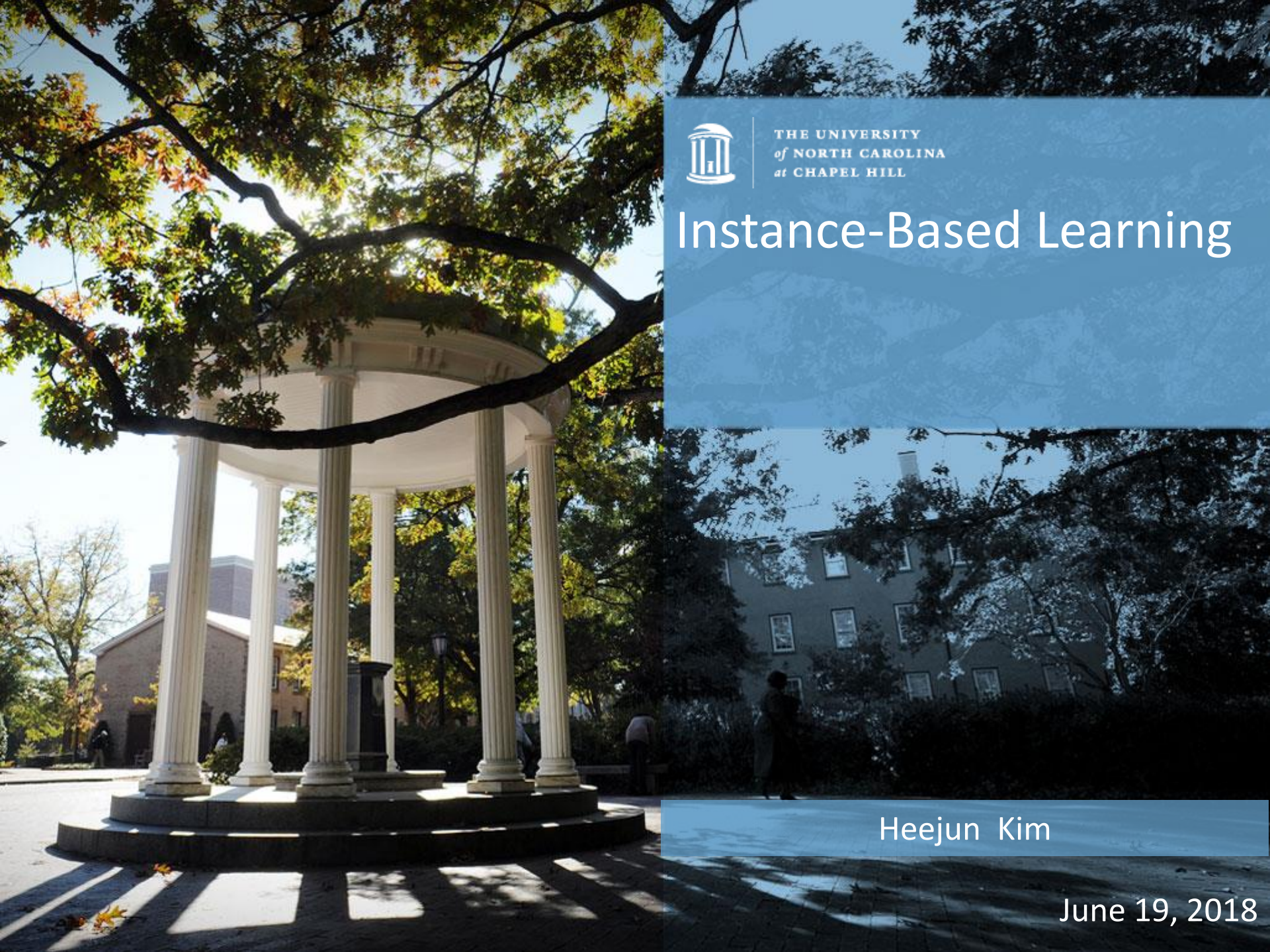


THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Instance-Based Learning

Heejun Kim

June 19, 2018

Instance-Based Classification

Motivation

training data

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10	sentiment
1	0	1	0	1	0	0	1	1	0	positive
0	1	0	1	1	0	1	1	0	0	negative
0	1	0	1	1	0	1	0	0	0	negative
0	0	1	0	1	1	0	1	1	1	positive
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	0	1	1	0	0	1	0	1	positive

test instance

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10	sentiment
1	0	1	0	1	0	0	1	1	0	?

Instance-Based Classification

Motivation

training data

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10	sentiment
1	0	1	0	1	0	0	1	1	0	positive
0	1	0	1	1	0	1	1	0	0	negative
0	1	0	1	1	0	1	0	0	0	negative
0	0	1	0	1	1	0	1	1	1	positive
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	0	1	1	0	0	1	0	1	positive

test instance

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10	sentiment
1	0	1	0	1	0	0	1	1	0	?

Instance-Based Classification

Motivation

training data

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10	sentiment
1	0	1	0	1	0	0	1	1	0	positive
0	1	0	1	1	0	1	1	0	0	negative
0	1	0	1	1	0	1	0	0	0	negative
0	0	1	0	1	1	0	1	1	1	positive
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	0	1	1	0	0	1	0	1	positive

test instance

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10	sentiment
1	0	1	0	1	0	0	1	1	0	positive

Instance-Based Classification

Motivation

training data

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10	sentiment
1	0	1	0	1	0	0	1	1	0	positive
0	1	0	1	1	0	1	1	0	0	negative
0	1	0	1	1	0	1	0	0	0	negative
0	0	1	0	1	1	0	1	1	1	positive
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	0	1	1	0	0	1	0	1	positive

test instance

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10	sentiment
1	0	1	0	1	0	0	1	0	0	?

