1. Consider the set of 6 documents (5 for training and 1 for testing).

|  |  |  |  |
| --- | --- | --- | --- |
| **docID** | | **words in document** | **in class = cricket ?** |
| 1 | | *sachin dravid sachin* | Yes |
| 2 | | *dravid federer schools* | No |
| 3 | | *sachin sachin neville* | Yes |
| 4 | | *sachin drive* | Yes |
| 5 | | *saina sushil sachin* | No |
| 6 (test set) |  | *sachin dravid neville federer* | ? |
|  | | | |

Using the TF-IDF weight formula *wt,d* = (1+ log10*tft,d*)log10(*N*/*dft*), where *N* is the no. of documents in the collection, determine whether the Rocchio classification will assign document 6 to sports? Why?

1. We wish to build an SVM classifier that categorizes a point into Class A or B based on 3 data points: (1, 1) {Class A), {2, O) (Class A) and (2, 3) (Class B).

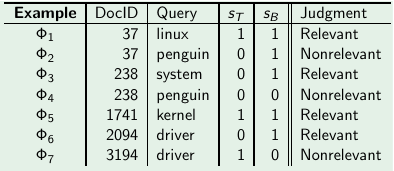
(a) Find the optimal separating hyperplane w.

(b) Find the margin p.

(c) Plot the points on a small graph and depict the separator geometrically. Mark points, lines and distances clearly.

1. Below is a training example where sT and sB represents whether the query word exits in title and body respectively. Assign weights to sT and sB to calculate the score of document-query pair such that the error is minimized. Error is defined as : *ϵ*(*g*,Фj ) = (*r* (*dj*, *qj* ) − score(*dj* , *qj* ))2

Consider r(*dj*, *qj*) as 1 for Relevant and 0 for Nonrelevant.



1. Show that Rocchio classification can assign a label to a document that is different from its training set label.