Data Mining for Business Analytics

Lecture 10: Similarity and Nearest Neighbors

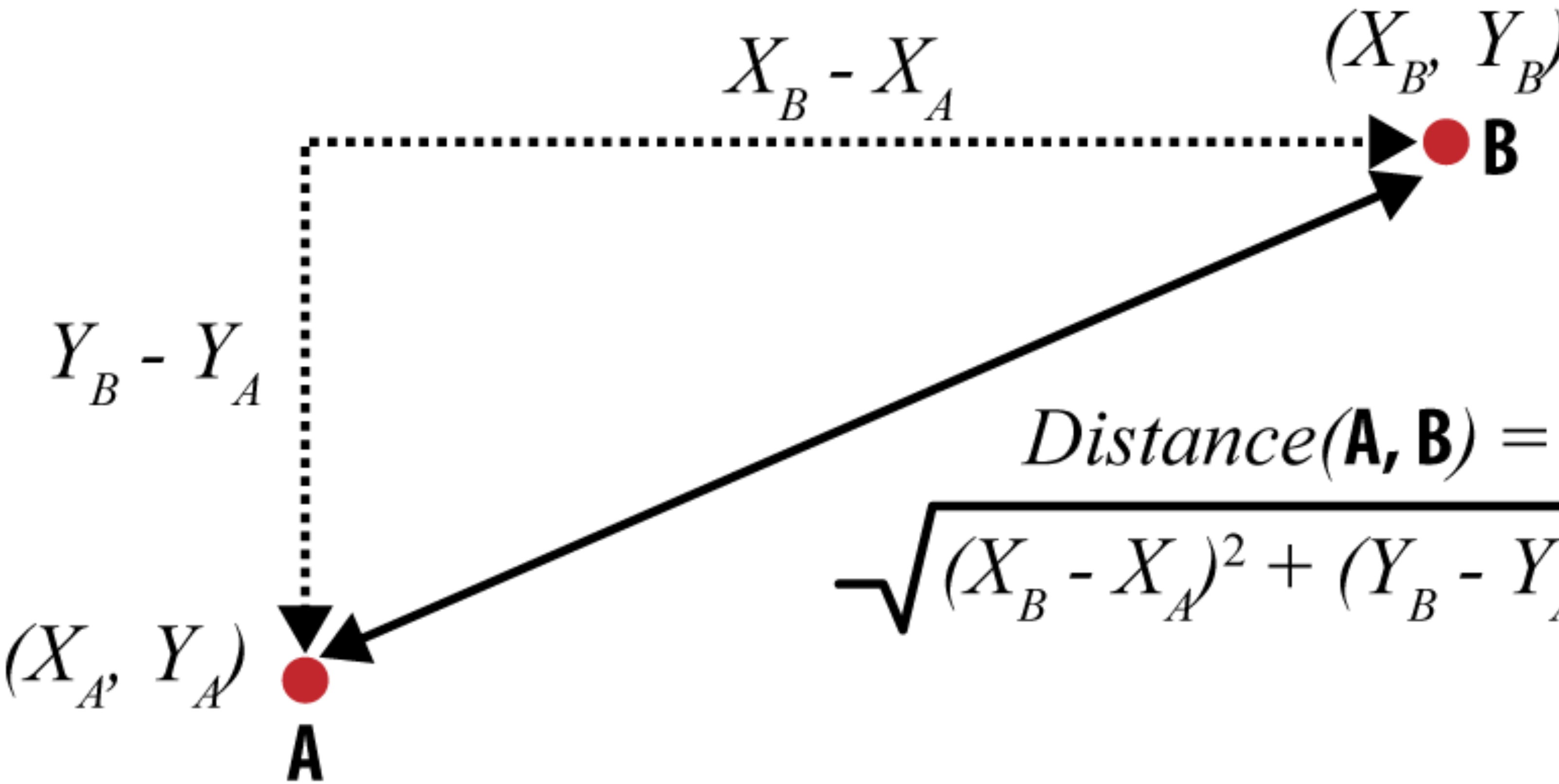
Stern School of Business
New York University
Spring 2014

Similarity and Distance

- If two objects can be represented as feature vectors, then we can compute the distance between them

Attribute	Person A	Person B
Age	23	40
Years at current address	2	10
Residential status (1=Owner, 2=Renter, 3=Other)	2	1

