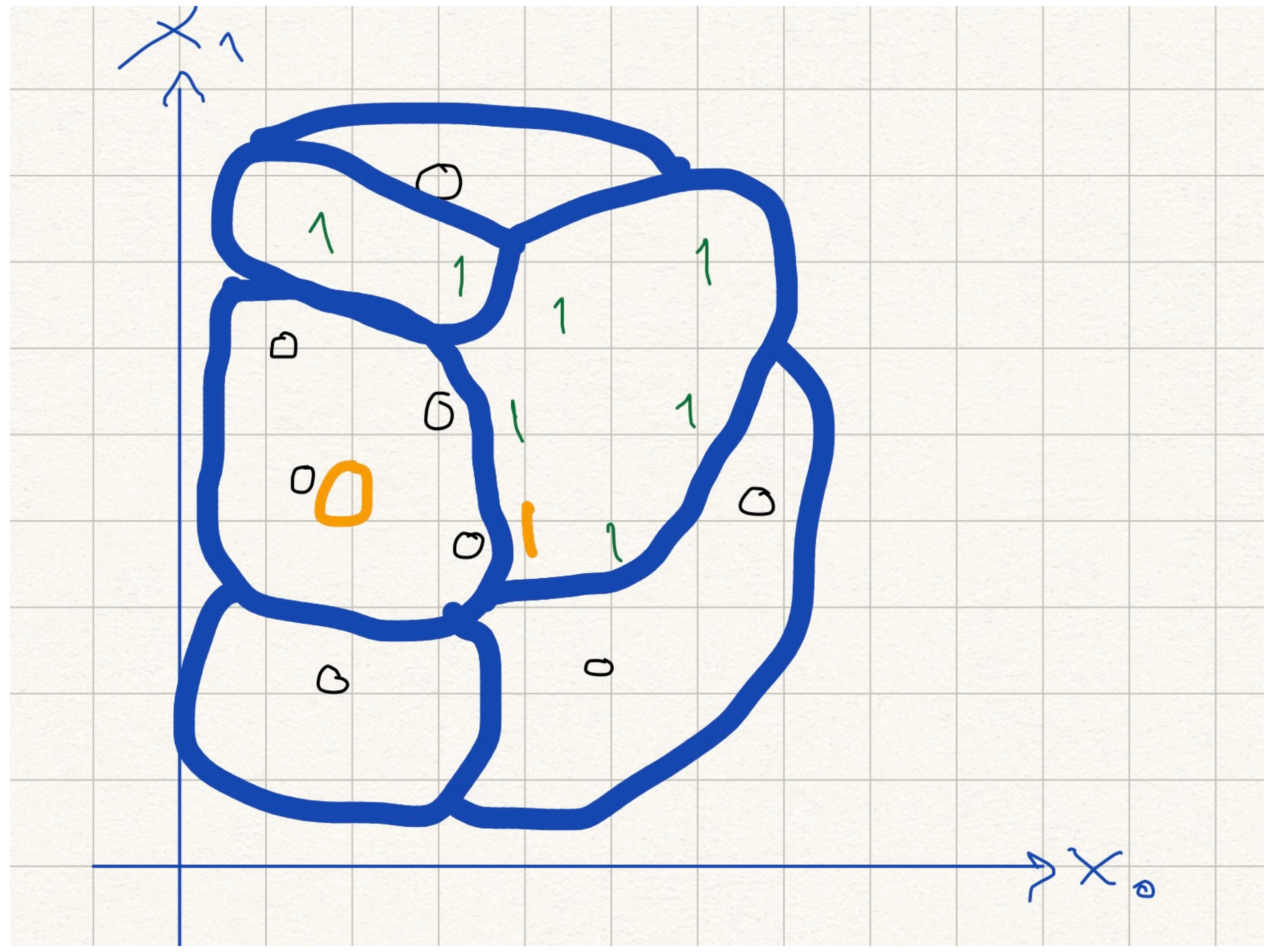
- Computing distance to the N members of the dataset is time consuming

- N may need to be large for the classifier to be effective

- Data points lie in high-dimensional spaces

- Some approaches to speed up nearest neighbors computation
  - kd-tree
  - locality-sensitive hashing





Nearest Neighbors

- Additional refinements
  - weight the contribution of the k neighbors according to their distance to the query
- Sensitive to potentially irrelevant features

Nearest Neighbor Learning

- Nearest Neighbor Learning
- k-Nearest Neighbor Learning
- How to address some issues of the NN algorithms

Applied Machine Learning

Classification - Nearest Neighbor Learning