

Write your answers on the pages; use the back of the pages if needed. Don't forget your name.

**1. Jamie's question [20 points]**

## 2. Naïve Bayes Classification [20 points]

Let  $n(t, d)$  be the count of term  $t$  being present in test document  $d$ .

- a) Specify the category scoring function in a multinomial Naïve Bayes classifier and the decision rule, using model parameters  $P(c)$ , the probability for a random document belonging to category  $c$ , and  $P(t|c)$ , the conditional probability of term  $t$  given category  $c$ . [5 points]

Answer:

$$\begin{aligned}\hat{c} &= \arg \max_c \left\{ P(c) \prod_{t \in d} P(t | c)^{n(t,d)} \right\} \\ &= \arg \max_c \left\{ \log P(c) + n(t,d) \sum_{t \in d} \log P(t | c) \right\}\end{aligned}$$

- b) Given training data  $D = \{(x_1, y_1^{(c)}), \dots, (x_N, y_N^{(c)})\}$  with  $x$  as a document vector and  $y$  as 0/1 valued, indicating the membership of a document with respect to category  $c$ , show how to use the training data to compute the maximum likelihood estimates  $\hat{P}(c)$  and  $\hat{P}(t | c)$  of the model parameters. [5 points]

Answer:

$$\begin{aligned}\hat{P}(c) &= \frac{\sum_{i=1}^N y_i^{(c)}}{N} \\ \hat{P}(t | c) &= \frac{\sum_{x_i \in c} n(t, x_i)}{\sum_{t' \in V} \sum_{x_i \in c} n(t', x_i)} \quad V : \text{vocabulary}\end{aligned}$$

- c) Define two common approaches (formulae) to smoothing in NB. [5 points]

Answer:

- Laplace smoothing:

$$\tilde{P}(t | c) = \frac{\sum_{x_i \in c} n(t, x_i) + 1}{\sum_{t' \in V} \sum_{x_i \in c} n(t', x_i) + |V|} \quad |V| : \text{vocabulary size}$$

- 2-class mixture:

$$\tilde{P}(t | c) = \lambda \hat{P}(t | c) + (1 - \lambda) \hat{P}(t), \quad \hat{P}(t) = \frac{\sum_c \sum_{x \in c} n(t, x)}{\sum_{t \in V} \sum_c \sum_{x \in c} n(t, x)}, \quad 0 \leq \lambda \leq 1$$

d) Let vector  $p(c) = (p_1(c), \dots, p_{|V|}(c))$  be the term distribution conditioned on category  $c$  where  $p_j(c) \equiv P(t_j | c)$ , and vector  $q(x) = (q_1(x), \dots, q_{|V|}(x))$  be the estimated term distribution conditioned on document  $x$ , where  $q_j(x)$  is the normalized term frequency using the document length, i.e., the count of within-document word occurrences as the denominator. The cross entropy between the two distributions is defined as:

$$H(q(x) \parallel p(c)) = -\sum_{j=1}^{|V|} q_j(x) \log p_j(c)$$

This quantity enable us to choose the category that minimizes the cross entropy for each new document  $x$ .

Show the similarity and difference between the minimum-entropy decision rule and that in multinomial NB method. Which method is more advantageous than the other? Justify your answer. [5 points]

**Answer:** From the NB decision rule we have:

$$\begin{aligned} \hat{c} &= \arg \max_c \left\{ P(c) \prod_{t \in d} P(t | c)^{n(t,d)} \right\} \\ &= \arg \max_c \left\{ \log P(c) + \sum_{t \in d} n(t,d) \log P(t | c) \right\} \\ &= \arg \max_c \left\{ \frac{\log P(c)}{n(x)} + \sum_{t \in d} \underbrace{\frac{n(t,d)}{n(x)}}_{q_t(x)} \underbrace{\log P(t | c)}_{p_t(c)} \right\} \\ &= \arg \max_c \left\{ \frac{\log P(c)}{n(x)} + \sum_{t \in d} q_t(x) \log p_t(c) \right\} \\ &= \arg \max_c \left\{ \frac{\log P(c)}{n(x)} - H(q(x) \parallel p(c)) \right\} \\ &= \arg \min_c \left\{ -\frac{\log P(c)}{n(x)} + H(q(x) \parallel p(c)) \right\} \end{aligned}$$

The common part of the two methods is that both take minimum cross entropy into account in scoring (and ranking) candidate categories. As for the difference, NB takes category prior into account while the minimum-entropy classifier does not. Thus, the former makes a better explanation of the observed training data than the latter.

