Similarity and Neighbors: K-NN

Source: Provost and Fawcett (2013)

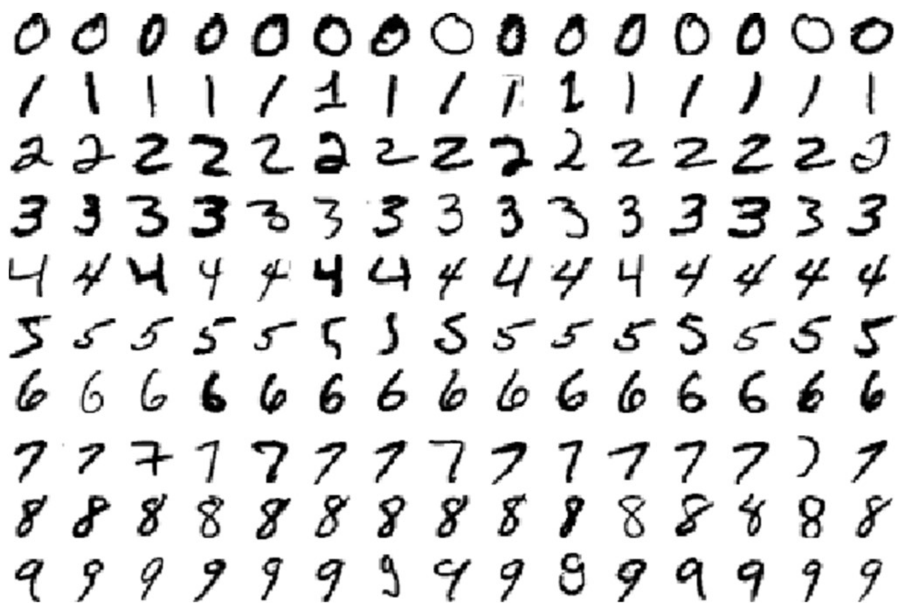
Similarity and Distance

- 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

Example: OCR for digits

- Classify images of handwritten digits by the (actual) digits they depict.
- Classification problem: $\mathcal{Y}$ = discrete set



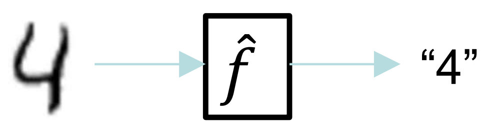
Nearest neighbor (NN) classifier

- **Given**: labeled examples $D \coloneqq \{(x_i, y_i)\}_{i=1}^{n} \subset \mathcal{X} \times \mathcal{Y}$

- **Predictor** $\hat{f}_D : \mathcal{X} \to \mathcal{Y}$:

  On input $x \in \mathcal{X}$:

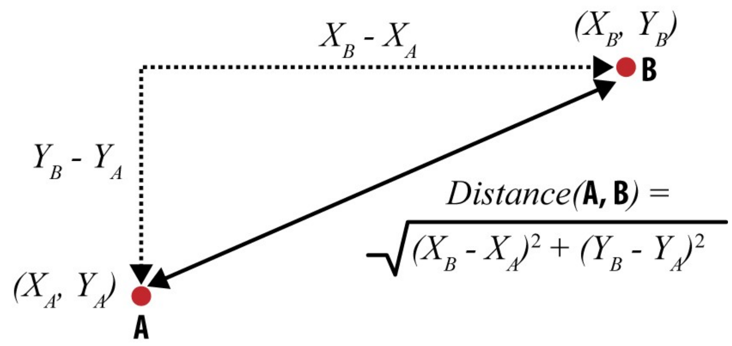
1. Find the point x_i among $\{x_i\}_{i=1}^{n}$ “closest” to x (nearest neighbor)
2. Return y_i

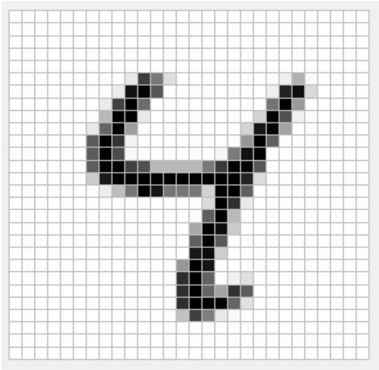


How to measure distance?

