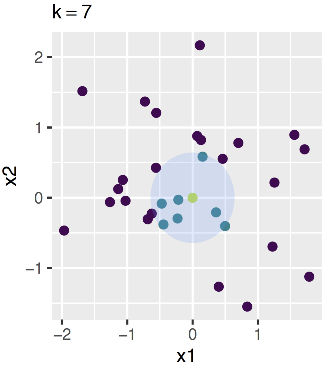
Introduction to Machine Learning

k-Nearest Neighbors Regression



Learning goals

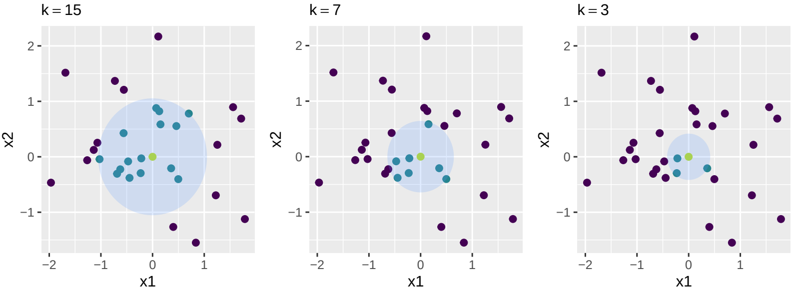
- Understand the basic idea of k-NN
- Know different distance measures for different scales of feature variables
- Understand that k-NN has no optimization step

NEAREST NEIGHBORS: INTUITION

- Say we know locations of cities in 2 different countries.
- Say we know which city is in which country.
- Say we don't know where the countries' border is.
- For a given location, we want to figure out which country it belongs to.
- Nearest neighbor rule: every location belongs to the same country as the closest city.
- k-nearest neighbor rule: vote over the k closest cities (smoother)

K-NEAREST-NEIGHBORS

- k-**NN** can be used for regression and classification
- It generates predictions $\hat{y}$ for a given $\mathbf{x}$ by comparing the k observations that are closest to $\mathbf{x}$
- "Closeness" requires a distance or similarity measure (usually: Euclidean).
- The set containing the k closest points $\mathbf{x}^{(i)}$ to $\mathbf{x}$ in the training data is called the k-**neighborhood** $N_k(\mathbf{x})$ of $\mathbf{x}$.



DISTANCE MEASURES

How to calculate distances?

- Most popular distance measure for numerical features: **Euclidean distance**
- Imagine two data points $\mathbf{x} = (x_1, ..., x_p)$ and $\tilde{\mathbf{x}} = (\tilde{x}_1, ..., \tilde{x}_p)$ with p features $\in \mathbb{R}$
- The Euclidean distance:

$$d_{Euclidean}(\mathbf{x}, \tilde{\mathbf{x}}) = \sqrt{\sum_{j=1}^{p} (x_j - \tilde{x}_j)^2}$$

DISTANCE MEASURES

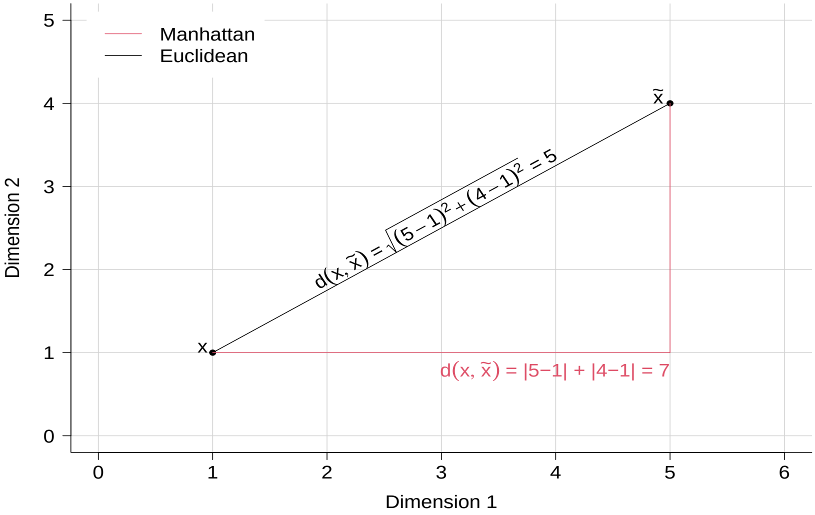
- Example:
  - Three data points with two metric features each: $a = (1, 3), b = (4, 5)$ and $c = (7, 8)$
  - Which is the nearest neighbor of b in terms of the Euclidean distance?
  - $d(b, a) = \sqrt{(4-1)^2 + (5-3)^2} = 3.61$
  - $d(b, c) = \sqrt{(4-7)^2 + (5-8)^2} = 4.24$
  - $\Rightarrow$ a is the nearest neighbor for b.
- Alternative distance measures are:
  - Manhattan distance

$$d_{manhattan}(\mathbf{x}, \tilde{\mathbf{x}}) = \sum_{j=1}^{p} |x_j - \tilde{x}_j|$$

