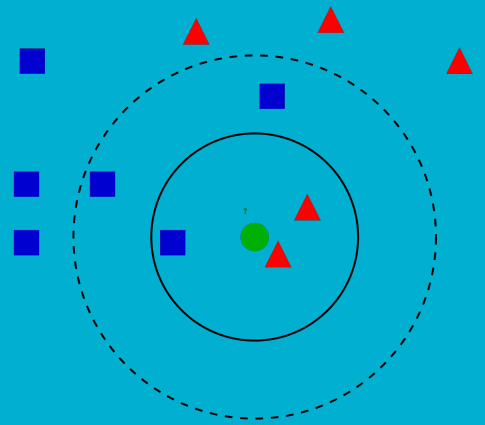
k Nearest-Neighbors (k-NN)



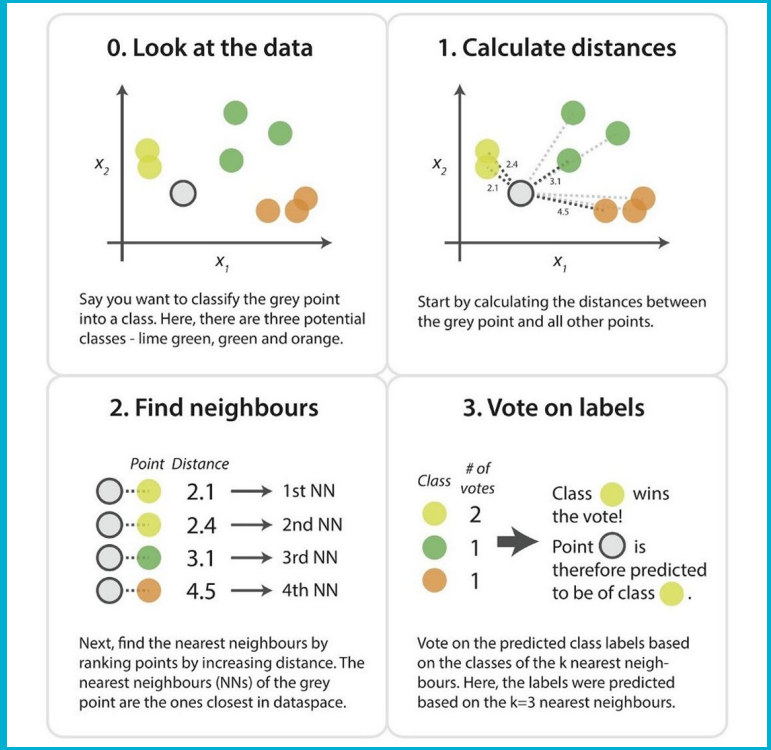
By Emily Atkinson, Vanessa Gleason, Anthony Rondos, and Eduard Stalmakov

Introduction to k-Nearest Neighbors (k-NN)

- Used for supervised learning
  - Classification - most common usage
  - Regression
- Works off of the assumption that similar points are spatially near each other
- Used for:
  - Simple recommendation systems
  - Pattern recognition
  - Data mining
  - Intrusion detection

How the algorithm works

- Finds the distance between a query and all other examples in the data
- Selects only the specified number examples closest to the query (K)
- If you're using it for:
  - Classification: Votes for the most frequent label and assigns it to that group
  - Regression: Averages the labels and assigns it that value



Classification Example: Where does sweet potato go?

- kNN classifier is to **classify unlabeled observations** (sweet potato) by assigning them to the class of **the most similar** labeled examples.
- The four nearest kinds of food are apple, green bean, lettuce, and corn.
- Because the vegetable wins the **most votes**, sweet potato is assigned to the class of **vegetable**.