Euclidean Distance



$$Distance(\mathbf{A}, \mathbf{B}) = \sqrt{(X_B - X_A)^2 + (Y_B - Y_A)^2}$$

Euclidean Distance

$$\sqrt{\left(d_{1,A} - d_{1,B}\right)^2 + \left(d_{2,A} - d_{2,B}\right)^2 + \cdots + \left(d_{n,A} - d_{n,B}\right)^2}$$

$||A,B||_2$ represents the L2 norm

$$d(A,B) = \sqrt{(23-40)^2 + (2-10)^2 + (2-1)^2} = 18.8$$

Other Distance Functions

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

(L1-norm, taxicab-distance)

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

where $||\cdot||_2$ represents the L2 norm, or Euclidean length, of each feature vector (for a vector this is simply the distance from the origin).

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

where, X and Y are sets

Example: "Whiskey Analytics"

1.	**Color:** *yellow, very pale, pale, pale gold, gold, old gold, full gold, amber,* etc.	(14 values)
2.	**Nose:** *aromatic, peaty, sweet, light, fresh, dry, grassy,* etc.	(12 values)
3.	**Body:** *soft, medium, full, round, smooth, light, firm, oily.*	(8 values)
4.	**Palate:** *full, dry, sherry, big, fruity, grassy, smoky, salty,* etc.	(15 values)
5.	**Finish:** *full, dry, warm, light, smooth, clean, fruity, grassy, smoky,* etc.	(19 values)

Consequently there are 68 binary features of each whiskey.

Whiskey	Distance	Descriptors
Bunnahabhain	—	*gold; firm,med,light; sweet,fruit,clean; fresh,sea; full*
Glenglassaugh	0.643	gold; firm,light,smooth; sweet,grass; fresh,grass
Tullibardine	0.647	gold; firm,med,smooth; sweet,fruit,full,grass,clean; sweet; big,arome,sweet
Ardbeg	0.667	sherry; firm,med,full,light; sweet; dry,peat,sea;salt
Bruichladdich	0.667	pale; firm,light,smooth; dry,sweet,smoke,clean; light; full
Glenmorangie	0.667	p.gold; med,oily,light; sweet,grass,spice; sweet,spicy,grass,sea,fresh; full,long

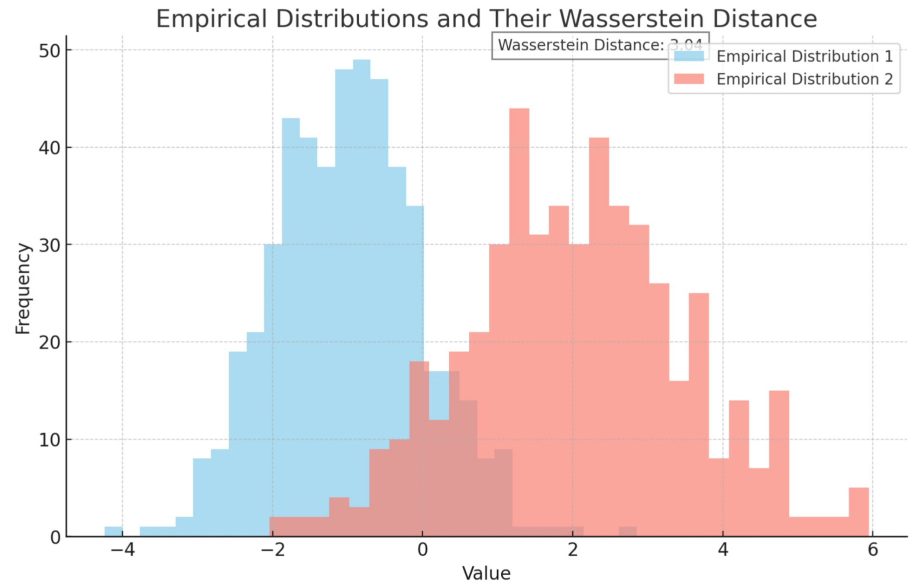
Introduction to Wasserstein Distance

What is Wasserstein Distance?

- Also known as the Earth Mover's Distance (EMD).
- Measures the distance between two probability distributions over a given metric space.
- Intuitively, it represents the minimum cost of transporting mass to transform one distribution into the other.