- For points in $\mathbb{R}^d$, a default choice for distance is the *Euclidean distance* (also called ℓ_2 distance).

$$\|u - v\|_2 = \sqrt{\sum_{j=1}^{d} \left(u_j - v_j\right)^2}$$





Grayscale 28×28 pixel images.
Treat as *vectors* (of 784 features) that live in $\mathbb{R}^{784}$.

Other Distance Functions

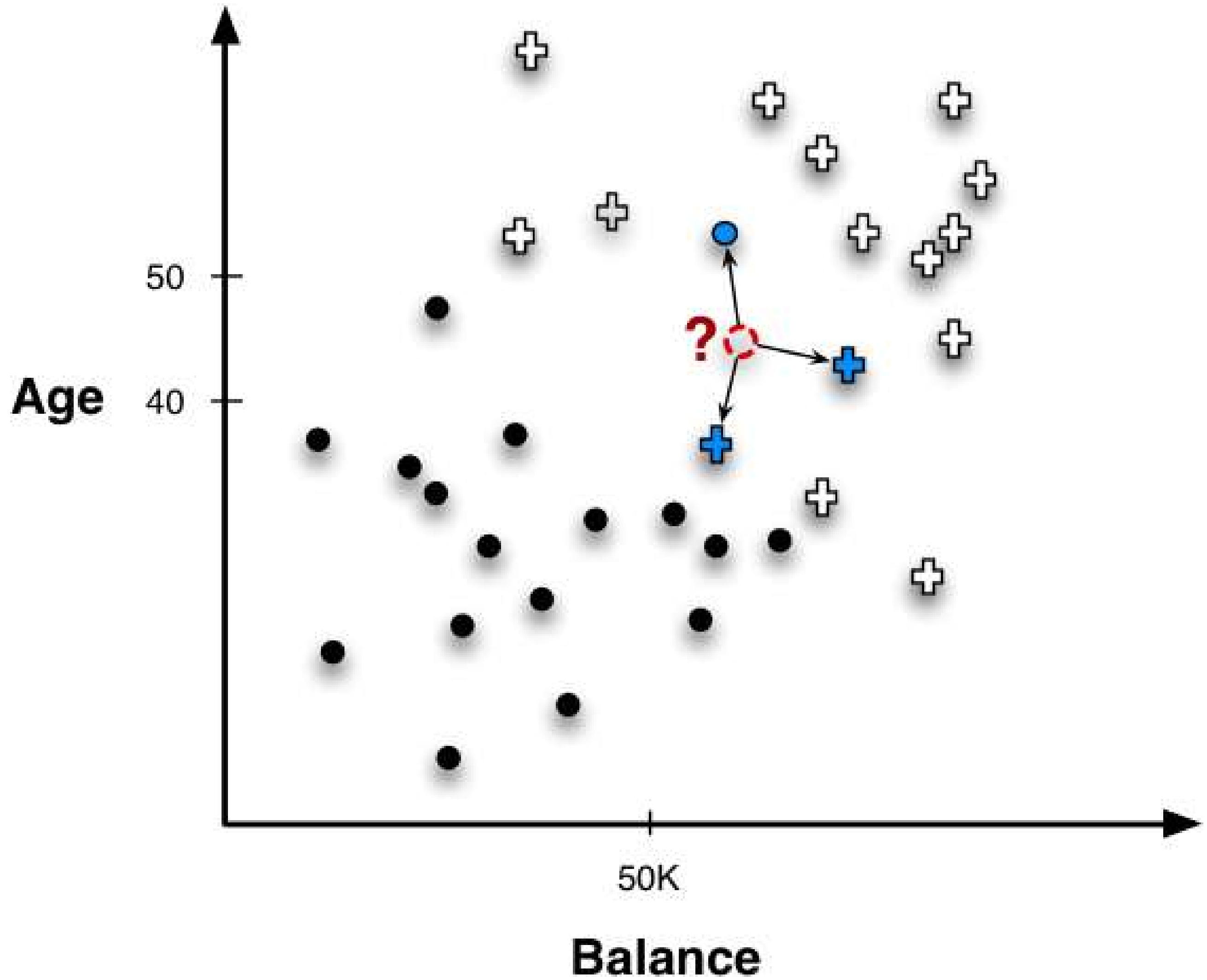
- $d_{Manhattan}(\boldsymbol{X}, \boldsymbol{Y}) = \|\boldsymbol{X} - \boldsymbol{Y}\|_1 = |x_1 - y_1| + |x_2 - y_2| + \cdots$
- $d_{Jaccard}(X, Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}$
- $d_{Cosine}(\boldsymbol{X}, \boldsymbol{Y}) = 1 - \frac{\boldsymbol{X} \cdot \boldsymbol{Y}}{\|\boldsymbol{X}\|_2 \cdot \|\boldsymbol{Y}\|_2}$
- $d(\boldsymbol{X}, \boldsymbol{Y})$ = # insertions/deletions/mutations needed to change x to y (Strings: edit distance)
- $d(\boldsymbol{X}, \boldsymbol{Y})$ = how much "warping" is required to change x to y (Images: shape context distance)

