Examination in IN3120/IN4120

UNIVERSITY OF OSLO
The Faculty of Mathematics and Natural Sciences
Written examination IN3120/IN4120
2024 Autumn
Duration: November 29, 15:00 - 19.00 (4 hours)
Permitted aids: Inspera calculator available
It is important that you read this front page before you start.

The different questions have different weights, as indicated.

You can answer in Norwegian or English. Please use the language that you are most comfortable with.

1 EVALUATION

(a) [5p] Consider a search engine that for a given user query retrieves 15 documents, of which 8 are relevant to the query. Assume that a set of 10 relevant documents exists for the query. Calculate the precision, recall, and F_1 score for the search engine's results.

(b) [5p] Let *R* denote a relevant document, and let *N* denote a non-relevant document. Consider a search system that for the query *burrito* produces the ranked result set *RRNRNRNR* and that for the query *shrimp cocktail* produces the ranked result set *RNNR*. Given these two queries, show how to compute the search system's MAP score.

(c) [5p] NDCG is a metric used in information retrieval to evaluate the quality of a ranked list of search results, particularly focusing on the order and relevance of documents presented to the user. Explain the idea behind the metric and how it is computed.

Fill in your answer here

Format

Words: 0

Maximum marks: 15

2 MIXED GRILL

(a) [5p] Consider a search engine tasked with indexing a set of documents for a news website. Users submit keyword queries, and the search engine's job is to return the most relevant documents to these queries. The current search engine ranking algorithm relies heavily on term frequency (TF) within documents and inverse document frequency (IDF) across the document collection. However, the news site has noticed that users often look for the most recent articles. Describe how you would modify the ranking algorithm to improve the ranking quality for time-sensitive news articles. Explain your reasoning.

(b) [4p] Document A is represented as the sparse document vector {“*science*”: 0.4, “*search*”: 0.8, “*moon*”: 0.2}, and document B as {“*search*”: 0.5, “*camembert*”: 0.3, “*science*”: 0.3}. Show how to calculate the cosine similarity between documents A and B.

(c) [6p] You were once tasked to create a search engine for people with expert knowledge in their field. Your index has now started to age, and since you built the index some documents have been deleted while new documents have appeared. You want your search engine to adapt to content changes as quickly as possible. How would you handle the deletion of old documents as well as the addition of newer documents?

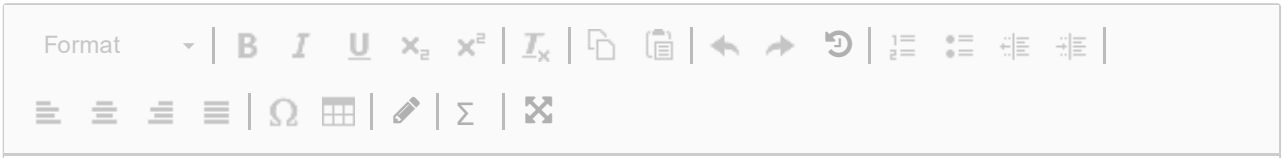
(d) [4p] Explain the beneficial effects that using a compressed index will have, and discuss some of the trade-offs you might have to make when designing such a system. You can assume that compression is applied both to the dictionary and to the posting lists.

(e) [6p] Describe how the following compression techniques work: Elias gamma encoding, variable-byte encoding, and Simple9. You can assume that gap encoding is already done.

(f) [5p] Clearly and succinctly explain why a Bloom filter does not allow for deletion of a previously added element.

Replace with question text

Fill in your answer here



Words: 0

Maximum marks: 30

3 STRINGORAMA

(a) [3p] Explain what a trie is and how it can be used to efficiently represent a set of strings in a dictionary.

(b) [8p] Given a query string *s* and a dictionary *D* represented in a trie, outline how you can efficiently find all entries in *D* that are *k* or fewer edit operations away from *s*.

(c) [4p] Discuss how your solution above scales with the value of *k*.

(d) [6p] You decide to use a suffix array to offer users the ability to do substring searches over a long string buffer, where matches can begin and end anywhere in the buffer. Show how to construct the suffix array of the string *superduper*.

(e) [4p] Demonstrate how to use your suffix array from the previous question to locate all substring matches for the query *up*. Clearly show how you arrive at your answer.