Key Properties

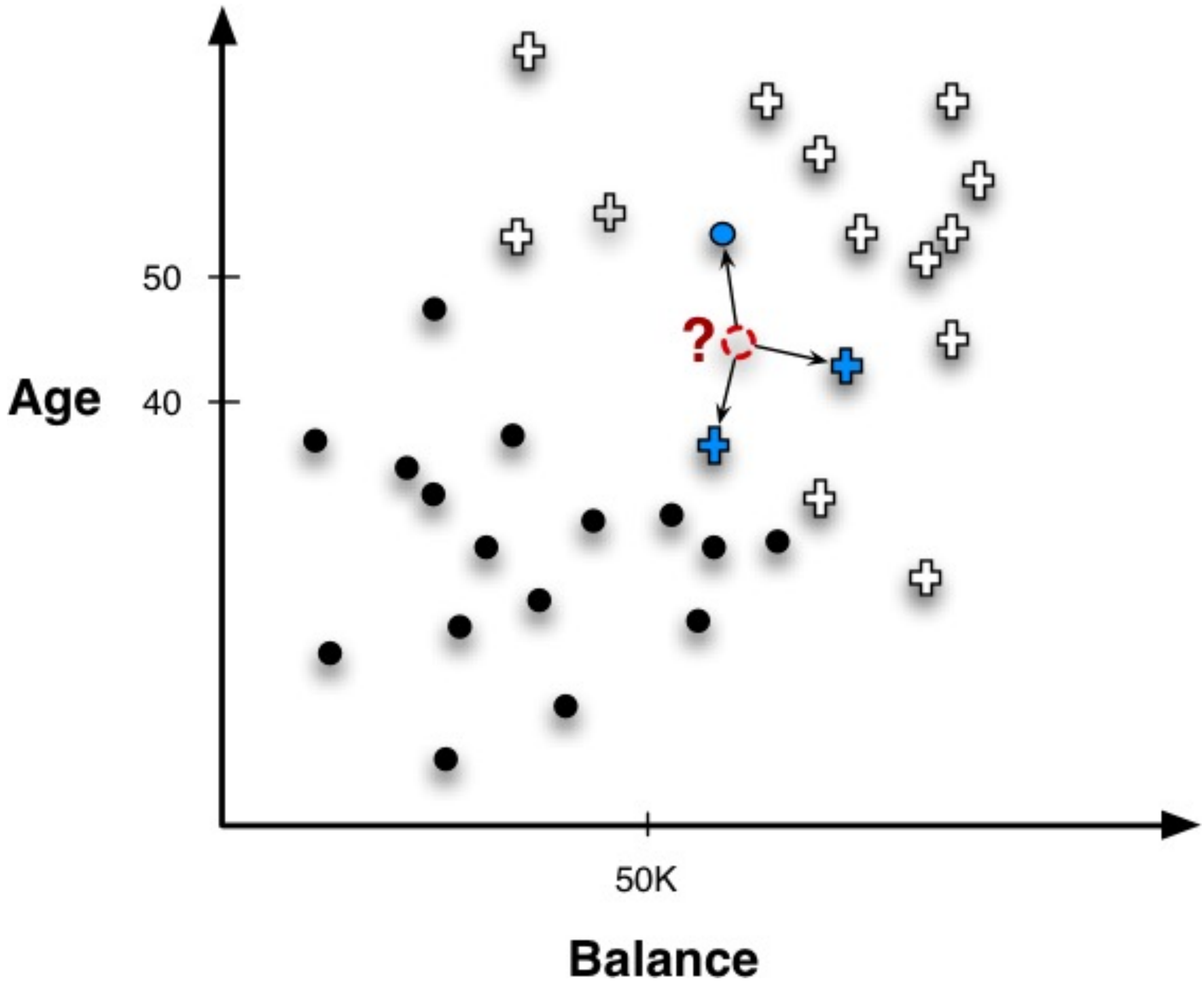
- **Metric:** Wasserstein distance is a proper distance metric, satisfying non-negativity, symmetry, and the triangle inequality.
- **Interpretability:** Offers a more intuitive and meaningful distance measure for probability distributions compared to other distances (e.g., Euclidean, KL divergence).
- **Applications:** Widely used in various fields such as optimal transport, machine learning (especially in generative adversarial networks - GANs), image retrieval, and more.



