

CS276 Problem Set #2

5 questions, 10 points each.

Assigned: Tuesday, May 10th 2011

Due: Thursday May 19th 2011

Delivery: Assignments must be submitted by 5 p.m. Pacific on the due date. Problem sets should be handed to TAs in class or left in the box outside of Professor Manning's office. All assignments not delivered in class must have the time and date of submission clearly marked on the front page. SCPD students: Fax to (650) 736-1266, (650) 725-4138, or email to scpd-distribution@lists.stanford.edu (cc to cs276-spr1011-staff@lists.stanford.edu)

Late policy: Refer to the course webpage.

Honor code: Please review the collaboration and honor code policy on the course webpage.

Also note: because some questions may be drawn from previous years' problem sets or exams, students are forbidden to consult solution sets from previous years unless we explicitly provide them.

1. Naive Bayes

The sales group of your new startup MadCorp has done some demographic studies and indicated that web users fall under two distinct groups: *losers* and *winners*. Using your Stanford contacts, your team has managed to get information of cookies of sites around the Internet (i.e., you know exactly which page a user visits from a tracked set of pages D). Furthermore, careful research has produced representative browsing histories for "typical" users in each class, below. Treat these browsing histories as a bag of URLs. These profiles were manually constructed by inspecting a set of five user histories. Your research team's task is to train a classifier that, for a new browsing history, will determine how likely the user is to be part of each group.

The typical browsing history for a *winner* (as aggregated from three histories) is:

Docs: D0 D1 D0 D4 D5 D6

The typical browsing history for a *loser* (as aggregated from two histories) is:

Docs: D3 D8 D1 D1 D7 D8 D6

(a) Estimate the following probabilities as used by a Multinomial Naive Bayes classifier over documents in the browsing history. Assume that the set of documents D is the vocabulary of "words" and that $|D| = 100$. Assume add-one smoothing.

- (i) $P(D0 | \text{winner})$
- (ii) $P(D0 | \text{loser})$
- (iii) $P(\text{winner})$
- (iv) $P(\text{loser})$

(b) Classify the following (unclassified) browser history according to its most likely class using the multinomial naive bayes classifier you started to train in (a):

Query history: D3 D0 D9

(c) The MadCorp sales team calls to say that too many irrelevant users are being inadvertently classified as *winners*. They mention that in fact only 2% of web users are actually *winners* on average. How can this information be integrated into the Naive Bayes classifier?

(d) Extra credit: Explain why a naive implementation of the Bernoulli naive Bayes model may be slower than Multinomial naive Bayes and suggest how it could be made equally fast. Next, confirm or refute the following interpretation: the Multinomial model accumulates exponentiated probabilities, while the Bernoulli model accumulates odds ratios.

2. Text Classification

(a) kNN

Consider the following supervised corpus of news headlines.

Class	Headline
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 that 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) **Text clustering with k-means**

Consider the problem of clustering the following documents using K-means with $K = 2$ and cosine similarity.

- Doc1: go Longhorns go
- Doc2: go Texas
- Doc3: Texas Longhorns
- Doc4: Longhorns Longhorns

Assume Doc1 and Doc3 are chosen as the initial seeds. Assume simple term-frequency weights (no IDF). Show all similarity calculations needed to cluster the documents, centroid computations for each iteration, and the final clustering.

3. Feature Selection

(a) Consider the following training set:

Class	Document	Context			
C1	Document 1	A	B	C	
	Document 2		B	C	
C2	Document 3	A	B		
	Document 4				D
	Document 5	A	B		D

- (i) Use mutual information to determine which letter(s) best discriminates class C2. (That is, find a letter (letters) that has the highest mutual-information value for class C2.)
- (ii) What is the chi-square value of the letter “A” as a discriminator for class C2?
- (iii) What is the chi-square value of the letter “B” as a discriminator for class C2?
- (iv) Using chi-square, determine which of “A” and “B” is a better discriminator for class C2.
- (v) Using chi-square, which letter (letters) is the best discriminator for class C2?

(b) What are the values of Mutual Information and Chi-square if term and class are completely independent? What if there are completely dependent?

4. Rocchio Classification

This question concerns the Rocchio Classification algorithm. Consider the following points in R_2 as partitioned into two classes, C1 and C2:

C1: (2, 7), (3, 8), (1, 6)

C2: (5, 9), (7, 11), (4, 7), (8, 9)

(a) Compute vectors u_1 and u_2 – the centroids of C1 and C2. (Each of u_1 and u_2 should be an (x,y) pair.)

(b) Compute the Rocchio decision boundary as a function $y \geq m x + b$. Make sure to specify which class is the output label if the inequality holds. Does this classifier misclassify any training points? If so, which ones? (Note: In computing the decision boundary, you should assume that the Rocchio classifier uses Euclidean distance to the centroids in making its classification decision.)

(c) Online Rocchio. In online learning, we are interested in learning a classifier for a dataset that is not entirely available at initial training time. For instance, if the training instances above were derived from a web crawl, we would like a simple way to update parameters u_1 and u_2 to account for new pages found and old pages removed.

- (i) Write an update rule for generating u_i' for an additional instance z in terms of the existing u_i . You may use the count N_i of documents currently in class i if necessary.
- (ii) Write an update rule for generating u_i' for an instance z to be removed in terms of the existing u_i . You may use the count N_i of documents currently in class i if necessary.

(d) Use your update rule to compute a new centroid for C2 after adding the point (6, 4). Include both the formula for the update rule and the new centroid in your answer.

5. Hierarchical Clustering

Assume we have n points on the x -axis, with the i th point being at $x = 2^{-i}$. Consider a run of Hierarchical Agglomerative Clustering on these n points with centroids used to determine the closest pair of clusters at a step. Write down the centroids of each cluster that results at the end of t steps, as a function of t .