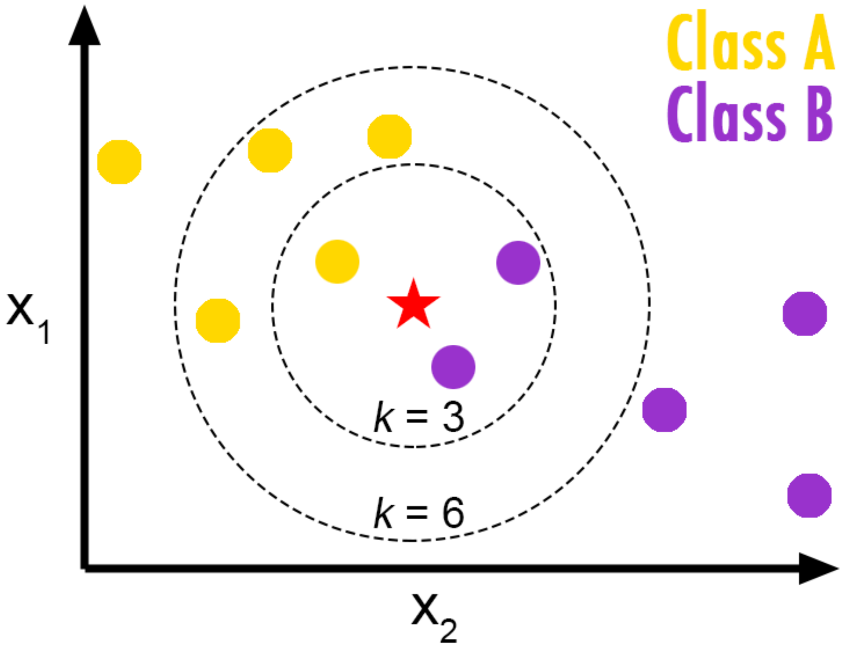
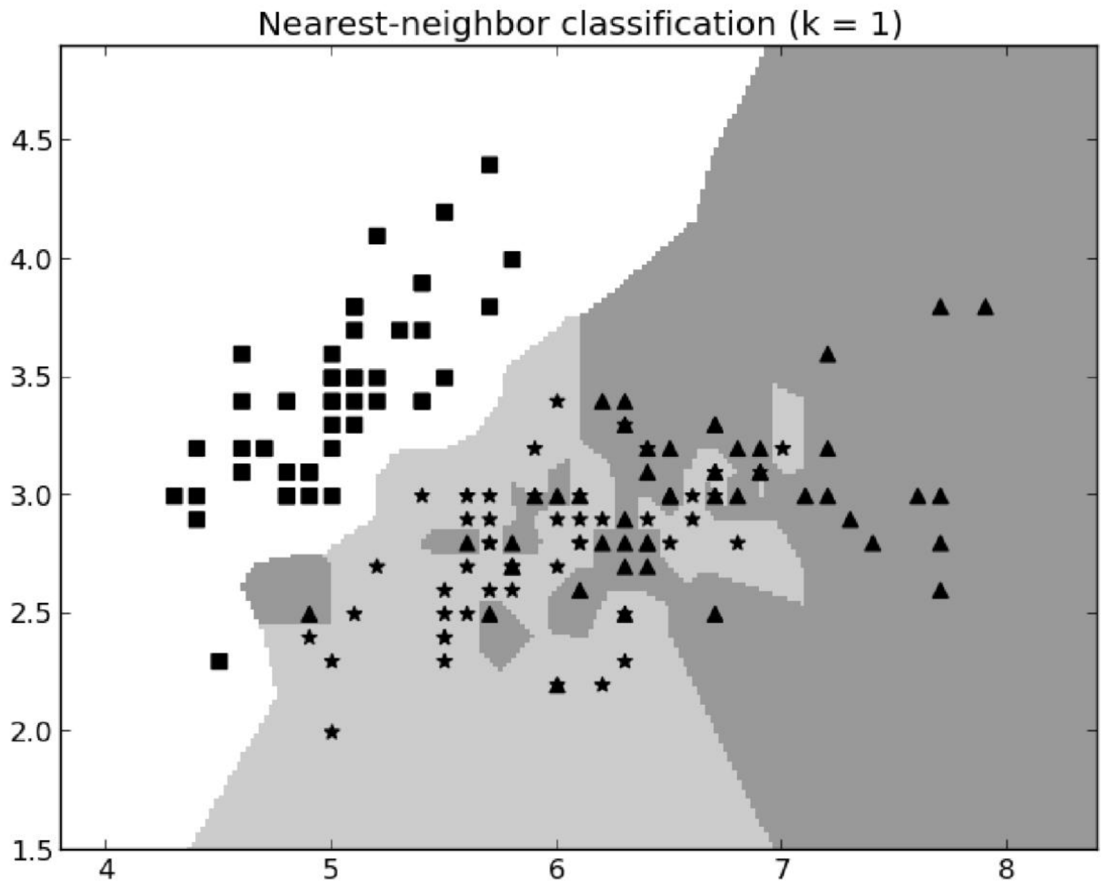
- K-Nearest Neighbors