### 3. Text Categorization Evaluation [20 points]

Consider a test set of documents D1, D2 and D3 where D1 have *one* relevant category, D2 has *three* relevant categories and D3 has *four* relevant categories for D3 as the truth. Systems A and B produced a ranked list of 10 categories for each document as shown in Table 1 where the black boxes indicate the locations of relevant categories, and white boxes indicate the locations of irrelevant categories.

Table 1. Ranked categories for each test document by systems A and B

Rank	A			B		
	D1	D2	D3	D1	D2	D3
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						

- a) Fill the slots in Table 2 with the test statistics at the decision threshold at top 3, 5 and 10 on all the ranked lists by each system, i.e., A and B respectively. (Note: when recall and precision are both zero, we define F1 to be zero.) [8 points]

Table 2. Evaluation statistics for systems A and B on three test documents

	top 3	top 5	top 10	top 3	top 5	top 10
Total True Positive	3	5	8	5	5	8
Total False Positive	6	20	92	4	20	92
Total False Negative	5	3	0	3	3	0
Micro-avg Recall	0.38	0.63	1.00	0.63	0.63	1.00
Micro-avg Precision	0.33	0.20	0.08	0.56	0.20	0.08
Micro-avg F1	0.35	0.30	0.15	0.59	0.30	0.15

- b) Treating each test document for categorization as a “query” in ad hoc retrieval, we can evaluate the system-produced ranked lists using the Mean Average Precision (MAP). Fill up the open (non-dark) slots in the table with the corresponding precision values, and compute the MAP values for system A and B, respectively. [8 points]

Table 3. Precision scores computed at positions of relevant categories

	A			B		
	D1	D2	D3	D1	D2	D3
at 1st relevant	0.50	1.00	0.50	0.25	0.50	1.00
at 2nd relevant		0.25	0.40		0.67	0.67
at 3rd relevant		0.33	0.50		0.38	0.43
at 4th relevant			0.44			0.40
Average Precision	0.50	0.53	0.46	0.25	0.52	0.63

- MAP of A =
- MAP of B =

- c) Based on the evaluation results in “a” and “b”, which system is better, in what sense? Which metric makes more sense to you? Justify your answer. [4 points]

Answer:

System A is better than system B in terms of ranking all the relevance categories on average, which is reflected in the MAP scores. On the other hand, if we only focus on top-3 categories (the higher precision end), system B is better than system A, as reflected in the averaged F1 scores at threshold on rank 3.

#### 4. Significance Tests [19 points]

Two research papers reported comparative evaluations with classification methods A and B. The per-category F1 scores were reported for each system as the following:

1. Paper 1 used a test set of documents in 10 categories which are the most common subset from a larger classification scheme of 100 categories. For *two* out of the 10 categories, both methods A and B had the perfect F1 value of 100%. For the remaining categories, method A was better than method B for *6* categories while method B was better than method A for *2* categories.
  2. Paper 2 used the total set of 100 categories in the evaluation set. For *20* out of the 100 categories, both methods A and B had the worst F1 value of 0% due to extremely small numbers of training examples for rare categories. For the remaining categories, method A was better than method B for *60* categories while method B was better than method A for *20* categories.
- a) Choose a proper type of statistical significance test for the case of paper 1, and fill up the slots Table 4. [8 points]
  - b) Choose a proper type of statistical significance test for the case of paper 1, and fill up the slots Table 5. [8 points]
  - c) Provide a generic description about the meaning of a p-value. [3 points]

**Table 4. Significance Test for Paper 1**

Choice of the test, and why	Sign test (macro-level) is the only possible choice given the provided information.
Test Statistic	$Y \sim \text{Bin}(8, p)$ , the number of times for A to outperform B in $n (=8)$ trials.
H0	$p = EY = 0.5$
H1	$p > 0.5$
Extreme supporting H1	Large
P-value (formula & score)	$\Pr(Y \geq 6 \mid p = 0.5) = \frac{1}{2^8} \sum_{j=k}^8 \frac{8!}{k!(8-k)!} = \frac{1}{256} (1 + 8 + 28) = 0.14 > 0.1$
Conclusion and concerns	The p-value does not provide statistically strong evidence for H1. One concern is that the number of trials ( $n = 8$ ) is rather too small for a sign test. I would evaluate on a larger number of categories, i.e., the fullest of 90 categories in the Reuters-21578 benchmark dataset instead. Also, it is important to know who the categories were selected. For example, if their the subset of the most common categories in a large classification system with many categories, then the subset is not necessarily representative of the majority of the categories (less common). As a result, the “randomness” (i.i.d) assumption in the sign test would be