Instance-Based Classification

Motivation

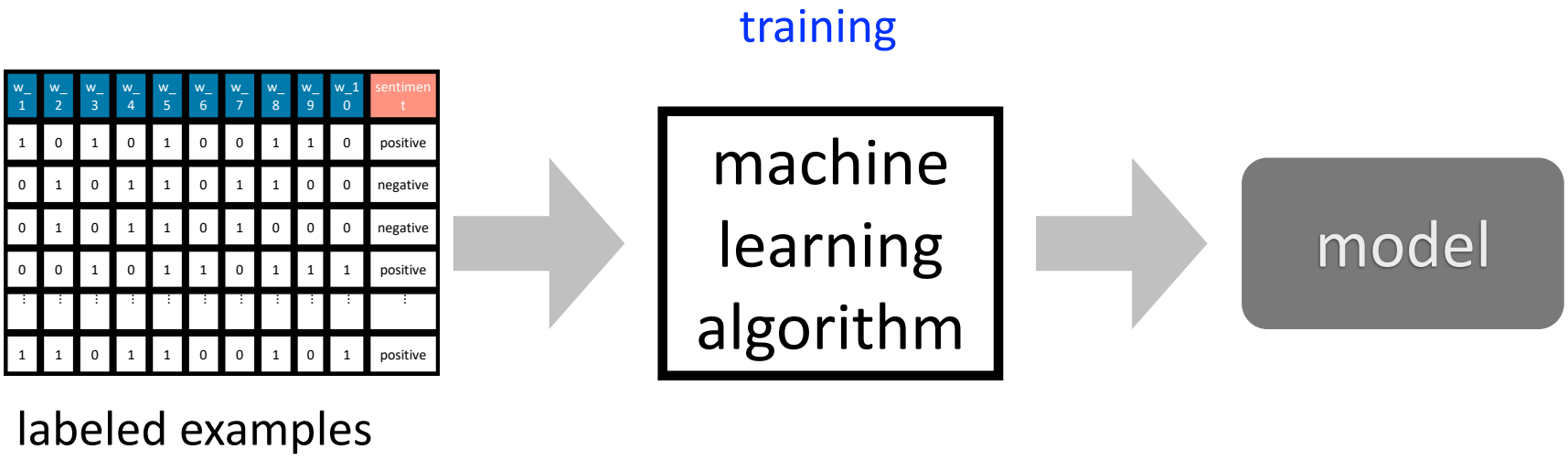
training data

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10	sentiment
1	0	1	0	1	0	0	1	1	0	positive
0	1	0	1	1	0	1	1	0	0	negative
0	1	0	1	1	0	1	0	0	0	negative
0	0	1	0	1	1	0	1	1	1	positive
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	0	1	1	0	0	1	0	1	positive

test instance

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10	sentiment
1	0	1	0	1	0	0	1	0	0	positive

Typical Supervised Classification

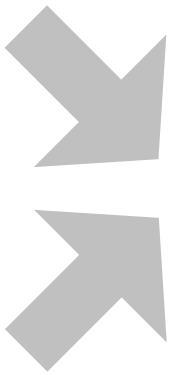


Instance-based Classification

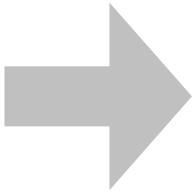
w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10	sentiment
1	0	1	0	1	0	0	1	1	0	positive
0	1	0	1	1	0	1	1	0	0	negative
0	1	0	1	1	0	1	0	0	0	negative
0	0	1	0	1	1	0	1	1	1	positive
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	0	1	1	0	0	1	0	1	positive

labeled examples

testing



instance-based algorithm



w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10	sentiment
1	0	1	0	1	0	0	1	1	0	positive

prediction

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10	sentiment
1	0	1	0	1	0	0	1	1	0	???

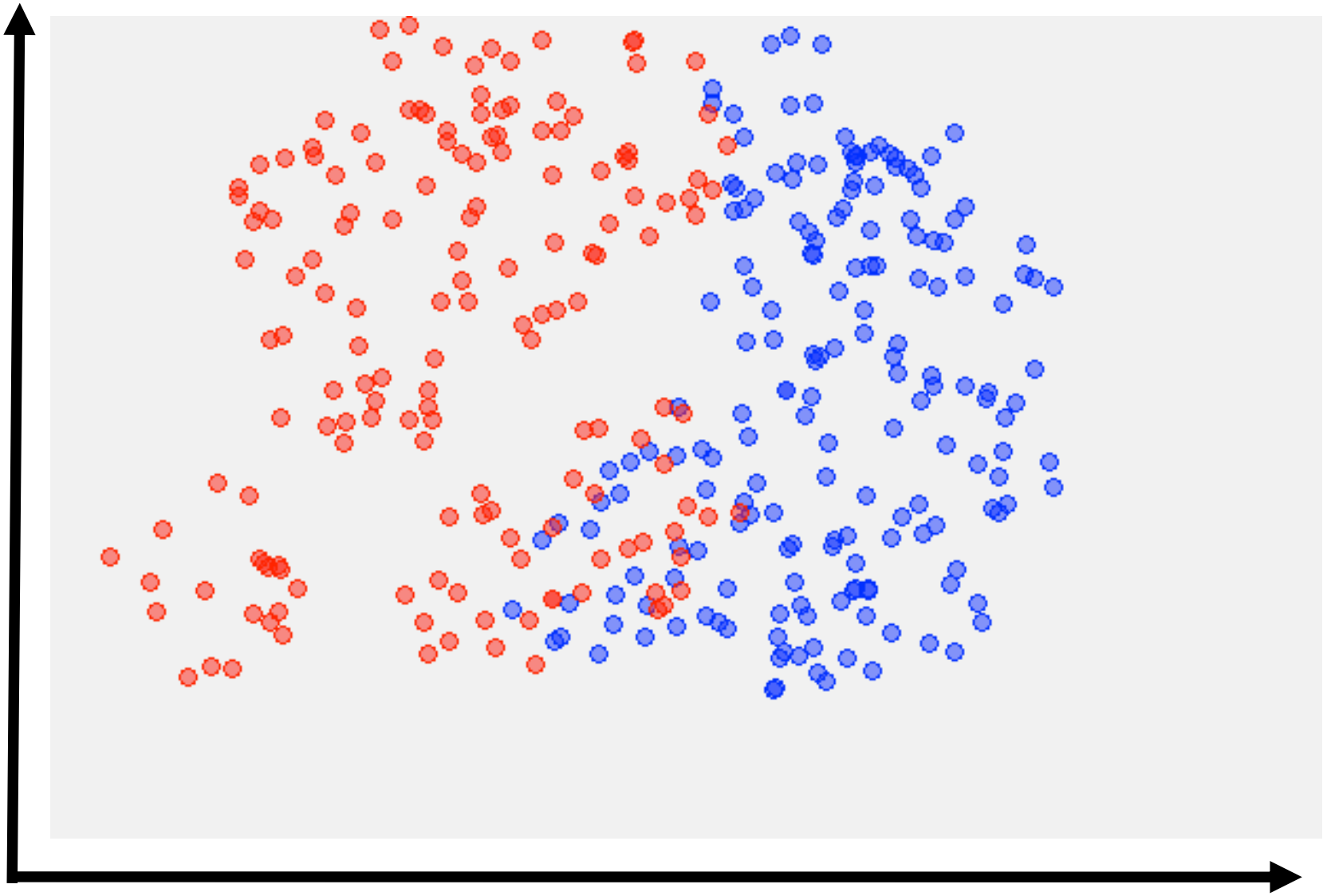
new, unlabeled example

No explicit generalization!!