I've updated the plot once more, this time adjusting the legend to avoid overlap with the histogram. This should provide a clearer view of both empirical distributions and the Wasserstein distance between them.

Nearest Neighbors

Nearest Neighbors for Predictive Modeling



Nearest Neighbors for Predictive Modeling

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	$\sqrt{(25-37)^2+(40-50)^2+(4-2)^2}=15.74$

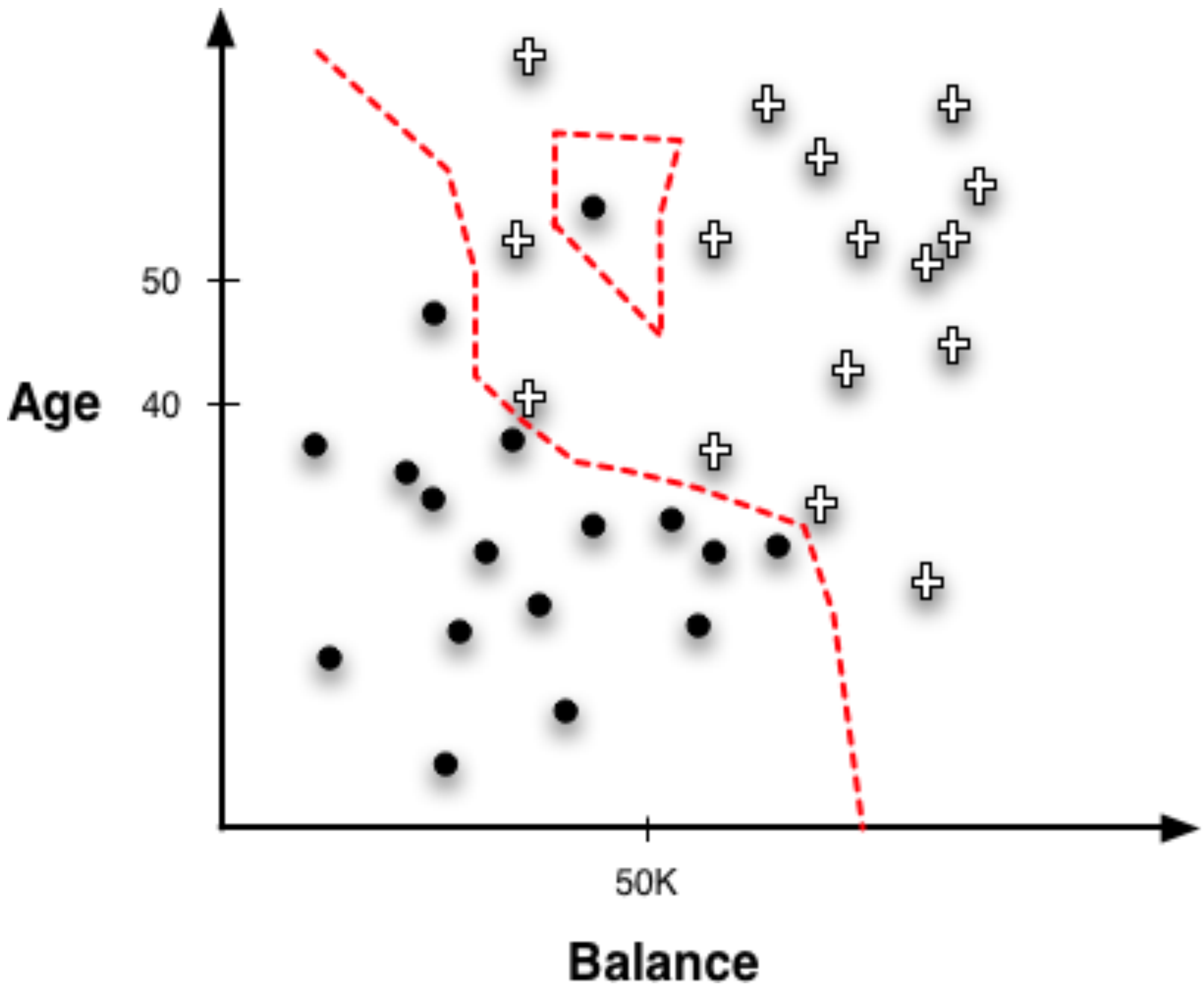
How Many Neighbors and How Much Influence?

