CS5154/6054 Final Exam, 12/7/2021

Key: bccacbda, bdbbbca, acbaabc, abcacbdc, dbcabad

1. After df['year'] = df['date'].dt.year < 2020, X_train, X_test, Y_train, Y_test = train_test_split(df['text'], df['year'], test_size=0.2), Y_train has
   a. 1 values.
   b. 2 values.
   c. 13 values.
   d. no values.

2. Suppose confusion_matrix(Y_test, Y_pred) generates a 3x3 matrix [[10, 2, 0], [8, 6, 4], [0, 3, 7]] for a test set of 40 data points. What is the accuracy?
   a. 11/40.
   b. 19/40.
   c. 23/40.
   d. 25/40.

3. Let tfidf = TfidfVectorizer(min_df=10) and tfidf.fit(X_train). tfidf.get_feature_names() returns
   a. stopwords.
   b. bigrams.
   c. the vocabulary.
   d. idfs.

4. Let model = MultinomialNB(). Which method trains the model?
   a. model.fit()
   b. model.transform().
   c. model.fit_transform().
   d. model.predict().

5. After model=svm.SVC(kernel='linear') and model.fit(X, y), we have a linear classifier wx + b. How do we extract w from the model?
   a. model.classes_
   b. model.support_vectors_
   c. model.coef_
   d. model.intercept_

6. The margin learned by a linear SVC is
   a. $|w|/2$.
   b. $2/|w|$.
   c. $\min_i\{w_i\}$.
   d. $\max_i\{w_i\}$.

7. The default kernel of svm.SVC is 'rbf', or radial basis function. Using this kernel is equivalent to mapping data to and return the innerproduct from
   a. a two-dimensional space.
   b. a six-dimensional space.
   c. a space of the same dimensionality as that of the input data x.
   d. an infinite-dimensional Hilbert space.

8. The support vector machine is solved with quadratic programming on a dual problem and the solution includes $w = \sum_i \alpha_i y_i x_i$. Support vectors are those x_i's in the training set with
   a. $\alpha_i > 0$.
   b. $y_i > 0$.
   c. $x_i > 0$.
   d. $y_i x_i > 0$.

▶ **Table 13.1** Data for parameter estimation examples.

	docID	words in document
training set	1	Chinese Beijing Chinese
	2	Chinese Chinese Shanghai
	3	Chinese Macao
	4	Tokyo Japan Chinese
test set	5	Chinese Chinese Chinese Tokyo Japan

Let us label the training set as docID {1, 2} for class1 and {3, 4} for class2.

9. TrainMultinomialNB finds T_{ct} for c=class1 and t=Chinese to be
   a. 2.
   b. 4.
   c. 5.
   d. 6.

10. With add-one smoothing, TrainMultinomialNB estimates P(Tokyo|class2) as
    a. (1+1)/(2+6) = 2/8.
    b. (1+1)/(3+6) = 2/9.
    c. (1+1)/(4+6) = 2/10.
    d. (1+1)/(5+6) = 2/11.

11. For multinomial NB, P(classj|d5) will be proportional to
    a. a = P(Chinese|classj)P(Tokyo|classj)P(Japan|classj).
    b. b = P(Chinese|classj)^{3}P(Tokyo|classj)P(Japan|classj).
    c. a(1-P(Beijing|classj))(1-P(Shanghai|classj))(1-P(Macao|classj))
    d. b(1-P(Beijing|classj))(1-P(Shanghai|classj))(1-P(Macao|classj))

12. TrainBernoulliNB finds N_{ct} for c=class1 and t=Chinese to be
    a. 1.
    b. 2.
    c. 3.
    d. 4.

13. With add-one smoothing, TrainBernoulliNB estimates P(Tokyo|class2) as
    a. (1+1)/(1+2) = 2/3.
    b. (1+1)/(2+2) = 2/4.
    c. (1+1)/(3+2) = 2/5.
    d. (1+1)/(4+2) = 2/6.

14. For Bernoulli NB, P(classj|d5) will be proportional to
    a. a = P(Chinese|classj)P(Tokyo|classj)P(Japan|classj).
    b. b = P(Chinese|classj)^{3}P(Tokyo|classj)P(Japan|classj).
    c. a(1-P(Beijing|classj))(1-P(Shanghai|classj))(1-P(Macao|classj))
    d. b(1-P(Beijing|classj))(1-P(Shanghai|classj))(1-P(Macao|classj))

15. When binarized MNB is used, the only T_{ct} that will be different is
    a. c=class1 and t=Chinese.
    b. c=class2 and t=Chinese.
    c. c=class1 and t=Tokyo.
    d. c=class2 and t=Tokyo.

16. When binarized MNB is used, P(classj|d5) will be proportional to
    a. a = P(Chinese|classj)P(Tokyo|classj)P(Japan|classj).
    b. b = P(Chinese|classj)^{3}P(Tokyo|classj)P(Japan|classj).
    c. a(1-P(Beijing|classj))(1-P(Shanghai|classj))(1-P(Macao|classj))
    d. b(1-P(Beijing|classj))(1-P(Shanghai|classj))(1-P(Macao|classj))

17. One way to explain a learned linear classifier wx + b is to exhibit features associated with
    a. the most positive w values.
    b. the most negative w values.
    c. both the most positive and the most negative w values.
    d. w values in the middle range.