Instance-based Classification

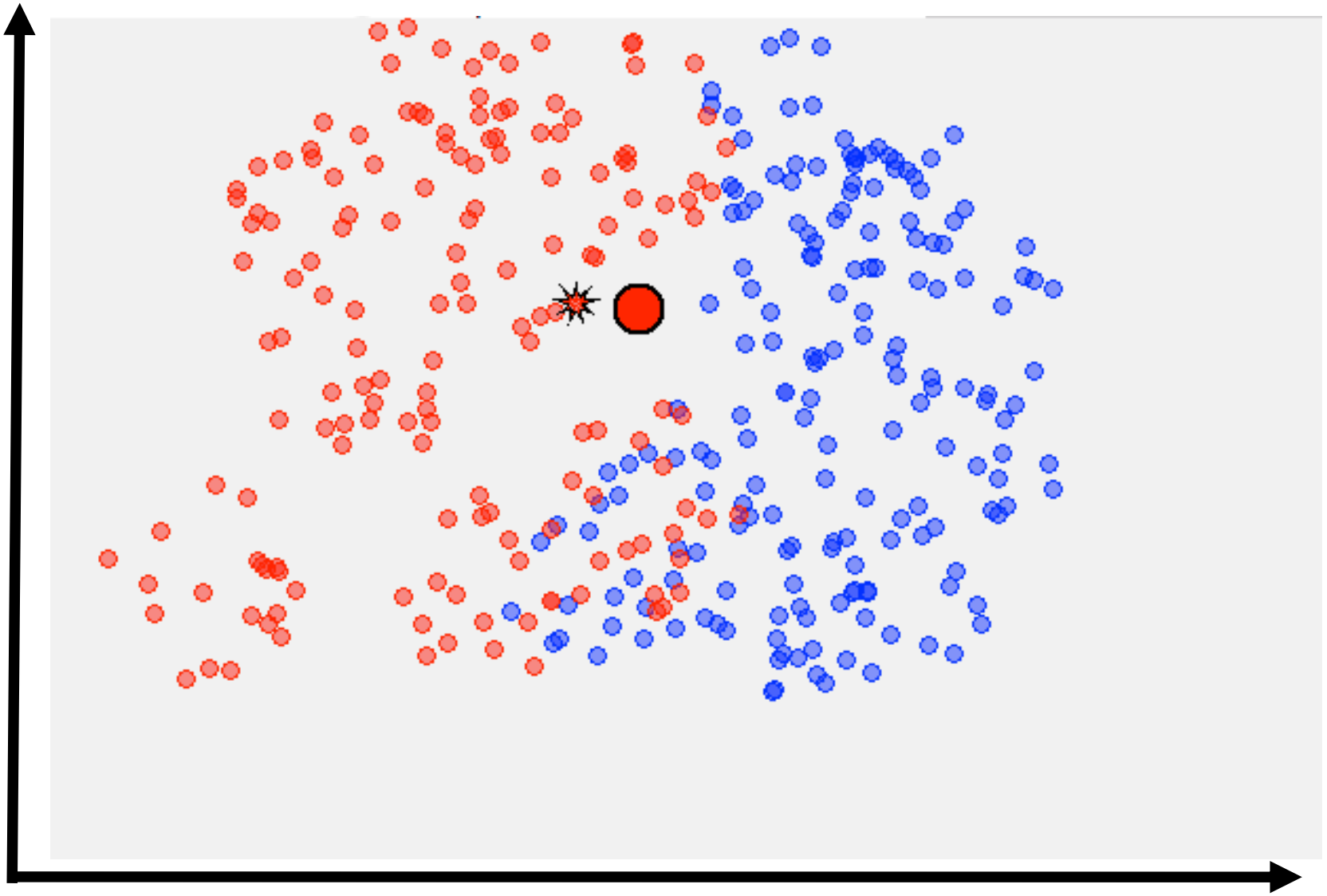
- Assumption: instances with similar feature values should have the same target label
- Necessary Ingredients:
  - a similarity/distance metric: a measure of similarity between instances
  - an averaging technique: a way of combining the labels from the most similar training instances

