Machine Learning
Lecture 2

Instructor

Anton Selitskiy

- **Office:** GOL-3701
- **Webpage:** `https://www.cs.rit.edu/~ams/`
- **Email:** amsvcs@rit.edu

- **Office Hours:**
  - Tuesday 3pm — 5pm
  - Thursday 10am — 12pm
  - Any time I'm in the office
  - By appointment

Similarity Principle and kNN

Left photo by David Iliff. License: CC BY-SA 3.0

Right photo `https://www.natesnursery.com/trees-of-the-high-desert-eldarica-pine/`

Similarity Principle and kNN

https://matterntrees.com/product/concolor-fir-cut

Similarity Principle and kNN

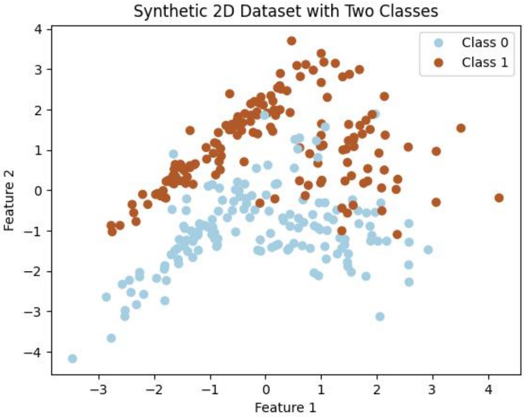
Fir: Branches upwards, trunk not visible, denser greenery, bright green color

Pine: Branches upwards, visible trunk, dull color

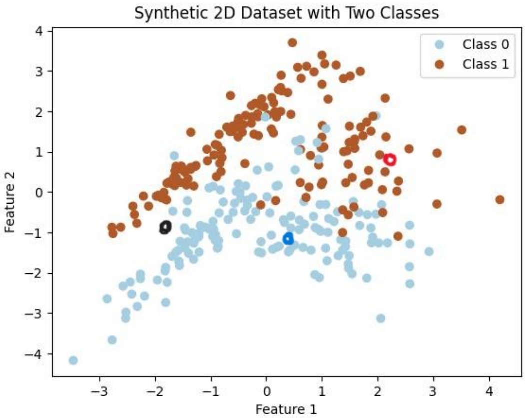
Similarity Principle and kNN

Branches upwards
trunk not visible
dense greenery
light color
$\Rightarrow$ fir

Classification Problem



Classification Problem and Similarity Principle



Classification Problem and Similarity Principle

We will assume that if the features are similar, then the objects are from the same class

k Nearest Neighbors (kNN): training

- We are given $\left(x^{(i)},\ y^{(i)}\right)$, $i = 1,\ 2, \ldots,\ N$.
- $\mathbb{Y} = \{1,\ 2,\ \ldots,\ K\}$ (classification problem).
- training $=$ memorizing of the given data.

k Nearest Neighbors (kNN): prediction

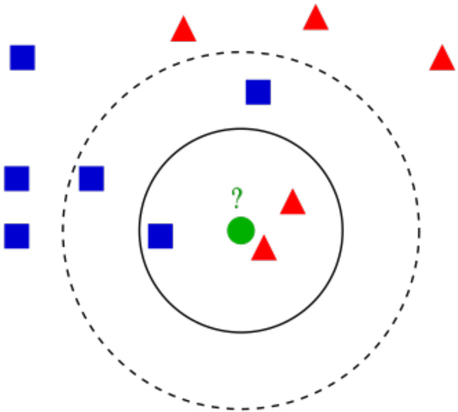
- We have a new feature x.
- Define the distance between new feature and $x^{(i)}$: $d\left(x,\ x^{(i)}\right)$.
- Rearrange the objects by the closeness to x :

$$d\left(x,\ x^{(i_1)}\right) \leq d\left(x,\ x^{(i_2)}\right) \leq \ldots \leq d\left(x,\ x^{(i_N)}\right).$$

- Look at the first k labels and assign the class with the highest number of representatives:

$$a(x) = \underset{c \in \mathbb{Y}}{argmax} \sum_{s=1}^{k} \left[x^{(i_s)} = c\right].$$