K-Nearest Neighbors (K-NN)

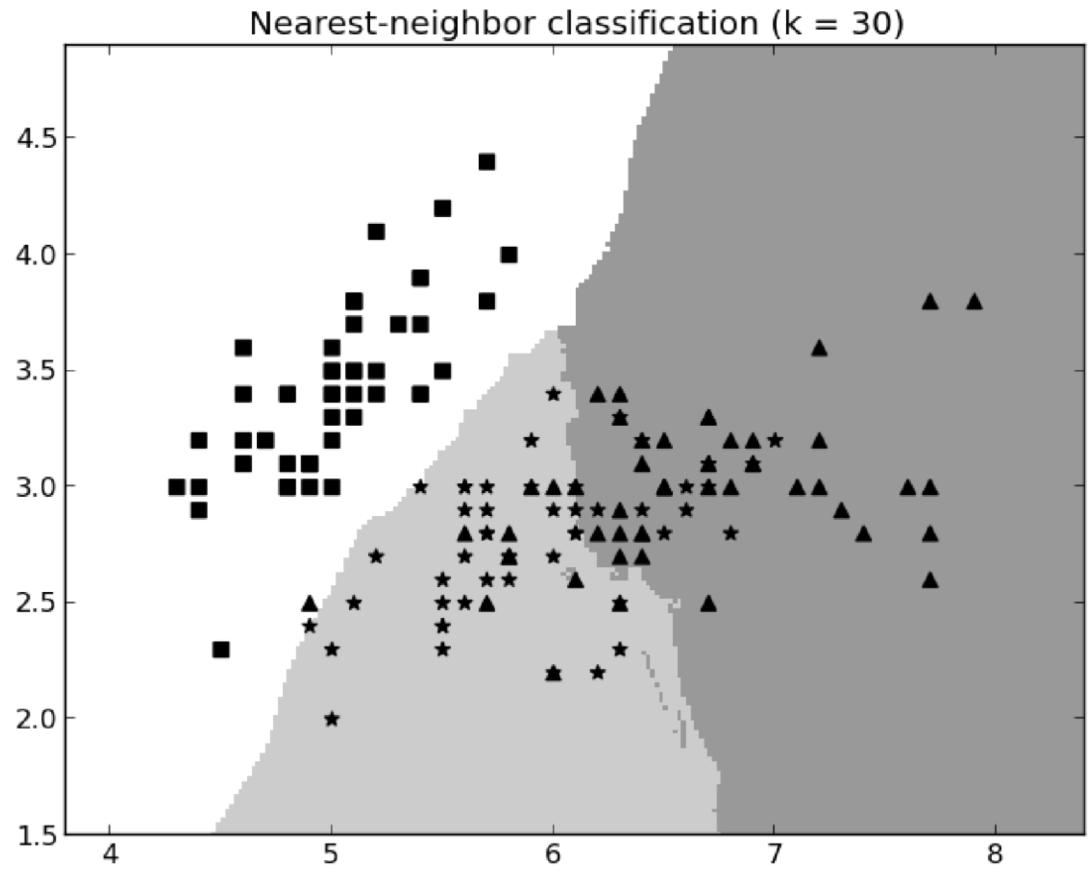


$k = ?$

1-Nearest Neighbor



30-Nearest Neighbors

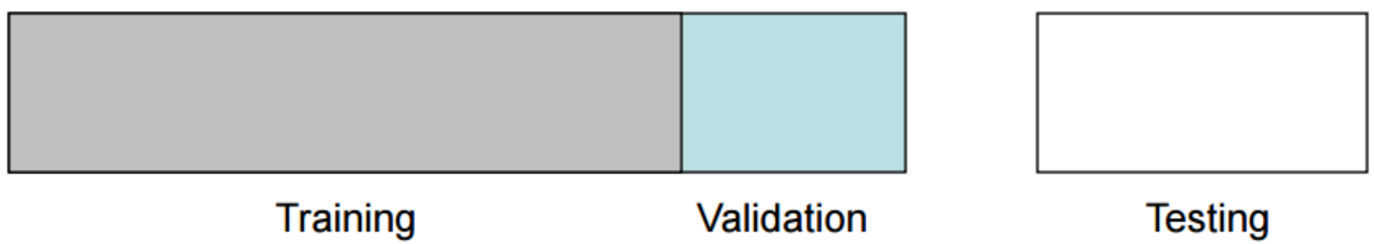


Different Setting of K

- Low values of k (1, 3 …) capture local structure in data (but also noise)
- High values of k provide more smoothing, less noise, but may miss local structure
  - Note: the extreme case of $k = |D|$ (i.e. the entire data set) is the same thing as "naïve rule" (classify all records according to majority class)