Nearest-Neighbor Classification

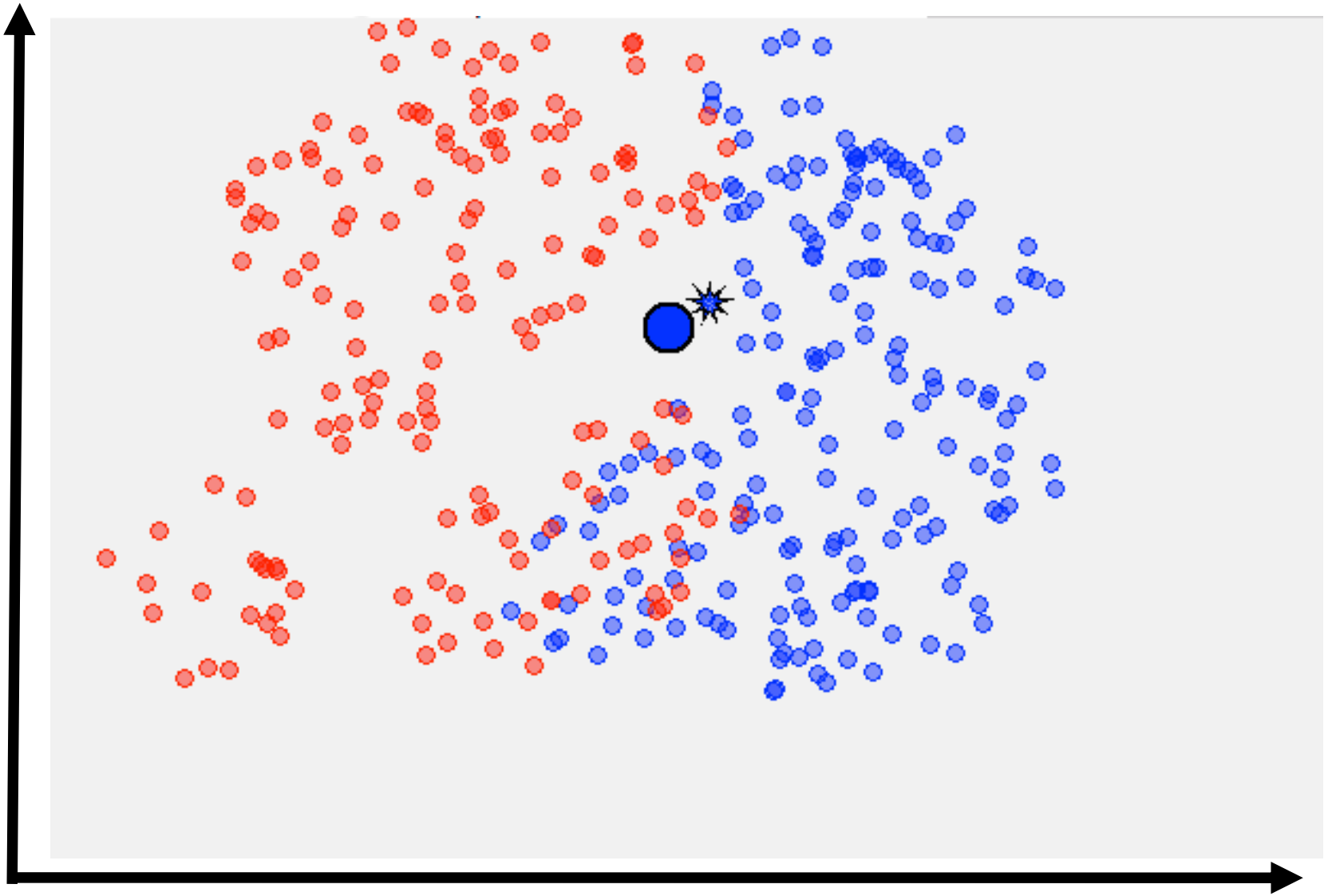


source: http://www.math.le.ac.uk/people/ag153/homepage/KNN/

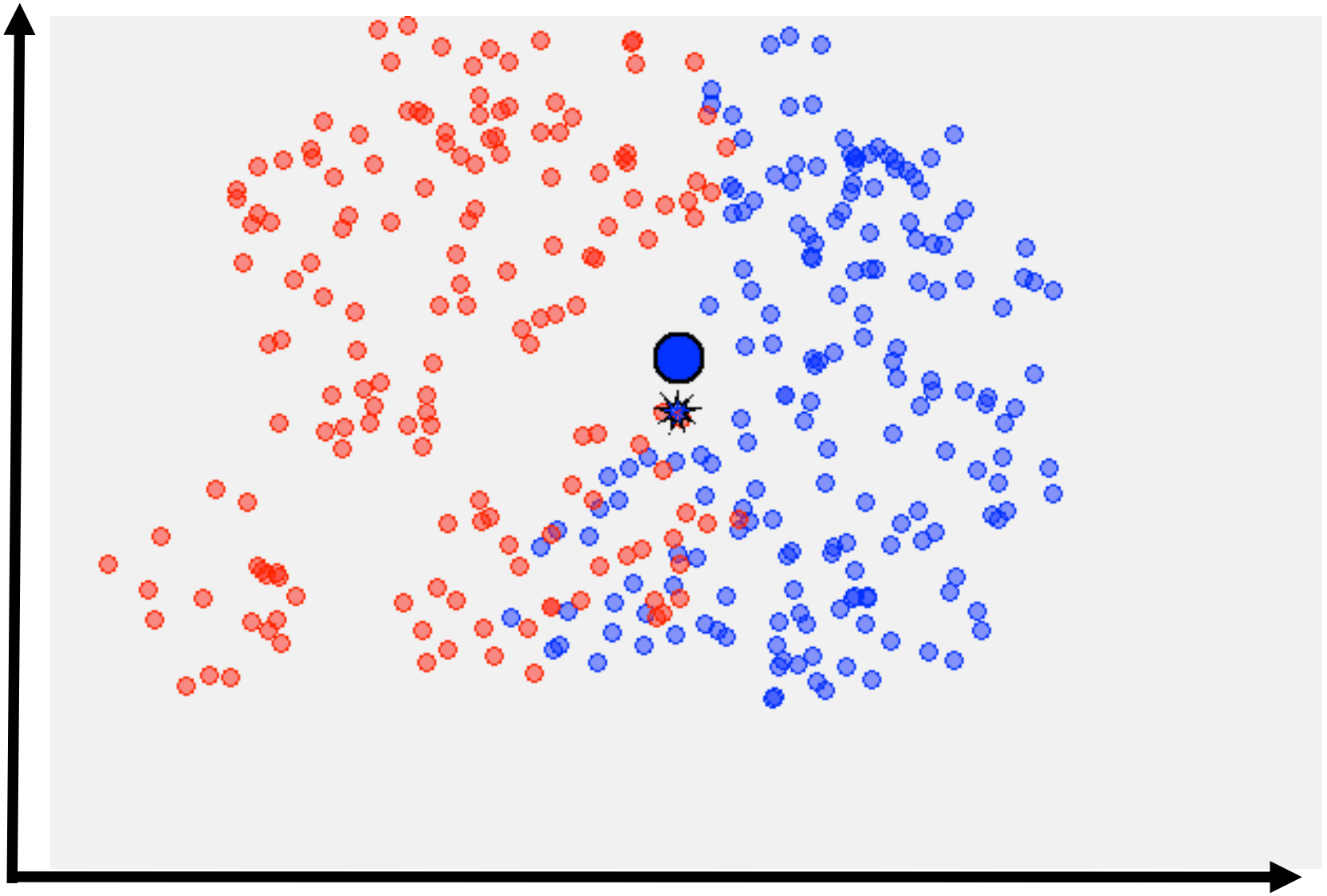
Nearest-Neighbor Classification



Nearest-Neighbor Classification



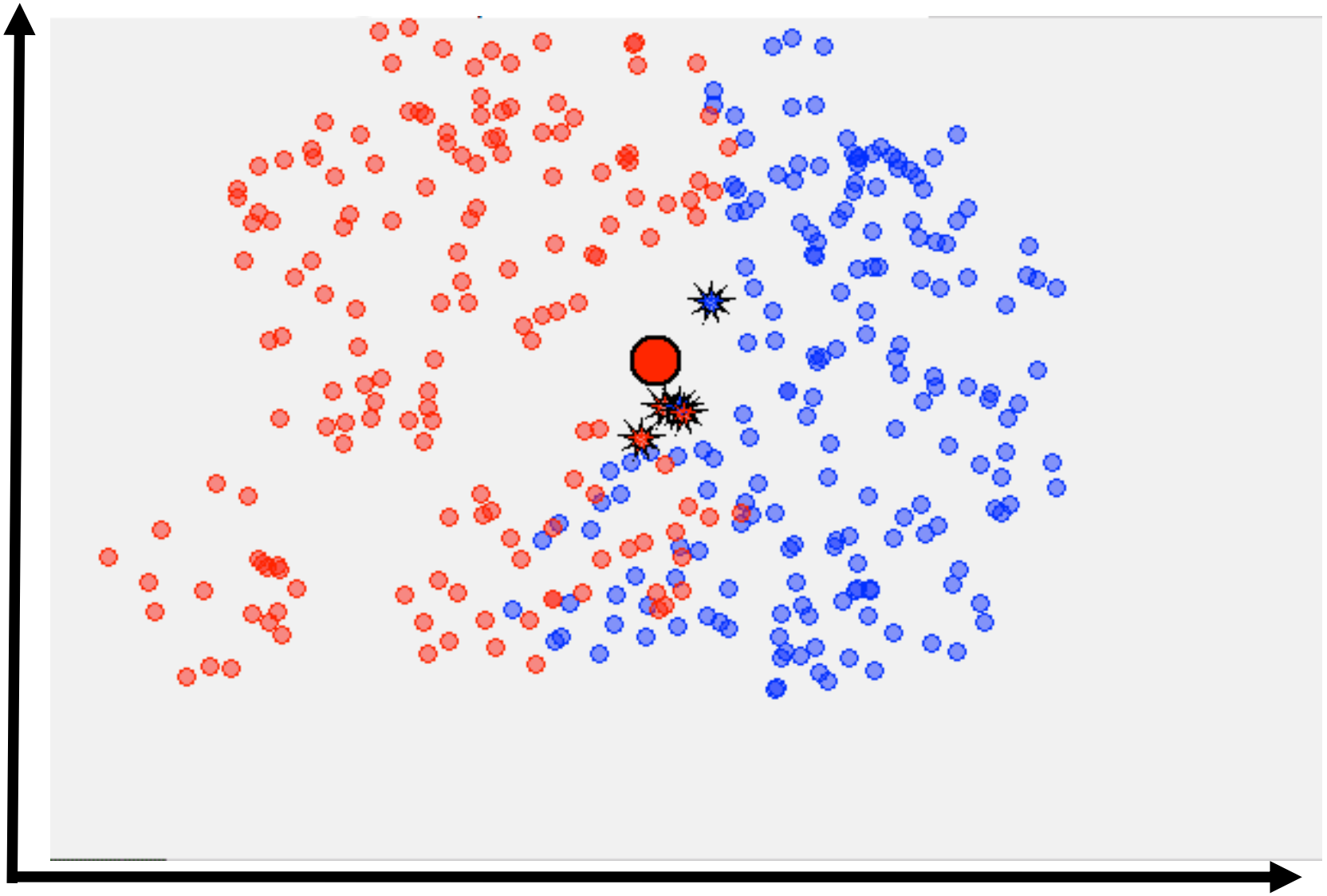
Nearest-Neighbor Classification



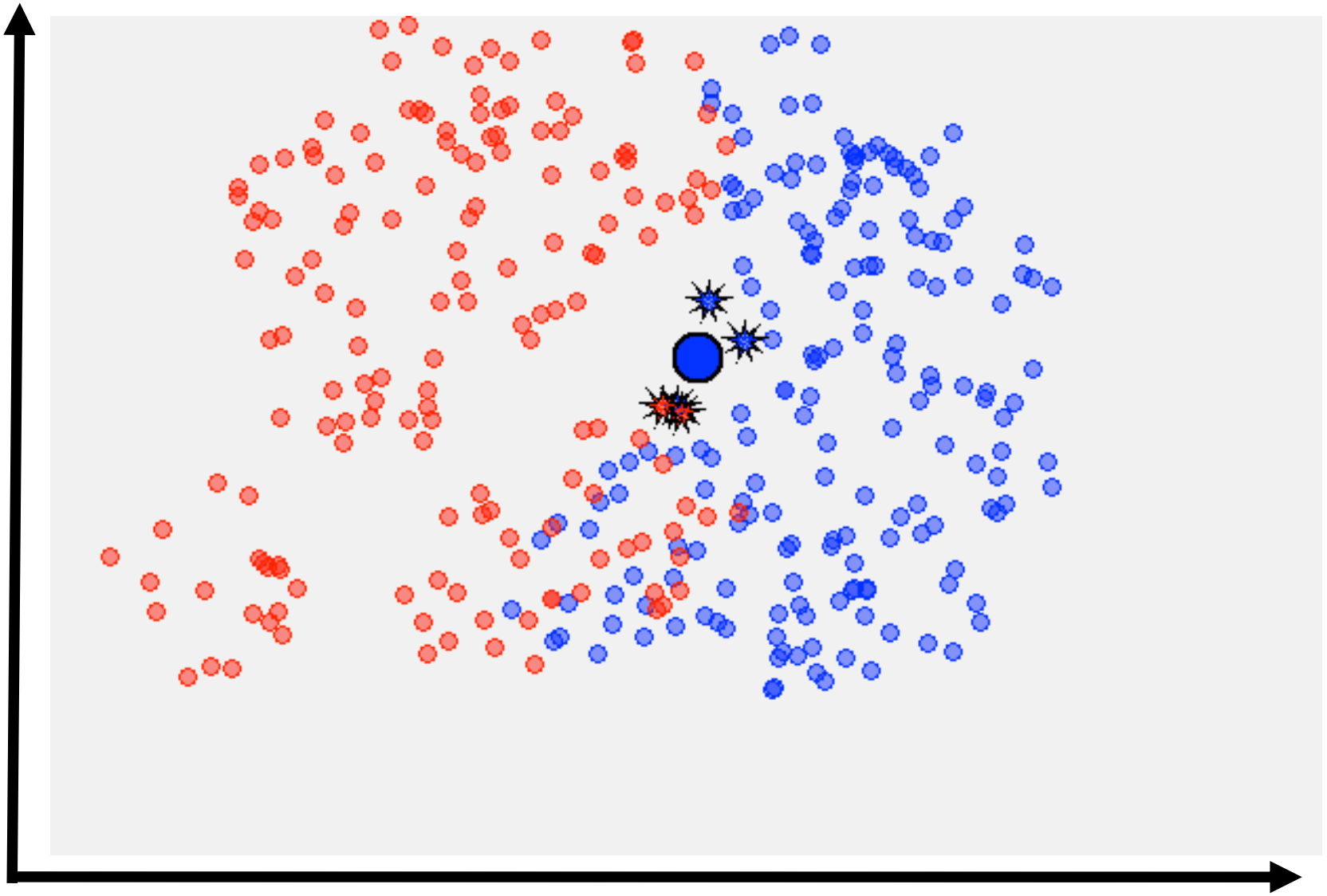
Nearest-Neighbor Classification

- Given a test instance, assign the label associated with the nearest training set instance
- What are a potential limitation of this approach?
- The nearest neighbor may be an outlier
- For example: a positive movie review with lots of negative words
- Solution: use the majority class associated with the K nearest neighbors

K Nearest-Neighbor (KNN)
(K = 5)



K Nearest-Neighbor (KNN)
(K = 5)



K Nearest-Neighbor Classification

- Given a test instance, assign the majority label associated with the K nearest training set instances
- What is a potential limitation of this approach?
- Nearest-neighbors that are far away have the same influence as nearest-neighbors that are close
- Solution: use some kind of weighted voting
- There are many, many variants
- Including one that does weighted voting using the entire training set

K Nearest-Neighbor (KNN)

practical matters

- Feature normalization
- Feature weighting
- Computational complexity

K Nearest-Neighbor (KNN)

practical matters: feature normalization

- KNN assumes that feature values (and differences in feature value) are comparable between features
- For example, TF.IDF term-weighting places more emphasis on rare terms over the documents
- In some cases, we want the opposite
- That is, we want features to be treated equally
- This can be tricky if feature values are not comparable

K Nearest-Neighbor (KNN)

practical matters: feature normalization

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10	sentiment
10.5	1.2	100.4	4.54	33.4	503.4	76.8	0.54	2.31	145.6	positive
13.5	1.5	101.4	5.65	34.5	400.3	79.7	0.36	5.35	353.3	negative
20.4	1.6	143.5	7.47	24.5	323.2	74.3	0.75	10.54	550.5	negative
12.4	1.4	164.2	5.76	65.6	543.2	43.4	0.23	1.65	365.2	positive
12.5	3.2	156.4	4.54	67.5	234.5	45.3	0.54	1.67	543.2	negative
15.7	1.8	154.6	8.67	65.7	156.5	55.5	0.45	5.64	300.4	positive

- Features that capture different types of evidence may have very different ranges
- What can we do so that they have roughly equal contribution?