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)

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

Nearest Neighbors for Predictive Modeling

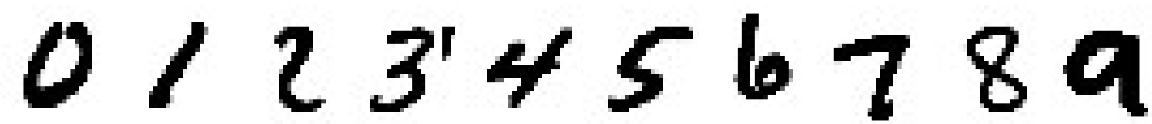


Nearest Neighbors for Predictive Modeling

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

Example: OCR for digits with NN classifier

- Classify images of handwritten digits by they digits they depict.



- $\mathcal{X} = \mathbb{R}^{768}$, $\mathcal{Y} = \{0,1,2,3,4,5,6,7,8,9\}$
- **Given**: labeled examples $D \coloneqq \{(x_i, y_i)\}_{i=1}^{n} \subset \mathcal{X} \times \mathcal{Y}$.
- Construct NN classifier $\hat{f}_D$ using D.
- **Question**: How good is this classifier?

Error rate

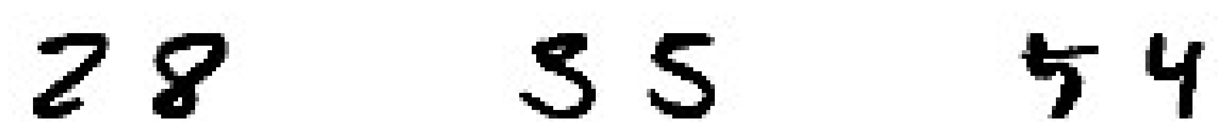
- *Error rate* of classifier f on a set of labeled examples D:

$$\mathrm{err}_D(f) \coloneqq \frac{|\{(x, y) \in D : f(x) \neq y\}|}{|D|}$$

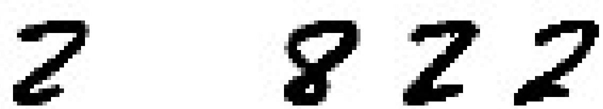
(on what fraction of D does f disagree with the paired label?)