  - Mahalanobis distance (takes covariances in $\mathcal{X}$ into account)

DISTANCE MEASURES

Comparison between Euclidean and Manhattan distance measures:



DISTANCE MEASURES

Categorical variables, missing data and mixed space:
The Gower distance $d_{gower}(\mathbf{x}, \tilde{\mathbf{x}})$ is a weighted mean of $d_{gower}(x_j, \tilde{x}_j)$:

$$d_{gower}(\mathbf{x}, \tilde{\mathbf{x}}) = \frac{\sum\limits_{j=1}^{p} \delta_{x_j,\tilde{x}_j} \cdot d_{gower}(x_j, \tilde{x}_j)}{\sum\limits_{j=1}^{p} \delta_{x_j,\tilde{x}_j}}.$$

- $\delta_{x_j,\tilde{x}_j}$ is 0 or 1. It becomes 0 when the j-th variable is ***missing*** in at least one of the observations ($\mathbf{x}$ or $\tilde{\mathbf{x}}$), or when the variable is asymmetric binary (where “1” is more important/distinctive than “0”, e. g., “1” means “color-blind”) and both values are zero. Otherwise it is 1.

DISTANCE MEASURES

- $d_{gower}(x_j, \tilde{x}_j)$, the j-th variable contribution to the total distance, is a distance between the values of x_j and $\tilde{x}_j$. For nominal variables the distance is 0 if both values are equal and 1 otherwise. The contribution of other variables is the absolute difference of both values, divided by the total range of that variable.

DISTANCE MEASURES

Example of Gower distance with data on sex and income:

index	sex	salary
1	m	2340
2	w	2100
3	NA	2680

$$d_{gower}(\mathbf{x}, \tilde{\mathbf{x}}) = \frac{\sum_{j=1}^{p} \delta_{x_j,\tilde{x}_j} \cdot d_{gower}(x_j, \tilde{x}_j)}{\sum_{j=1}^{p} \delta_{x_j,\tilde{x}_j}}$$

$$d_{gower}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \frac{1 \cdot 1 + 1 \cdot \frac{|2340-2100|}{|2680-2100|}}{1+1} = \frac{1+\frac{240}{580}}{2} = \frac{1+0.414}{2} = 0.707$$

$$d_{gower}(\mathbf{x}^{(1)}, \mathbf{x}^{(3)}) = \frac{0 \cdot 1 + 1 \cdot \frac{|2340-2680|}{|2680-2100|}}{0+1} = \frac{0+\frac{340}{580}}{1} = \frac{0+0.586}{1} = 0.586$$

$$d_{gower}(\mathbf{x}^{(2)}, \mathbf{x}^{(3)}) = \frac{0 \cdot 1 + 1 \cdot \frac{|2100-2680|}{|2680-2100|}}{0+1} = \frac{0+\frac{580}{580}}{1} = \frac{0+1.000}{1} = 1$$

DISTANCE MEASURES

Weights:

Weights can be used to address two problems in distance calculation:

- **Standardization:** Two features may have values with a different scale. Many distance formulas (not Gower) would place a higher importance on a feature with higher values, leading to an imbalance. Assigning a higher weight to the lower-valued feature can combat this effect.
- **Importance:** Sometimes one feature has a higher importance (e. g., more recent measurement). Assigning weights according to the importance of the feature can align the distance measure with known feature importance.

For example:

$$d_{Euclidean}^{weighted}(\mathbf{x}, \tilde{\mathbf{x}}) = \sqrt{\sum_{j=1}^{p} w_j (x_j - \tilde{x}_j)^2}$$

K-NN REGRESSION

Predictions for regression:

$$\hat{y} = \frac{1}{k} \sum_{i: \mathbf{x}^{(i)} \in N_k(\mathbf{x})} y^{(i)}$$

$$\hat{y} = \frac{1}{\sum_{i: \mathbf{x}^{(i)} \in N_k(\mathbf{x})} w_i} \sum_{i: \mathbf{x}^{(i)} \in N_k(\mathbf{x})} w_i y^{(i)}$$

with neighbors weighted according to their distance to $\mathbf{x}$: $w_i = \frac{1}{d(\mathbf{x}^{(i)}, \mathbf{x})}$

K-NN SUMMARY

- k-NN has no optimization step and is a very local model.
- We cannot simply use least-squares loss on the training data for picking k, because we would always pick $k = 1$.
- k-NN makes no assumptions about the underlying data distribution.
- The smaller k, the less stable, less smooth and more “wiggly” the decision boundary becomes.
- Accuracy of k-NN can be severely degraded by the presence of noisy or irrelevant features, or when the feature scales are not consistent with their importance.

K-NN SUMMARY

Hypothesis Space: Step functions over tesselations of $\mathcal{X}$.
Hyperparameters: distance measure $d(\cdot, \cdot)$ on $\mathcal{X}$; size of neighborhood k.

Risk: Use any loss function for regression or classification.

Optimization: Not applicable/necessary.
But: clever look-up methods & data structures to avoid computing all n distances for generating predictions.