K Nearest-Neighbor (KNN)

min/max normalization

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10	sentiment
0	0	0	0	0.21	0.9	0.92	0.6	0.07	0	positive
0.3	0.15	0.02	0.27	0.23	0.63	1	0.25	0.42	0.51	negative
1	0.2	0.68	0.71	0	0.43	0.85	1	1	1	negative
0.19	0.1	1	0.3	0.96	1	0	0	0	0.54	positive
0.2	1	0.88	0	1	0.2	0.05	0.6	0	0.98	negative
0.53	0.3	0.85	1	0.96	0	0.33	0.42	0.45	0.38	positive

$$w_{i,j}^{\text{norm}} = \frac{w_{i,j} - \min(w_{i,*})}{\max(w_{i,*}) - \min(w_{i,*})}$$

K Nearest-Neighbor (KNN)

practical matters: feature weighting

- In some cases, some features are more important than others
- TF.IDF assumption: the important features are the rare ones over the documents
  - A feature that distinguishes between instances will also distinguish between the target class values
- Alternative: learn feature weights from the training data

K Nearest-Neighbor (KNN)

practical matters: feature weighting

- Weighted Euclidean Distance:

$$D(x,y)\sqrt{\left(\sum_{i=1}^{|\mathcal{V}|} w_i(x_i - y_i)^2\right)}$$

K Nearest-Neighbor (KNN)

practical matters: feature weighting

- Split the training set into two sets
- Make predictions on the second set using the first set
- For each second-set instance that is misclassified based on its first-set nearest neighbor:
  - Find the features where the instances are the most similar
  - Increase their weights (i.e. accentuate their differences)

K Nearest-Neighbor (KNN)

practical matters: making predictions

- How fast/slow is KNN is making predictions?
- KNN can be very slow
- It needs to compute the similarity/distance between the test instance and every training instance
- Is there anything we can do to speed the process?

K Nearest-Neighbor (KNN)

kD-tree example

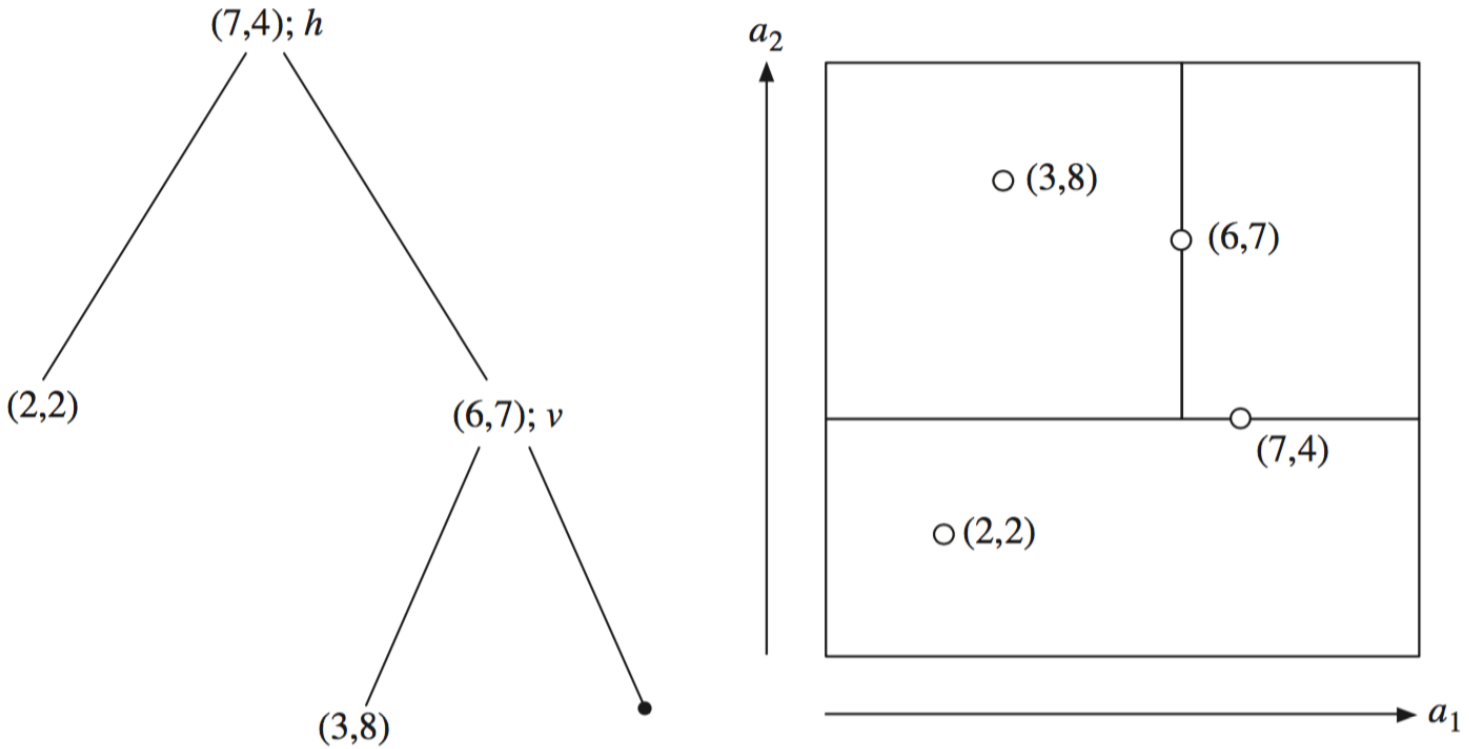
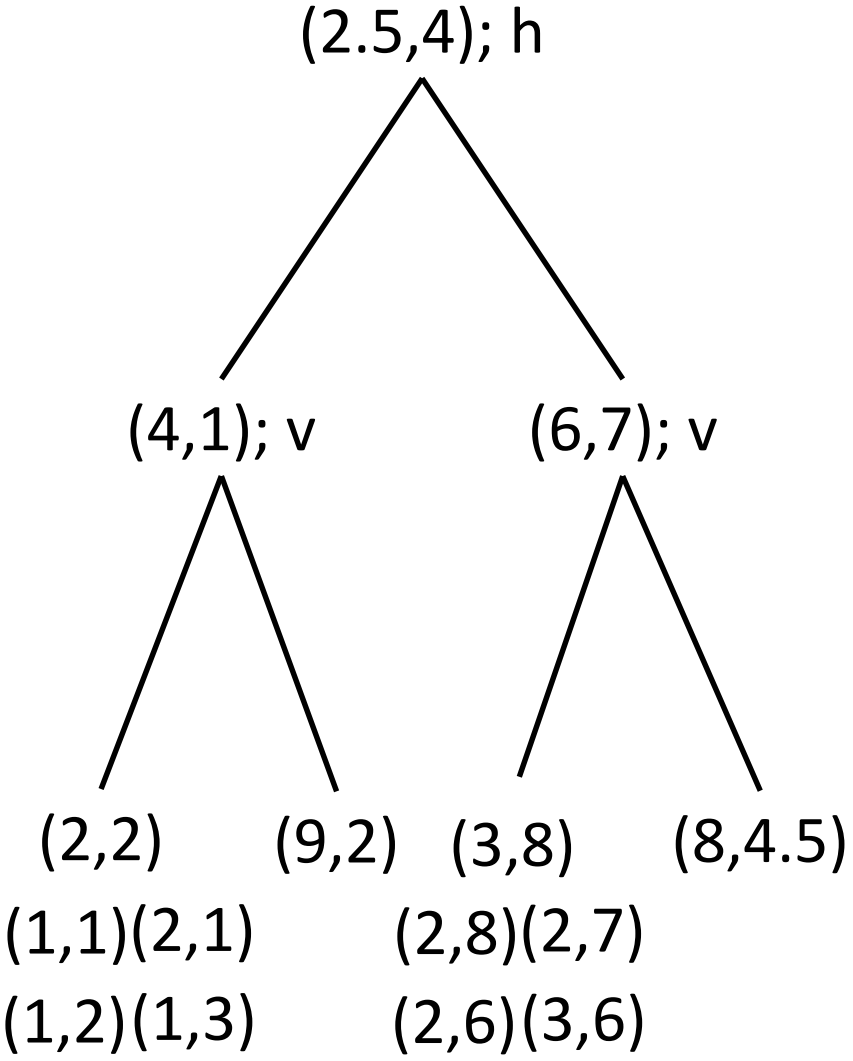


