

Classification

- Assume that the following table is the bag of word representation of 5 documents
- Normalise the document vectors
- Classify d5 using rocchio and 1-NN
- 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, 14.6

- 14.2: Show that Rocchio classification can assign a label to a document that is different from its training set label.
- 14.3: Prove that the number of linear separators of two classes is either infinite or zero.
- 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?