

4. Text classification (30 points)

(a) kNN

Consider the following supervised corpus of news headlines, where the first word is the document class

WorldNews	Iraq election
WorldNews	French executive injured
Business	Chief executive smiles
Business	Krispy Kreme executive resigns

(i) Consider now assigning a class to the following document using 3NN classification:

executive suite

What class is this document assigned to? Assume raw term frequency, no idf, and cosine similarity. Show the similarity calculations that justify your answer.

(ii) Would the same result be guaranteed using 1NN classification? Why or why not?

(b) Naïve Bayes

(i) We observed that Naïve Bayes classifiers, by their independence assumptions, can “double count” evidence. Show with a complete numerical example how this double counting can lead to the wrong classification decision. Base your example around a text collection which contains the name Mariah Carey a number of times, but where the individual terms Mariah and Carey never occur except together.

(ii) Conversely, we discussed upweighting zones in a document to improve classification performance. Is this a form of double counting of evidence? Explain why this can be a useful thing to do (perhaps with an example).

(c) SVM kernels

For the XOR problem over two features, a document is in class 1 if either feature is present, but not if both or neither are (e.g., you are interested in documents that mention linguistic(s) or relativity but not ones that discuss linguistic relativity or neither term). Illustrate explicitly a kernel function over binary incidence vectors for these two features that will allow this XOR decision to be made by a linear classifier. Give the mapping of the four possible vectors, and the decision rule, via specifying a decision boundary.

5. (Multinomial classifiers 30 points) Our task is to classify words as English or not English. Words are generated by a source with the following distribution:

event	word	English?	Probability
1	ozb	0	4/9
2	uzu	0	4/9
3	zoo	1	1/18
4	bun	1	1/18

(a) Compute the parameters (priors and conditionals) of a naïve bayes multinomial classifier that uses the letters b, n, o, u, and z as features. Assume a training set that reflects the probability distribution of the source perfectly. Make the same independence assumptions that are usually made for a multinomial classifier that uses words as features for text classification. Compute parameters using smoothing, in which computed-zero probabilities are smoothed into probability 0.01, and computed-nonzero probabilities are untouched. (This simplistic smoothing may cause $P(A) + P(-A) > 1$, which can be corrected if we correspondingly smooth all complementary probability-1 values into probability 0.99. For this problem, solutions may omit this correction to simplify arithmetic.)

(b) How does the classifier classify the word “zoo”?

(c) Classify the word “zoo” using a multinomial classifier as in part (a), but do not make the assumption of positional independence. That is, estimate separate parameters for each position in a word. You only need to compute the parameters you need for classifying “zoo”.