Figure 4.13 from Witten et al., 2017 (Data Mining: Practical Machine Learning Tools and Techniques (Fourth Edition)

K Nearest-Neighbor (KNN)

kD-tree example



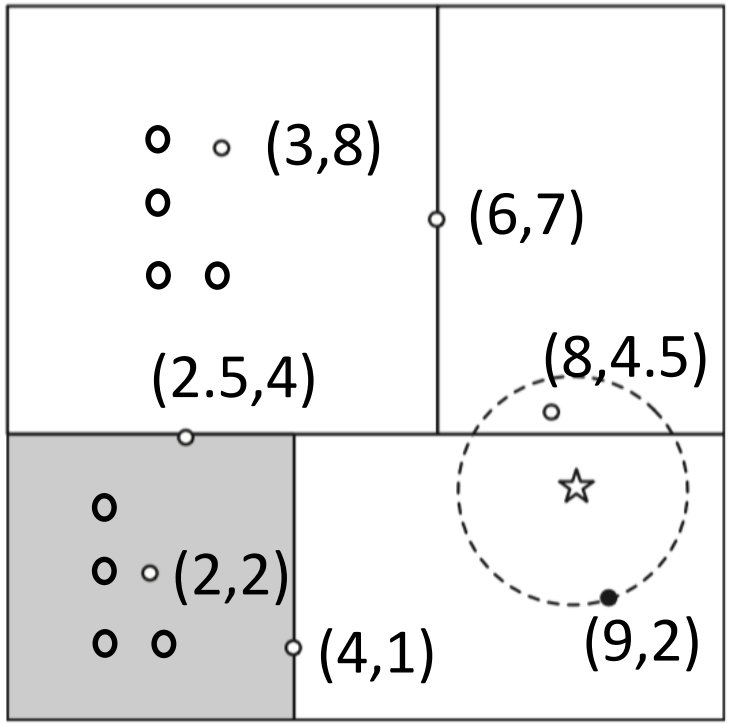
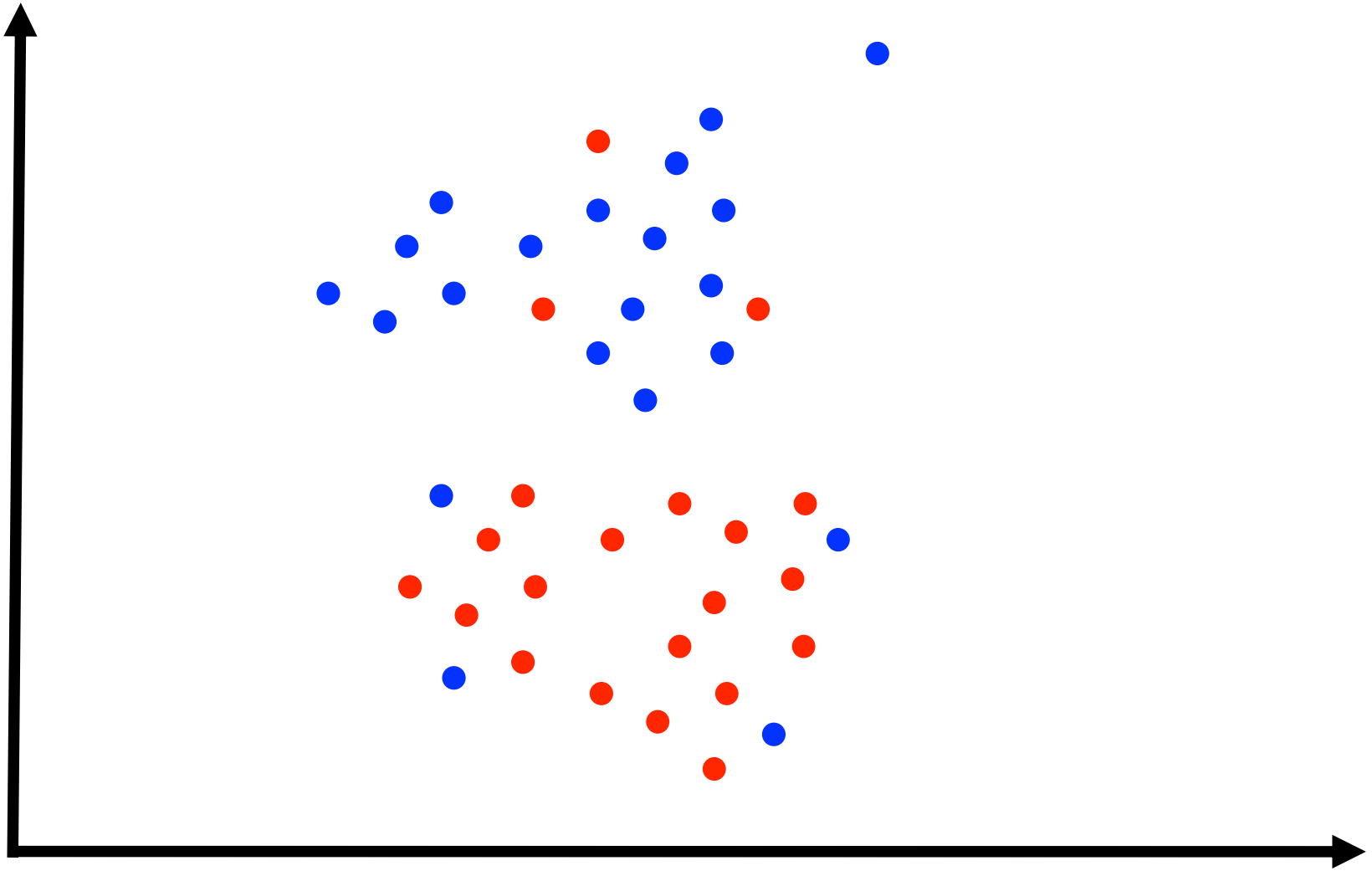