	violated.
--	-----------

**Table 5. Significance Test for Paper 2**

Choice of the test, and why	Sign test (macro-level) is the only possible choice given the provided information. Since $n$ is large, we need to use normal distribution to approximate the p-value in the binomial distribution.
Test Statistic	$Y \sim \text{Bin}(80, p)$ , the number of times for A to outperform B in $n = 80$ trials..
H0	$p = EY = 0.5$
H1	$p > 0.5$
Extreme supporting H1	Large
P-value (formula & score)	$z(Y = k   n = 80, p = 0.5) = \frac{k - np}{p\sqrt{n}} \approx \frac{60 - 40}{0.5 \cdot 9} = 4.4$ $\Pr(Z \geq 4.4   N(0,1)) < 0.01$
Conclusion and concerns	The p-value ( $< 0.01$ ) gives a strong statistical evidence for H1, i.e., method A is better than B.

In standard normal distribution, the cumulated one-sided tail-probabilities (p-value) is approximately 10% when  $Z = 1.28$ , 5% when  $Z = 1.64$ , and 1% when  $Z = 2.32$ , respectively.

**Answer for c)**

The p-value is the probability, under the null hypothesis, for the value of test statistic as what was observed or as more extreme in the direction of supporting the alternative hypothesis.

## 5. Clustering [21 points]

Consider the task of to provide online clustering of Google-returned web pages for each query to support user browsing. Notice that the number of documents is typically very large, e.g., over 2 millions for query “document clustering”.

- Describe the generic procedure for agglomerative hierarchical clustering. Discuss the scalability of Group Average Clustering (GAC) in big-O notation of time complexity; assuming cosine similarity is used for document pairs. [7 points.]
- Describe the Buckshot/GAC procedure as a speedy alternative, provide the time complexity analysis in big-O notation, briefly discuss potential short-comings (if any) and suggest a solution. [7 points.]
- Describe the k-means clustering procedure as a speedy alternative, provide the time complexity analysis in big-O notation, briefly discuss potential short-comings (if any) and suggest a solution. [7 points.]

Answer for a):

The AHC procedure consists of the following steps:

1. Compute the similarity for each pair of documents.
2. Place each document as a singleton cluster in set  $S$ .
3. Merge the two closest clusters in  $S$  into a new cluster, add the new on to  $S$  and remove the two clusters from  $S$ .
4. Compute the similarity of the new cluster with each of the remaining clusters in  $S$ .
5. Repeat step 3 and 4 until some termination conditions are met, e.g., when the size of  $S$  becomes one, a pre-specified number of iterations is reached, or the linkage of two closest clusters (in step 3) is sufficiently small.

Generally, AHC has at least  $O(n^2)$  time complexity. GAC using cosine similarity also has  $O(n^2)$  time complexity and  $O(n^2)$  space complexity, which may be scale well for online clustering of a collection with millions of web pages.

Answer for c):

GAC with Buckshot takes the steps as:

- 1) Randomly sample a subset of  $m$  documents from the full set of  $n$  documents.
- 2) Run GAC on the subset to obtain  $k$  cluster centroids.
- 3) Assign each document in the full set to the closest centroid.

Hopefully, the clusters from the sample would be sufficiently close to those that would be produced by using the full collection. The time complexity in step 1 is  $O(m)$ , in step 2 is  $O(m^2l)$  where  $l$  is the average number of non-zero features per document vector or cluster centroid), and in step 3 is  $O(nkl)$ . Thus the total complexity is  $O(m^2l) + O(nkl)$ .

A potential short-coming is that the small clusters would be easily missed in a small-sized sample. To fix that, one solution is to use a diversity-based sampling strategy (e.g., using the Maximal Marginal Relevance principle in text summarization or the like).

Answer for d):

K-means clustering steps are:

1. Randomly sample  $k$  documents from the input dataset as the initial cluster centroids.
2. Assign each document to the closest centroid.

3. Update the each centroid using its current member documents.
4. Repeat steps 2 and 3 until the centroids are stabilized.

The time complexity is dominated by step 2, which is  $O(knl)$ . A well-known limitation of k-means is that its local optimum is sensitive to the initial seeds; how to avoid outliers in the seed selection is a crucial question. One solution is to run k-means several times with random seeds, and select the clustering with the lowest “cost”, defined as the sum of the squared errors of members from corresponding centroids. Other solution is to use Buckshot/GAC to produce the initial seeds.