Diagnostics

- Some examples of NN classifier mistakes
  (test point in T, nearest neighbor in S)



- First mistake (correct label is “2”) could’ve been avoided by looking at the three nearest neighbors (whose labels are “8”, “2”, and “2”):



Test point   Three nearest neighbors

k-nearest neighbors (k-NN) classifier

- **Given**: labeled examples $D \coloneqq \{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$
- **Predictor** $\hat{f}_{D,k} \colon \mathcal{X} \to \mathcal{Y}$:

  On input $x \in \mathcal{X}$:

1. Find the k points $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ among $\{x_i\}_{i=1}^n$ “closest” to x (the k nearest neighbors)
2. Return plurality of $y_{i_1}, y_{i_2}, \ldots, y_{i_k}$

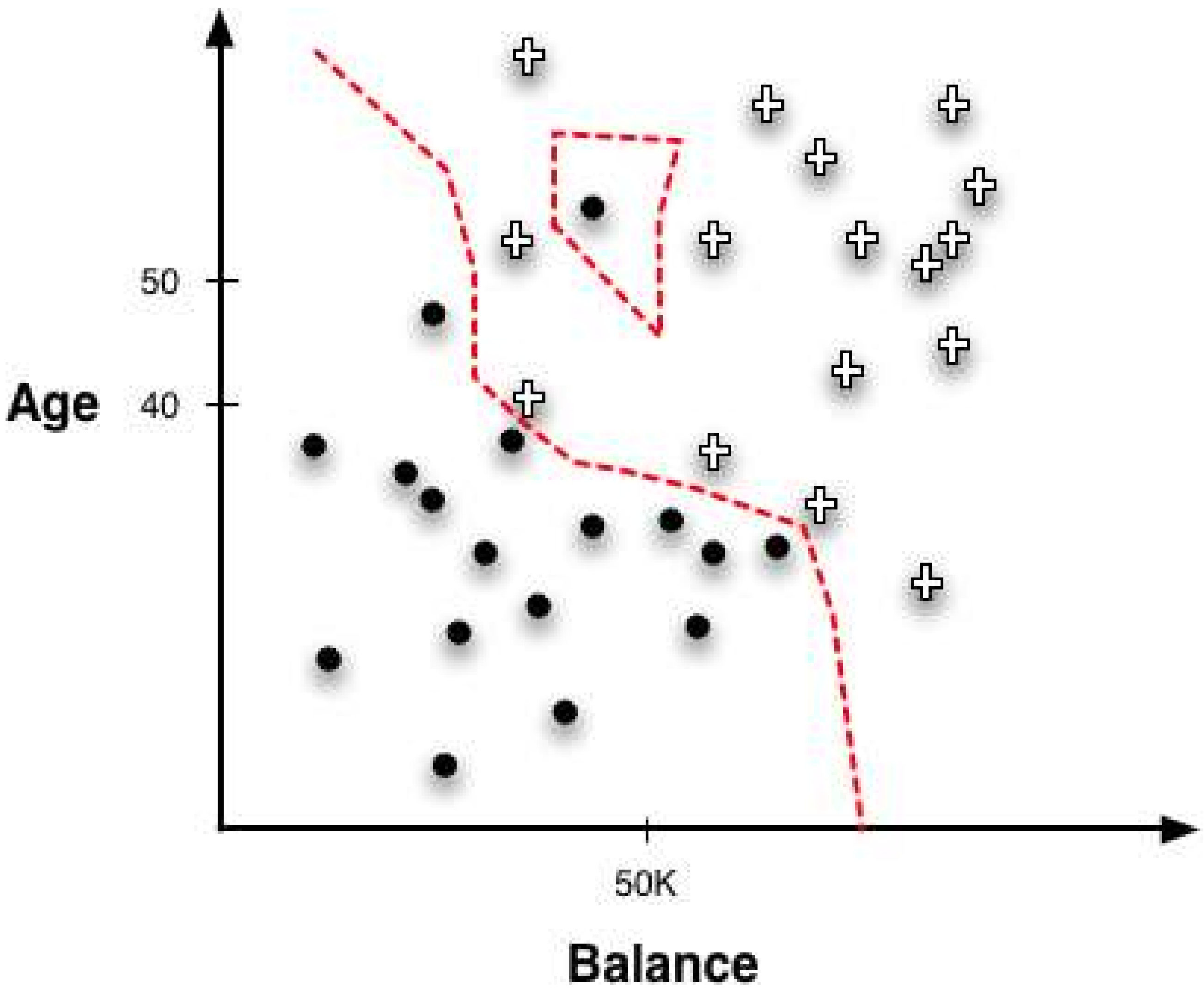
(Break ties arbitrarily in both steps.)