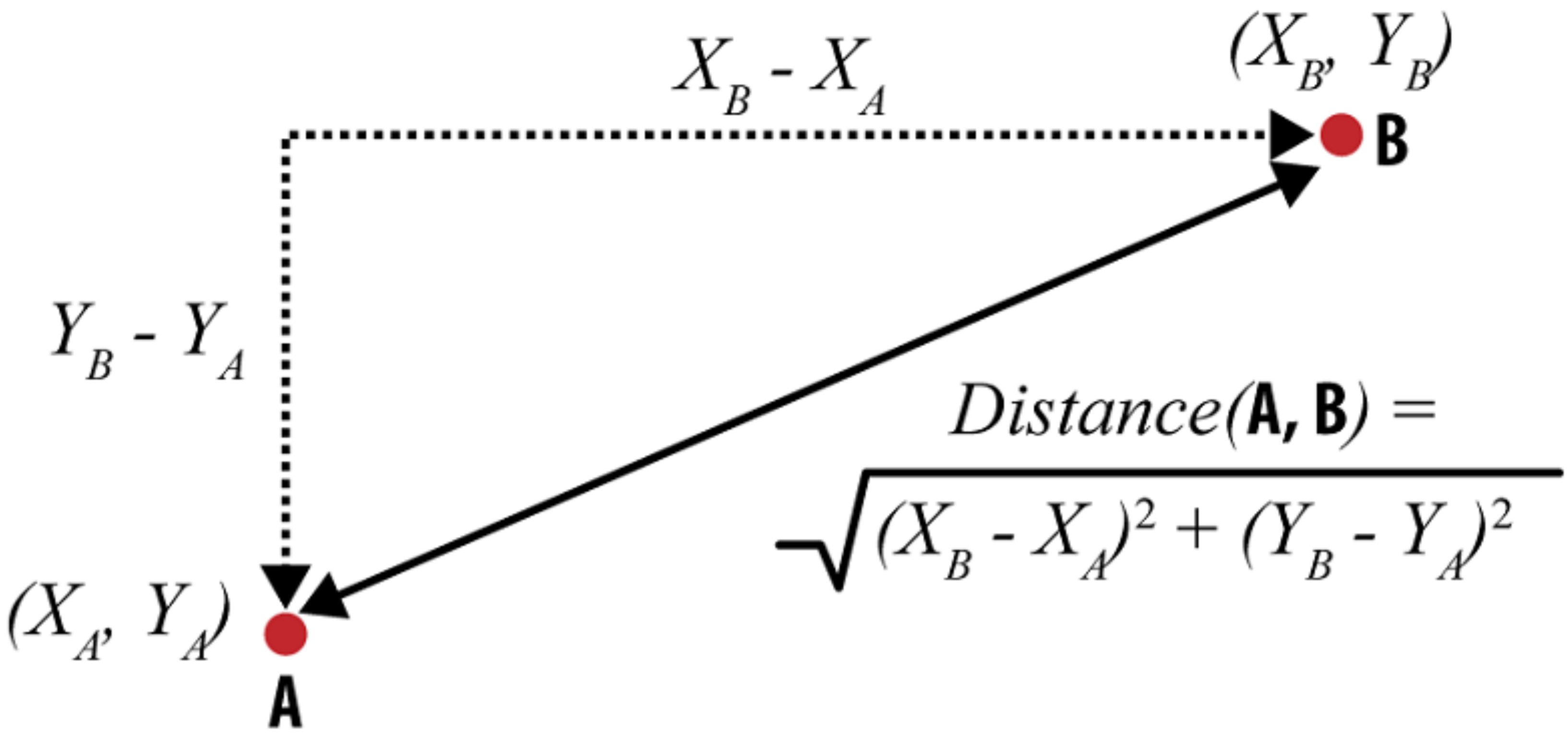
Training-Testing-Validation

- We can keep part of the labeled data apart as validation data
- Evaluate different k values based on the prediction accuracy on the validation data
- Choose k that minimize validation error



Validation can be viewed as another name for testing, but the name ***testing*** is typically reserved for final evaluation purpose, whereas ***validation*** is mostly used for model selection purpose.