Figure 4.14 from Witten et al., 2017

K Nearest-Neighbor (KNN)

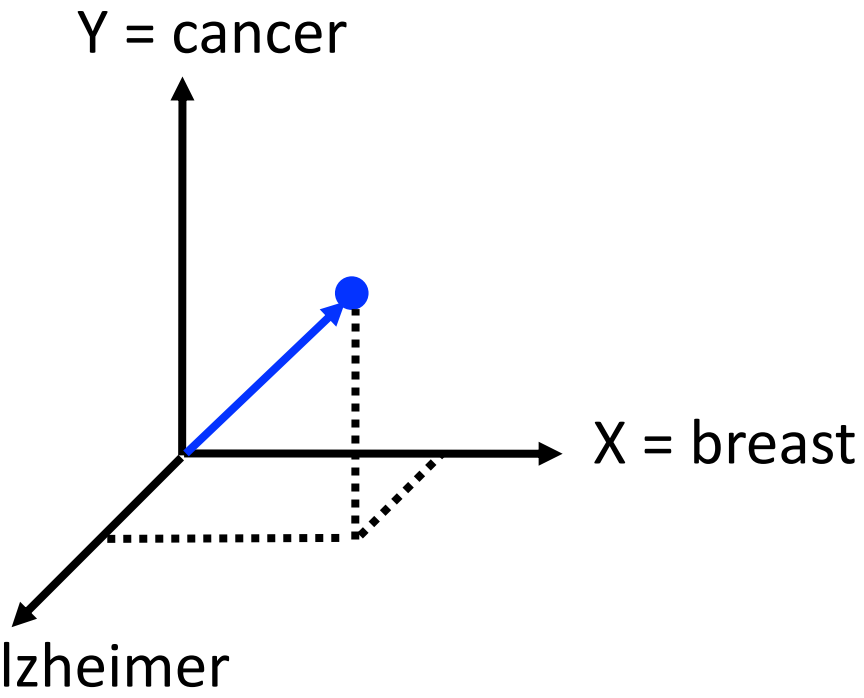
practical matters: making predictions



Independence Assumption

- The basis vectors (X, Y, Z) are linearly independent because knowing a vector's value on one dimension doesn't say anything about its value along another dimension

does this hold true for natural language text?



basis vectors for 3-dimensional space

Mutual Information

IMDB Corpus

- If this were true, what would these mutual information values be?

w1	w2	MI	w1	w2	MI
francisco	san	?	dollars	million	?
angeles	los	?	brooke	rick	?
prime	minister	?	teach	lesson	?
united	states	?	canada	canadian	?
9	11	?	un	ma	?
winning	award	?	nicole	roman	?
brooke	taylor	?	china	chinese	?
con	un	?	japan	japanese	?
un	la	?	belle	roman	?
belle	nicole	?	border	mexican	?

Mutual Information

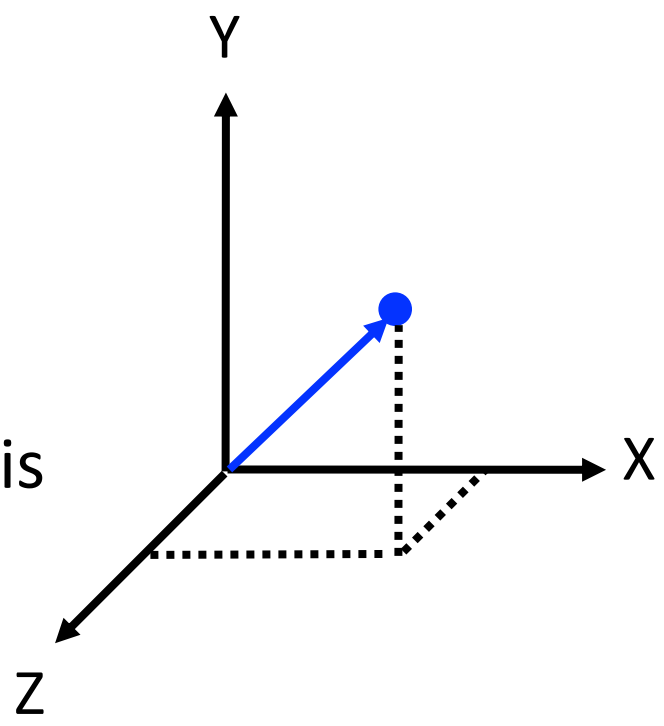
IMDB Corpus

- These mutual information values should be zero!

w1	w2	MI	w1	w2	MI
francisco	san	6.619	dollars	million	5.437
angeles	los	6.282	brooke	rick	5.405
prime	minister	5.976	teach	lesson	5.370
united	states	5.765	canada	canadian	5.338
9	11	5.639	un	ma	5.334
winning	award	5.597	nicole	roman	5.255
brooke	taylor	5.518	china	chinese	5.231
con	un	5.514	japan	japanese	5.204
un	la	5.512	belle	roman	5.202
belle	nicole	5.508	border	mexican	5.186

Independence Assumption

- Representing texts as vectors assumes that terms are independent
- The fact that one occurs says nothing about another one occurring
- This is viewed as a limitation
- However, the implications of this limitation are still debated
- A very popular solution



Summary

- Instance-based classification relies on one assumption:
  - similar instances should have the same label
- Ingredients:
  - similarity metric: to find the nearest neighbors
  - averaging technique: to combine their true labels into a final prediction
- K-NN: use the geometric distance to find the K nearest neighbors and take the majority label

Any Questions?

Evaluation

Next Class