How Many Neighbors and How Much Influence?

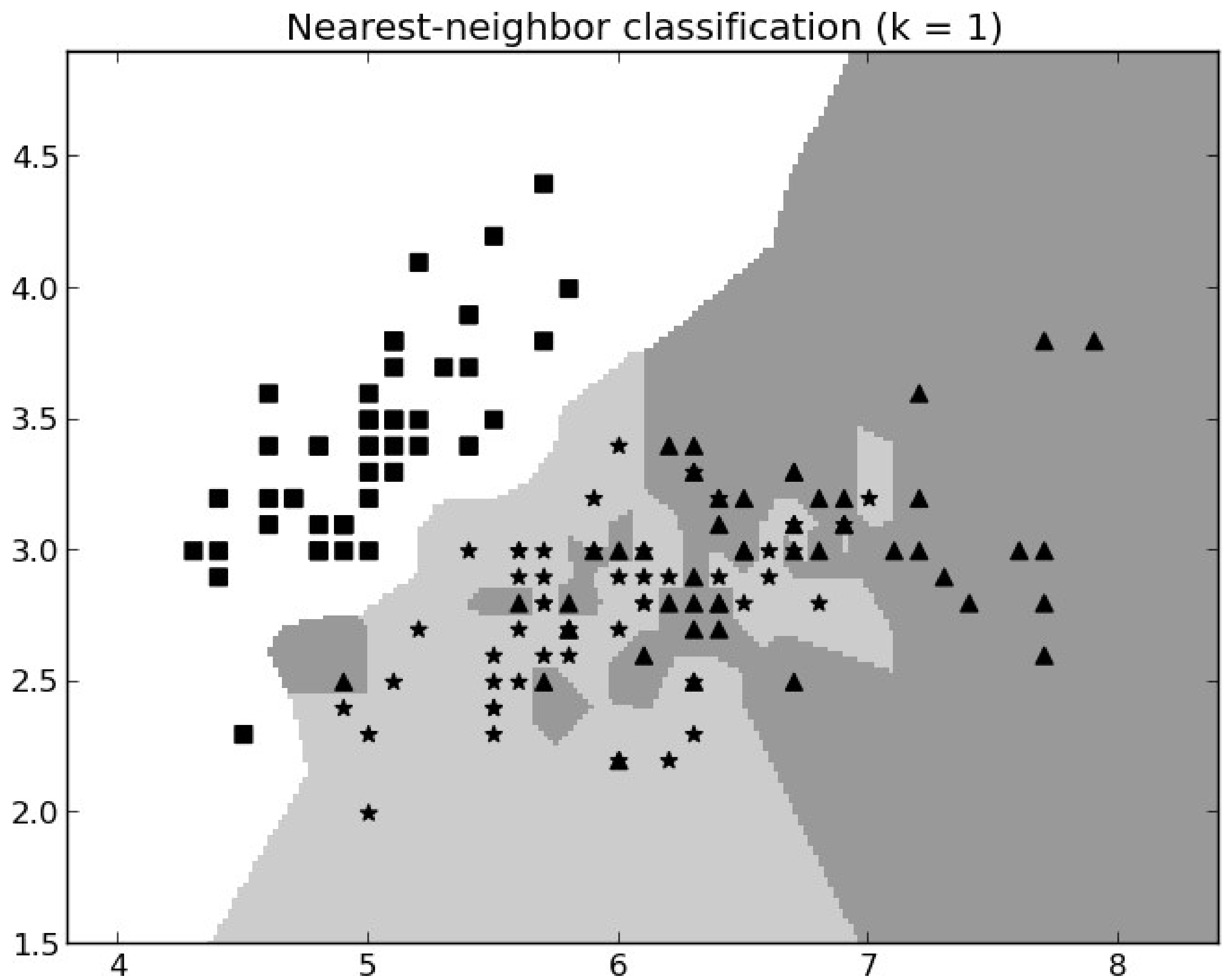
- **$\boldsymbol{k}$ Nearest Neighbors**