k Nearest Neighbors

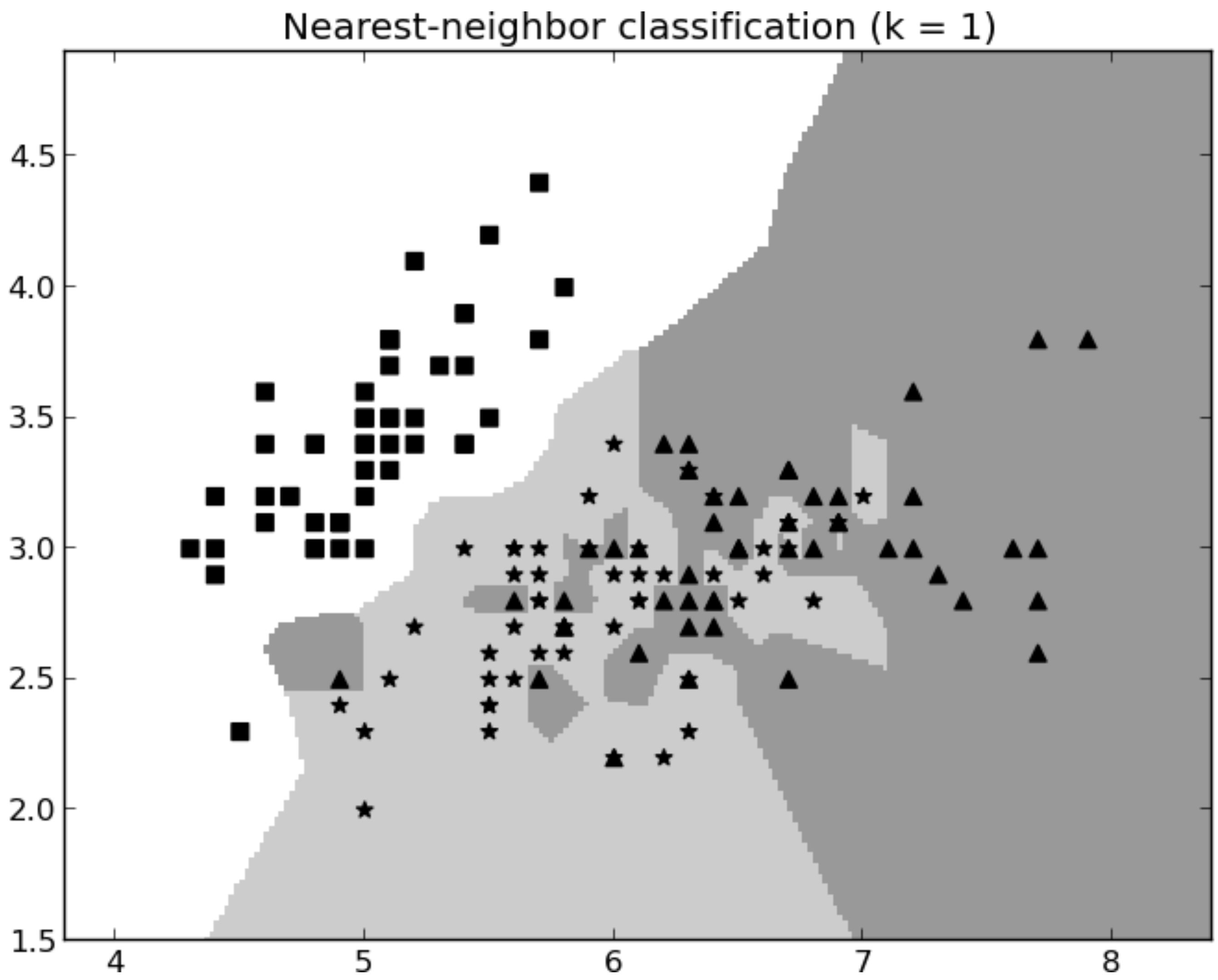
- $k = ?$
- $k = 1\ ?$
- $k = n\ ?$

Geometric Interpretation, Over-fitting, and Complexity

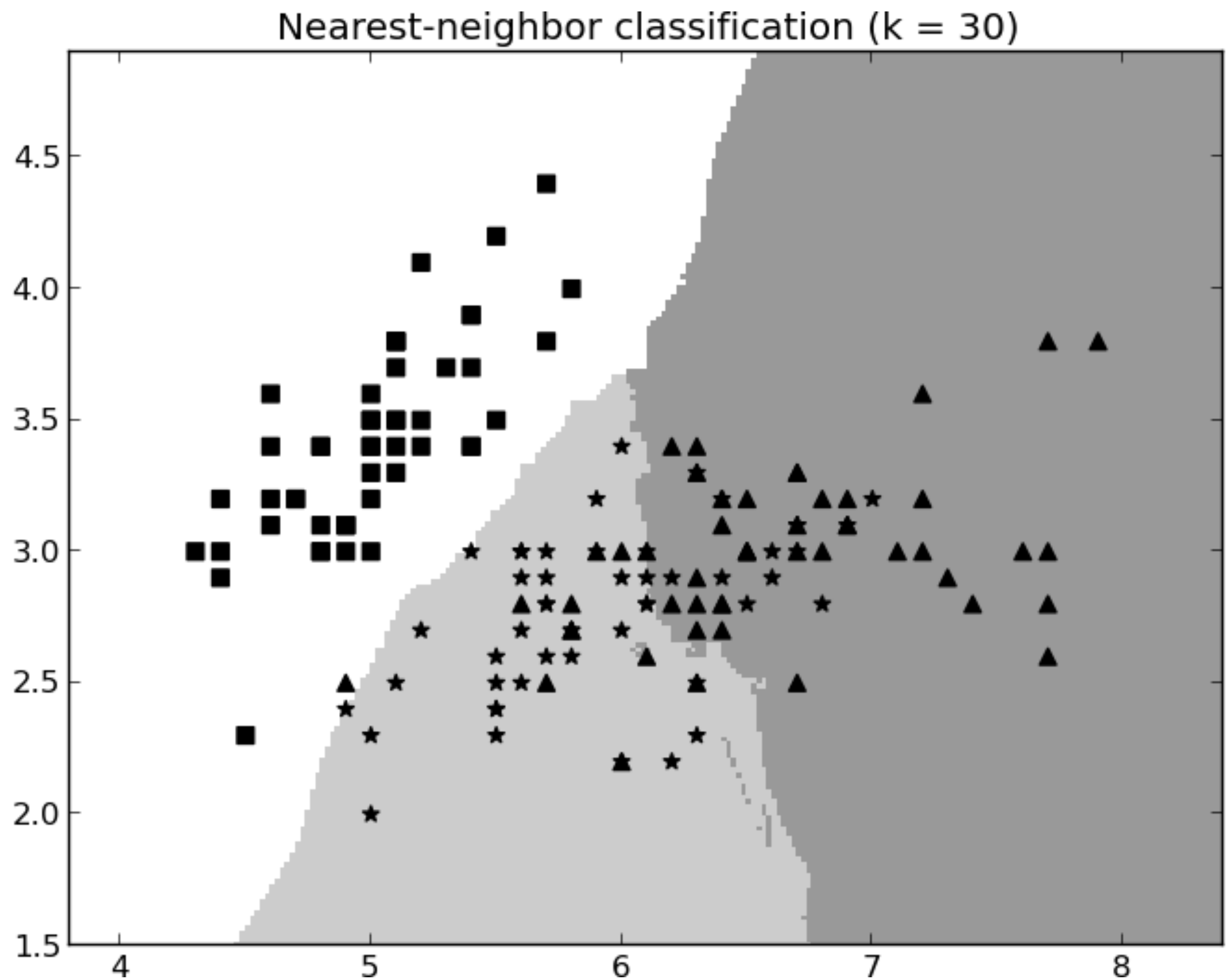
Boundaries created by a 1-NN classifier.



1-Nearest Neighbor



30-Nearest Neighbors



Issues with Nearest-Neighbor Models

- Dimensionality and domain knowledge
  - There might be too many features (and some are irrelevant)
  - The distance function need to consider the scale and importance of the features.

- Computational efficiency
  - Not suitable for online advertisement, whose decisions have to be made in a few tens of milliseconds.