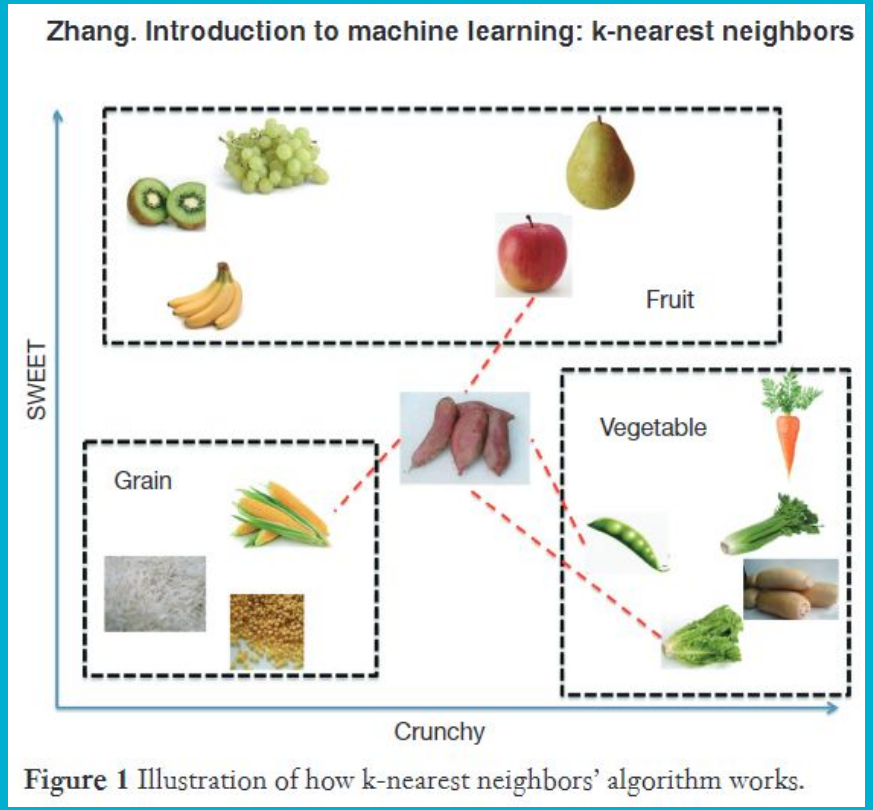
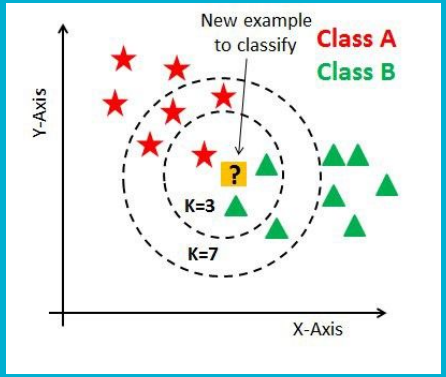


Figure 1 Illustration of how k-nearest neighbors' algorithm works.

https://atm.amegroups.com/article/view/10170/pdf

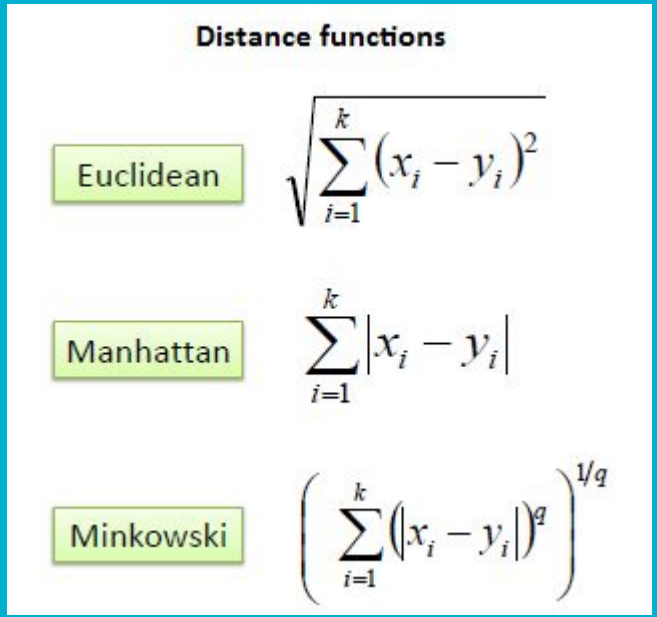
Tuning Hyperparameters

- k - this is the number of ‘neighbors’ the algorithm will view to classify the unknown value
- If k is too low, the algorithm will overfit the data
- If k is too high, the algorithm will underfit the data
- Best practice is k = √n, where n is the number of samples in the dataset