Euclidean Distance



What's problem here?

Customer	Age	Income (1000s)	Cards	Response (target)	Distance from David
David	*37*	*50*	*2*	*?*	0
John	35	35	3	Yes	$\sqrt{(35-37)^2+(35-50)^2+(3-2)^2}=15.16$
Rachael	22	50	2	No	$\sqrt{(22-37)^2+(50-50)^2+(2-2)^2}=15$
Ruth	63	200	1	No	$\sqrt{(63-37)^2+(200-50)^2+(1-2)^2}=152.23$
Jefferson	59	170	1	No	$\sqrt{(59-37)^2+(170-50)^2+(1-2)^2}=122$
Norah	25	40	4	Yes	$\sqrt{(25-37)^2+(40-50)^2+(4-2)^2}=15.74$

Dominate the distance calculation

Before using any model that measure distance, always Normalize

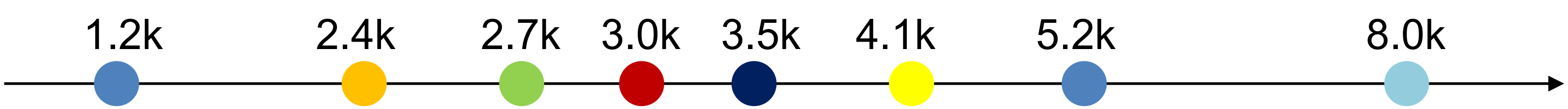
Normalization Approaches

- Scaling/normalizing is used to standardize the intervals before measuring distances
  - min-max scaling numerical attribute to interval [0,1]

$$z = \boxed{(x - \min) / (\max - \min)}$$

  - x := original value of the attribute
  - min := smallest value of the attribute
  - max := largest value of the attribute
  - z is the resulting (scaled) value of the attribute in the range [0,1]

Income:



Normalized:

0 0.18 0.22 0.26 0.33 0.42 0.59 1

Min-Max Approach Example

- Setting: Consider the age and income data of several employees

Name	Age	Income
Alice	70	100,000
Bob	25	50,000
Cindy	30	60,000
David	20	70,000
Earl	60	80,000

Name	Age(Norm)	Income(Norm)
Alice	1.0	1.0
Bob	0.1	0.0
Cindy	0.2	0.2
David	0.0	0.4
Earl	0.8	0.6

Why k-NN?

- "Lazy" learning approach
  - As opposed to "eager" approaches (e.g., decision trees)
  - No model building (data as model) → Faster to train but slower to estimate
- Enhancements
  - Weighted distance (closer neighbors have more impact)
- Strengths
  - Easy to implement and use
  - Robust (handles noisy data well, except for very low k values)
  - No statistical / distributional assumptions required
  - Captures complex interactions between variables without building models
- Weaknesses
  - Takes more time to perform estimation; computational efficiency
  - Requires a lot of storage
  - Dimensionality and domain knowledge

Other Distance Functions

$d_{Euclidean}(X, Y)$

$$d_{Manhattan}(\boldsymbol{X}, \boldsymbol{Y}) = \|\boldsymbol{X} - \boldsymbol{Y}\|_1 = |x_1 - y_1| + |x_2 - y_2| + \cdots$$

For Numeric Variables

$$d_{Jaccard}(X, Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}$$

For Categorical Variables

$$d_{Cosine}(\boldsymbol{X}, \boldsymbol{Y}) = 1 - \frac{\boldsymbol{X} \cdot \boldsymbol{Y}}{\|\boldsymbol{X}\|_2 \cdot \|\boldsymbol{Y}\|_2}$$

More Normalization Approaches

距離平均數幾個標準差

- $\boxed{\text{z-score}}$ scaling to standardize the intervals

$$\boxed{z = (x - m) / s}$$

  - m = mean value of the attribute
  - s = standard deviation (or mean absolute deviation, which is more robust to outliers than standard deviation)

- When attributes have different importance
  - Weighted distances may be used
  - E.g., $d(A, B) = w_1|a_1 - b_1| + \cdots + w_k|a_k - b_k|$