Classification

- Assume that the following table is the bag of words representation of 5 documents
- Normalise the document vectors
- Classify d5 using rocchio and 1-NN (with cosine similarity)
- Do they give the same result?

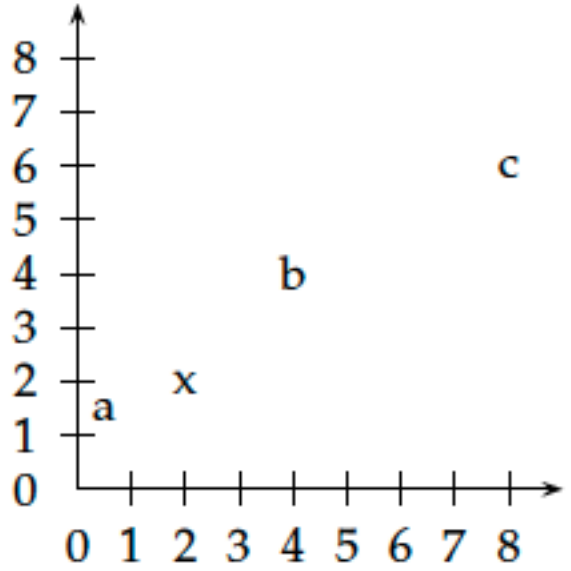
	Chinese	Japan	Tokyo	Macao	Beijing	Shanghai	class
d1	1	0	0	1	1	1	1
d2	1	1	0	0	0	1	1
d3	1	0	1	1	1	1	1
d4	0	1	1	1	1	0	0
d5	1	1	1	0	0	0	?

Exercise 14.2, 14.3

- 14.2: Construct a training set with documents in two classes where Rocchio classification assigns to one of the training documents the wrong class.
- 14.3: Prove that the number of linear separators of two classes is either infinite or zero.
  - Hint: write the hyperplane equation which states that for all the examples in class + the dot product of the vector orthogonal to the hyperplane and the document is larger than a threshold w_0

Exercise 14.6

- 14.6: In Figure 14.14, which of the three vectors a, b, and c is (i) most similar to x according to dot product similarity, (ii) most similar to x according to cosine similarity, (iii) closest to x according to Euclidean distance?



► **Figure 14.14** Example for differences between Euclidean distance, dot product similarity and cosine similarity. The vectors are $\vec{a} = (0.5 \;\; 1.5)^T$, $\vec{x} = (2 \;\; 2)^T$, $\vec{b} = (4 \;\; 4)^T$, and $\vec{c} = (8 \;\; 6)^T$.

Bayes error rate

- Examples in class 1 are uniformly distributed in the square $\{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$
- Examples in class 2 are uniformly distributed in the square $\{(x, y) : 1/2 \leq x \leq 3/2, 0 \leq y \leq 1\}$
- Assume that there is an equal number of examples in class 1 and 2
- What is the best class separator in this case?
- What is its Bayes error rate, i.e., the probability to make a misclassification using the best class separator?