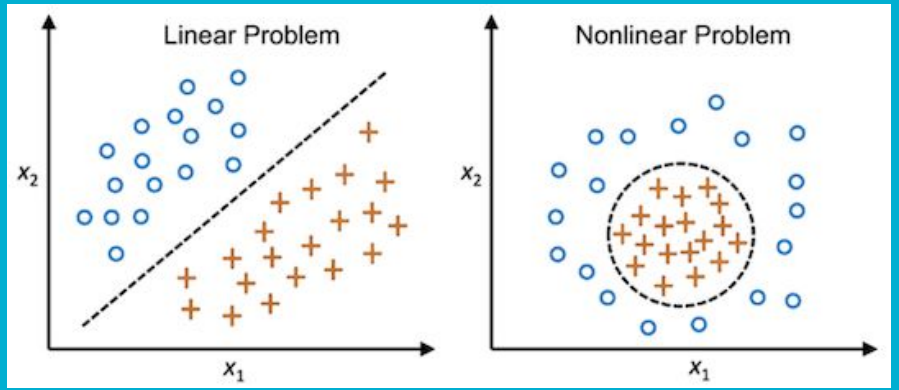
Tuning Hyperparameters

- Distance function
  - The most common are pictured to the right
- The Manhattan is optimal for when the data has more dimensions
- The Euclidean is best for when the data has fewer dimensions
- The Euclidean function is the most commonly used
- The Minkowski is the general form of the distance functions



Advantages

- Intuitive
- Simple and easy to implement with no training time needed
- Only needs k-value and distance metric hyperparameters
- Predictions automatically adjust to new data
- Works in non-linear situations



Disadvantages

- Highly dependent on the user setting an appropriate k-value
- Does not scale well with large amounts of data and data with many dimensions
  - Entire training set is used in every prediction
  - As data gains more dimensions, the distance between points and costs of calculating that distance increases
- Data needs to be standardized and/or normalized before use
- Algorithm is sensitive to outliers
- The algorithm places equal importance in all features
  - Points that are very close in many dimensions, but far away in a few may not be picked up in the points k nearest neighbors

Pre-processing

- Handle missing values
  - Interpolate unknown values
  - Remove unknown values
- Remove unique identifiers
- Remove outliers
- Convert categories to dummies
- Normalize or standardize data

```
# standardizing data
sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.transform(X_test)
```

```
# define the model
classifier = KNeighborsClassifier(n_neighbors=11, p=2, metric='euclidean')
```

```
# fit the model
classifier.fit(X_train, y_train)
```

```
KNeighborsClassifier(metric='euclidean', n_neighbors=11)
```

```
# predict results
y_pred = classifier.predict(X_test)
y_pred
```

Appendix 1

Link	Type (PDF, Video, Article, etc.)
sklearn.neighbors.KNeighborsClassifier — scikit-learn 1.1.2 documentation	Documentation
Introduction to machine learning: k-nearest neighbors	Introduction with working example
StatQuest: K-nearest neighbors, Clearly Explained	Quick video introducing KNN algorithm
Machine Learning Basics with the K-Nearest Neighbors Algorithm	Article that dives into an example of regression and classification.
What is the k-nearest neighbors algorithm?	IBM article that includes general info, advantages and disadvantages, and example code.
K-Nearest Neighbors in Python + Hyperparameters Tuning Data for code along	General overview of KNN with code along.

Appendix 2

Machine Learning Tutorial Python - 18: K nearest neighbors classification with python code	Video with theory explanation, and code example
Machine Learning | KNN Algorithm | Machine Learning with Python | KNN Algorithm 2022	Video overview of what the KNN algorithm is, how it works, and how to Implementation in Python. Code along at the end!
K-Nearest Neighbor(KNN) Algorithm for Machine Learning	Article that shows an example of classification
The Professionals Point: Advantages and Disadvantages of KNN Algorithm in Machine Learning	Breaks down advantages and disadvantages of KNN
How does KNN algorithm work ? What are the advantages and disadvantages of KNN ? - Machine Learning Interviews	Breaks down advantages and disadvantages of KNN
Why is scaling required in KNN and K-Means?	KNN is affected by the scale of the variables. So all variables have equal weightage, the variables need to be in the same range