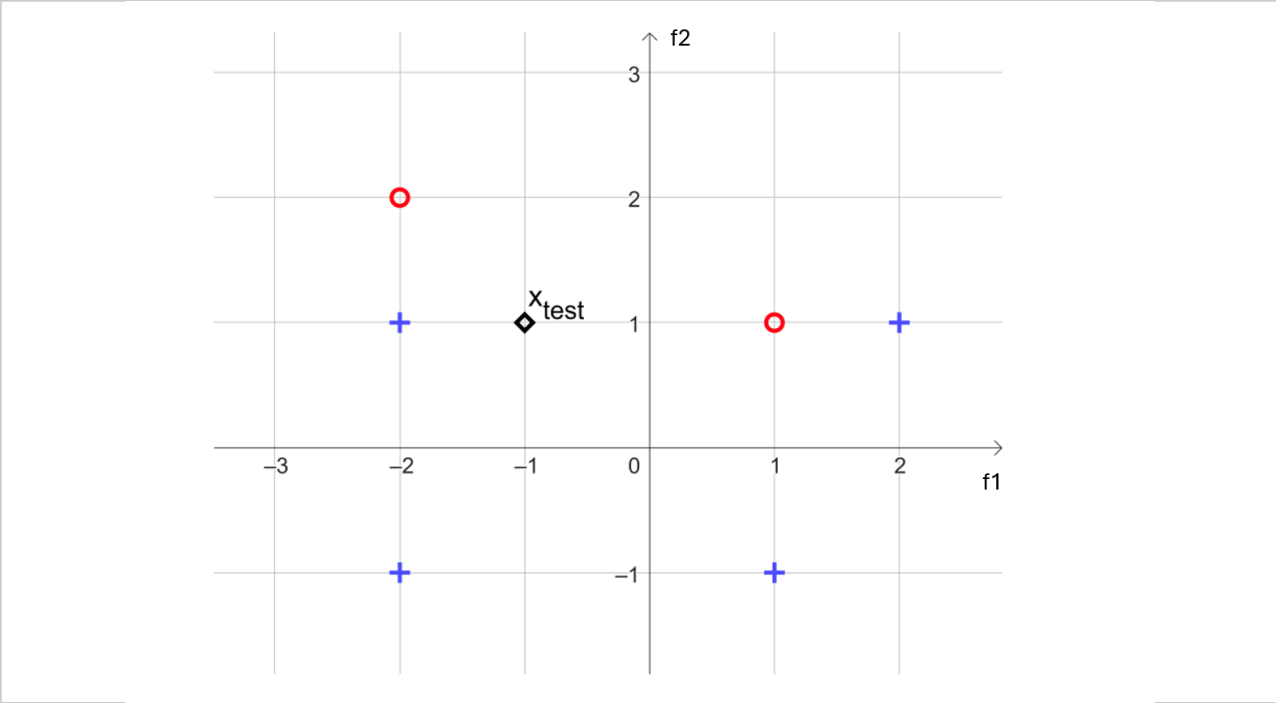
Fill in your answer here

Format

Words: 0

Maximum marks: 25

4 CLASSIFICATION



John is trying to classify objects into two classes, “+” and “o”. He has six labeled example objects, where each example is represented by a feature vector having two numerically valued features (f_1, f_2). In addition to his six labeled examples he has a seventh object x_{test} that he wants to classify.

John's data is attached (in graphical form), and is also listed below (in tabular form.)

Example	Feature f_1	Feature f_2	Label
x_1	-2	2	o
x_2	-2	1	+
x_3	-2	-1	+
x_4	1	1	o
x_5	1	-1	+
x_6	2	1	+
x_{test}	-1	1	?

(a) [6p] John first tries a 3-nearest neighbor classifier with simple unweighted voting, and the Euclidean norm as the distance metric.
(i) [3p] What prediction will the model make on x_{test}, and why? Clearly show how you arrive at your answer.
(ii) [3p] Discuss the impact that weighted voting might have for the prediction of x_{test}.

(b) [6p] John then tries a Rocchio classifier, also with the Euclidean norm as the distance metric. What prediction will the Rocchio model make on x_{test}, and why? Clearly show how you arrive at your answer.

(c) [6p] John considers using an SVM classifier, but looking at the data he’s not quite sure what type of SVM classifier to use. Which type(s) of SVM classifier would you advise John to use for his classification problem? Explain your reasoning.

(d) [12p] John then tries a naive Bayes classifier. He isn't quite sure how to apply naive Bayes to a problem where the features are numerical (although this is absolutely possible), but recalls from IN3120 how the presence of words (and their counts) in a category were used. He therefore decides to transform his examples into having simple Boolean features: Instead of (f_1, f_2) they become (b_1, b_2), where b_i is 1 if and only if $f_i > 0$ and 0 otherwise. In other words, (b_1, b_2) basically indicates which quadrant in (f_1, f_2) space the example falls into. John's transformed data is listed below (in tabular form.) Given the transformed data table, and given that John applies add-one Laplace smoothing when estimating probability estimates, what prediction will the naive Bayes model make on x_{test}, and why? Clearly show how you arrive at your answer.

Example	Feature b_1	Feature b_2	Label
x_1	0	1	o
x_2	0	1	+
x_3	0	0	+
x_4	1	1	o
x_5	1	0	+
x_6	1	1	+
x_{test}	0	1	?

Fill in your answer here

Format

Words: 0

Maximum marks: 30