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

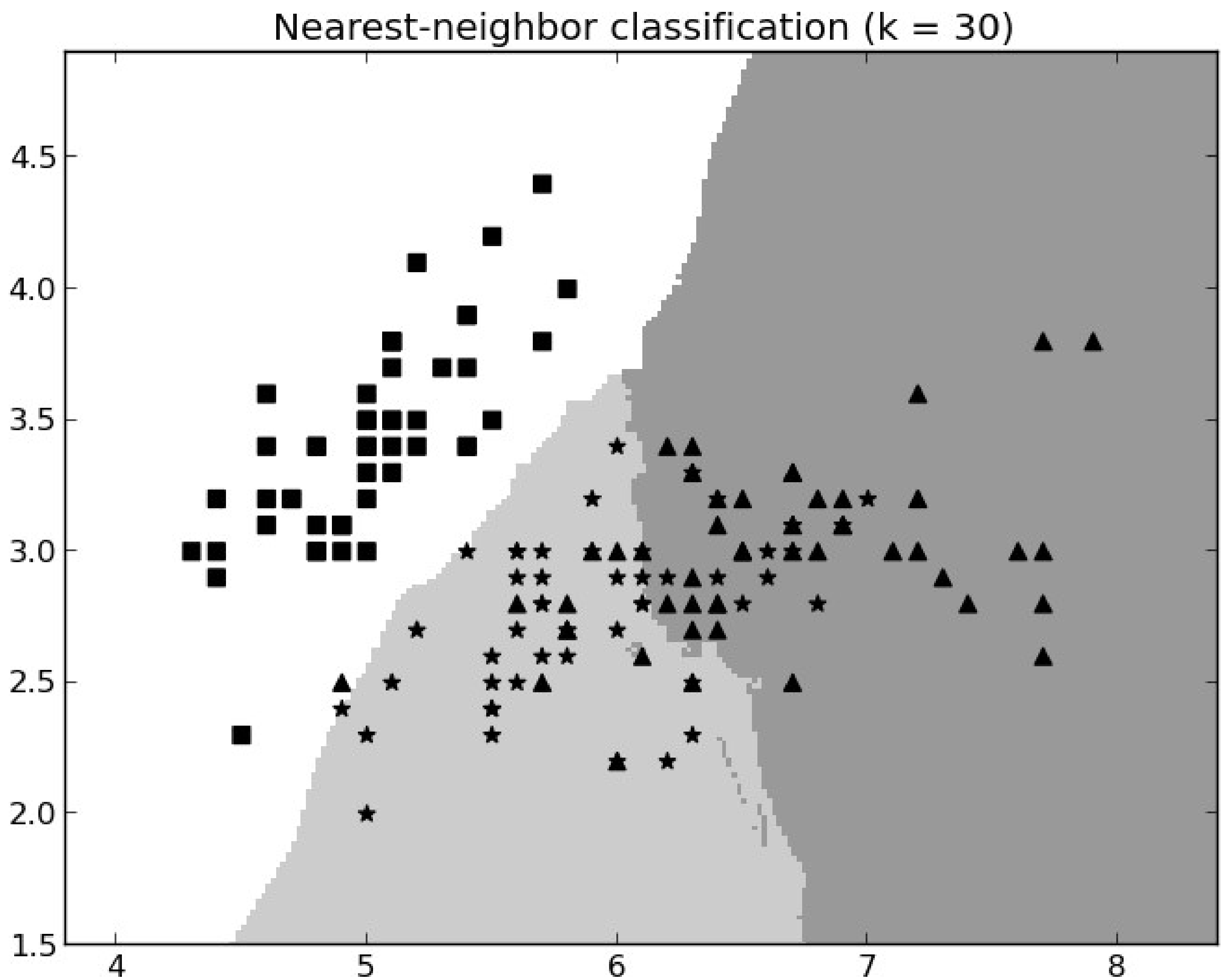
Geometric Interpretation, Over-fitting, and Complexity