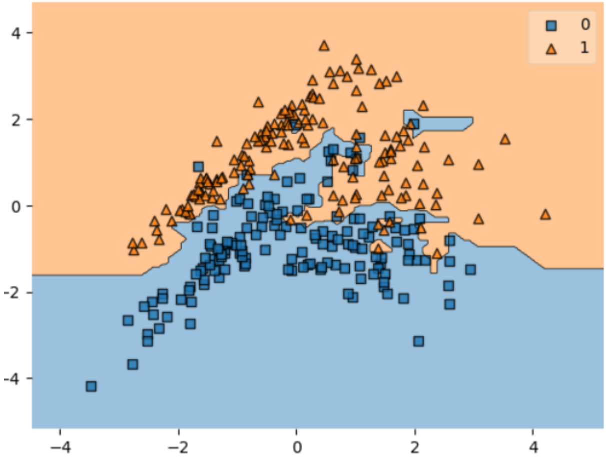
k Nearest Neighbors (kNN): prediction



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

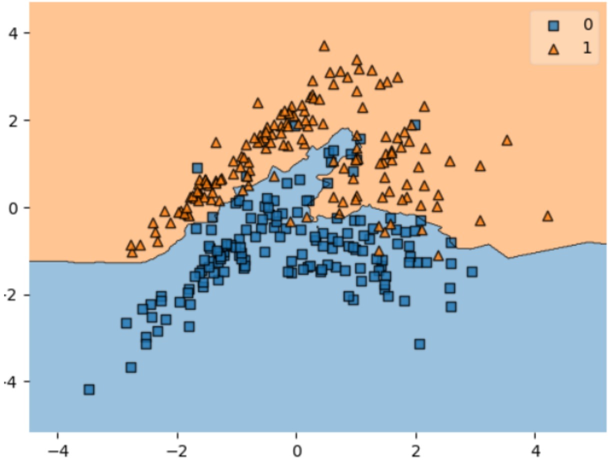
k Nearest Neighbors (kNN): prediction

$$k = 1,\ p = 1$$



k Nearest Neighbors (kNN): prediction

$$k = 5, \;\; p = 2$$

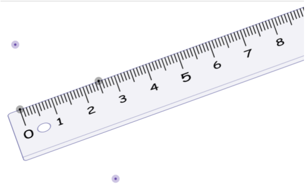


Example

How many trips during a day	Average expenses on taxi	How often premium was used	Age	Agreed to upgrade to premium
1	6	87837	787	yes
2	7	78	5415	no
...	...	...	...	...

How likely is the person to upgrade to the premium?

Euclidean Distance (or L_2)

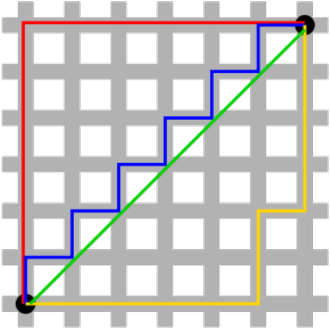


$$d\left(x^{(1)},\ x^{(2)}\right) = \sqrt{\sum_{j=1}^{d}\left(x_j^{(1)} - x_j^{(2)}\right)^2}$$

https://teacher.desmos.com/activitybuilder/custom/60527c839f5bc7445f2ac793?collections=5f2c6c2c3a7f6a21c8c5d7e8

Manhattan Distance (or L_1)

$$d\left(x^{(1)},\ x^{(2)}\right) = \sum_{j=1}^{d} \left| x_j^{(1)} - x_j^{(2)} \right|$$



L_p Distance

$$d\left(x^{(1)},\ x^{(2)}\right) = \sqrt[p]{\sum_{j=1}^{d} \left|x_j^{(1)} - x_j^{(2)}\right|^p}$$

Counting Distance (Hamming Distance)

Usually uses	City	Payment	Agreed to upgrade to premium
Premium	Buffalo	cash	yes
Econom	Rochester	card	no
...	...	...	...

$$d\left(x^{(1)},\ x^{(2)}\right) = \sum_{j=1}^{d} \left[x_j^{(1)} \neq x_j^{(2)}\right]$$

Model Comparison

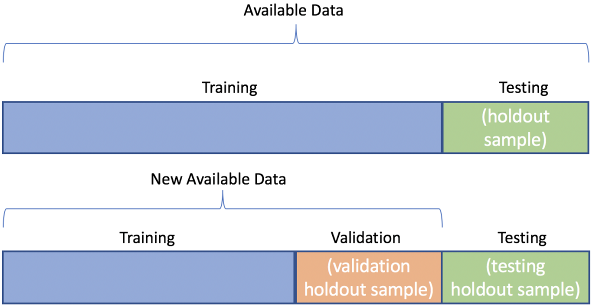
- How to compare models?
- How to choose k?

Accuracy

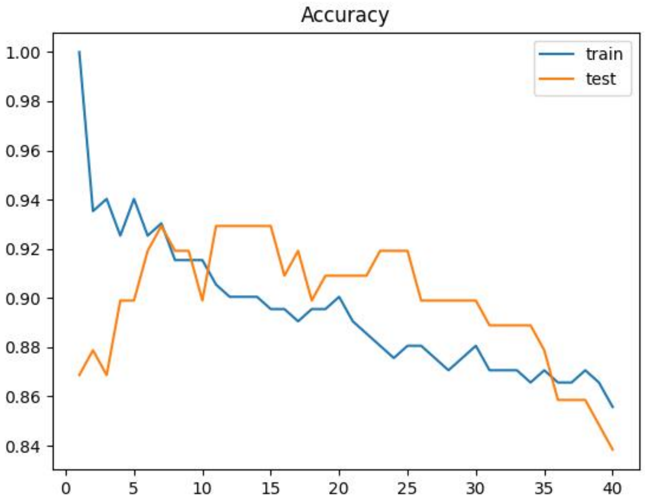
$$Accuracy = \frac{1}{N} \sum_{i=1}^{N} \left[a(x^{(i)}) = y^{(i)} \right]$$

Hyperparameters

k can not be tuned on the training data!



How to choose k?



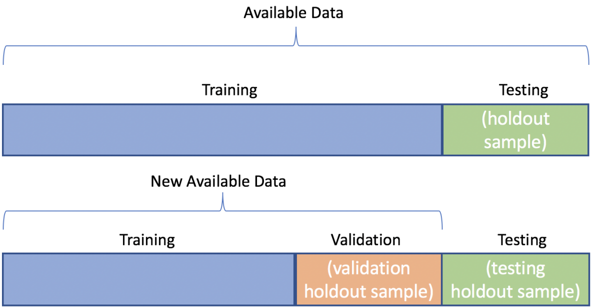
Overfitting and Generalization

Exam preparation:

Memorized lectures

Understand material

Typical Split



- Train set: build the model
- Validation set: tune hyperparameters
- Test set: evaluate quality of your model