18. One way to explain prediction errors from a model like SVC is to use predict_proba() to show that for a misclassified sample, the outcome probabilities to Y_test and Y_pred are
    a. very different.
    b. very similar.
    c. both lower than other classes.
    d. both zero.

19. Non-negative matrix factorization (NMF) approximates an n x m nonnegative document-term matrix C with the product of two non-negative matrices W (n x k) and H (k x m). If each of the k intermediate dimensions is a topic, then the prominent documents and terms associated with the topic are determined with
    a. the most positive elements in a W column and an H row.
    b. the most negative elements in a W column and an H row.
    c. middle-range elements in a W column and an H row.
    d. both the most positive and most negative elements in a W column and an H row.

20. NMF is often solved with an algorithm that alternates between finding the best H with W fixed and finding the best W with H fixed, each solved with
    a. a non-negative least squares solver.
    b. quadratic programming.
    c. maximum likelihood estimation.
    d. non-negative singular value decomposition.

21. sklearn.decomposition.TruncatedSVD() can perform SVD on the document-term matrix and then allow the user to retrieve the U (or the V) matrix, like the H matrix from NMF, from the attribute
    a. singular_values_.
    b. components_.
    c. support_vectors_.
    d. coef_.

22. Which sklearn.decomposition algorithm requires much more time than the others?
    a. NMF.
    b. TruncatedSVD.
    c. LatentDirichletAllocation.
    d. PCA.

23. Let C be the term-document incidence matrix (rows are terms). What is a diagonal entry of CC^T?
    a. document frequency
    b. term frequency
    c. document length
    d. size of the bag of words

24. For the same C in Problem 23, what is a non-diagonal element of the matrix C^TC?
    a. number of documents in which two given terms cooccur
    b. number of terms two documents share
    c. number of terms occurring in one or both documents
    d. number of documents containing one or both terms

25. The eigenvalues of CC^T are the
    a. singular values of C.
    b. square roots of the singular values of C.
    c. squares of the singular values of C.
    d. logarithms of the singular values of C.

26. Let SVD of C, the term-document matrix in Problem 23, be $U\Sigma V^T$. Terms can be embedded into a 2-dimensional space with the first two
    a. rows of U.
    b. columns of U.
    c. rows of V.
    d. columns of V.

27. The k-means algorithm alternates between "centroid computation" and "membership assignment". These steps indeed are
    a. TrainMultinomialNB and ApplyMultinomialNB.
    b. TrainBernoulliNB and ApplyBernoulliNB.
    c. TrainRocchio and ApplyRocchio.
    d. Train-kNN and Apply-kNN.

28. The k-means, as an unsupervised learning algorithm, can be viewed as a supervised learning algorithm that repeatedly modifies the labels of the training samples that have been
    a. correctly classified.
    b. misclassified.
    c. the support vectors.
    d. the centroids.

29. sklearn.metrics.cluster.contingency_matrix() takes two clusterings and generates the contingency table [[10, 2, 0], [8, 6, 4], [0, 3, 7]]. There are two ways to compute the purity: sum of row max or sum of column max (over the total, 40).
    a. In both case, purity is 23/40.
    b. In both case, purity is 25/40.
    c. purity is always symmetric.
    d. purity is not symmetric, as this example shows.

30. Which of the external criteria of clustering quality below is not symmetric?
    a. normalized mutual information
    b. Rand index
    c. the F5 measure
    d. the Fowlkes Mallows index

31. In computing the Rand index, we divide all pairs of data points into counts of TP, FP, FN, and TN where TP is the number of intra-cluster pairs in both clusterings, etc. It is easy to compute TP, A = TP+FP, B = TP+FN, and N = TP+FP+FN+TN. The Rand index (accuracy) is
    a. TP/N.
    b. (A+B)/N.
    c. (N-A-B+TP)/N.
    d. (N-A-B+2TP)/N.

32. The Fowlkes-Mallows index is defined as the geometric mean of the precision and recall and, in notations of Problem 31, is
    a. sqrt(AB).
    b. TP/sqrt(AB).
    c. sqrt(AB)/TP.
    d. sqrt(AB)/N.

33. After selector = sklearn.feature_selection.SelectKBest(chi2, k=100), and selector.fit(X, y), the 100 selected features can be printed through
    a. selector.transform(X).
    b. selector.fit_transform(X, y).
    c. selector.get_support().
    d. selector.get_params().

34. Feature selection in Problem 33 is different from Problem 17's feature selection in that this (Problem 33) is
    a. before model training.
    b. during model training.
    c. after model training.
    d. only good for a specific type of models.

35. HITS makes the hub and authority scores of terms and documents of matrix C converge to
    a. Rand index.
    b. eigenvectors of CC^T and C^TC.
    c. singular values of C.
    d. the purity score.

36. Rows and columns of a 0-1 matrix that have top hub and authority scores after HITS may represent a submatrix
    a. dense with 1's.
    b. full of 0's.
    c. full of holes.
    d. with the same density as that of the matrix.

37. MultinomialNB for two classes c and ~c is a linear classifier wx + b where w_t=
    a. $\log(T_{ct}/T_{\sim ct})$.
    b. $\log((T_{ct}+1)/(T_{\sim ct}+1))$.
    c. log(P(c)/P(~c)).
    d. log(P(t|c)/P(t|~c)).