1-Nearest Neighbor



30-Nearest Neighbors



Effect of k

- Smaller k: smaller training error.
- Larger k: higher training error, but predictions are more “stable” due to voting.

OCR digits classification:

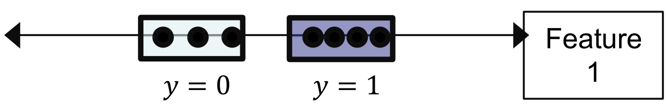
k	1	3	5	7	9
Test error rate	3.09%	2.95%	3.12%	3.06%	3.41%

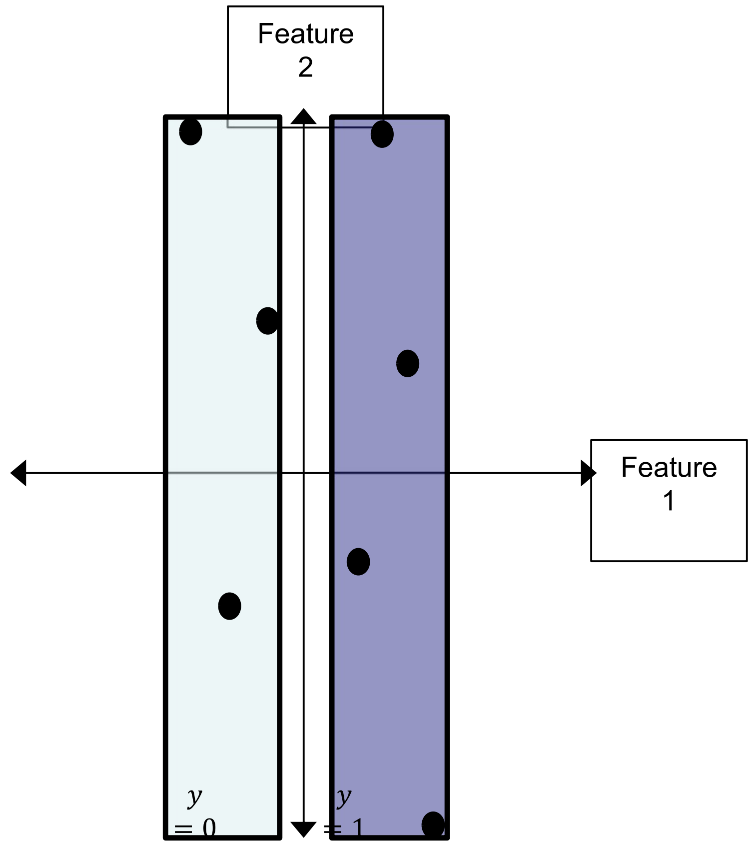
Picking k

- **Simplest approach: use a *hold-out set***:
  1. Pick a subset $V \subset S$ (*hold-out set, or validation set*).
  2. For each $k \in \{1,3,5,\dots\}$:
     - Construct k-NN classifier $\hat{f}_{S\setminus V,k}$ using $S \setminus V$
     - Compute error rate of $\hat{f}_{S\setminus V,k}$ on V ("hold-out error rate")
  3. Pick the k that gives the smallest hold-out error rate.

Noisy features

Caution: nearest neighbors can be broken by noisy features!





Issues with Nearest-Neighbor Models

- Dimensionality and domain knowledge
- Computational efficiency