Harvard University Division of Continuing Education
Harvard Extension School ● Harvard Summer School
51 Brattle Street, Cambridge, Massachusetts 02138-3722
617 495-0977

Online Exam Cover Page

Course Name: CSCI E-81 (14728)

Exam Length: 120 minutes

Number of exam pages: 3

Please note the timer for your exam has begun. You have 120 minutes to complete and upload your exam.

Good luck!

For technical support please call: (617) 998-8571

For all other questions please call: (617) 495-0977

CSCI E-81 Machine Learning & Data Mining

Fall 2015 Exam

2 hours

Instructions:

- Write answers into a word or text document for submission. You do not have to repeat the question but it may be useful to paraphrase given the problems with automatic numbering.
- Choose 18 of the 23 questions to answer. If you answer all, we will skip the last 5 so no extra credit for answering more than 18
- Use any books, notes, online search, etc. to answer the questions
- Do not communicate with anyone during the test
- Do not post, send copies or discuss the exam until after Thanksgiving weekend
- Good luck!

Oct 1	Cloudy	Cool		Oct 11	Fair	Warm		Oct 21	Cloudy	Cool
Oct 2	Windy	Cool		Oct 12	Fair	Warm		Oct 22	Cloudy	Warm
Oct 3	Windy	Cool		Oct 13	Rain	Warm		Oct 23	Cloudy	Cool
Oct 4	Cloudy	Cool		Oct 14	Cloudy	Warm		Oct 24	Cloudy	Cool
Oct 5	Cloudy	Cool		Oct 15	Fair	Cool		Oct 25	Cloudy	Cool
Oct 6	Part-Cloudy	Cool		Oct 16	Fair	Cool		Oct 26	Cloudy	Cool
Oct 7	Part-Cloudy	Warm		Oct 17	Cloudy	Cool		Oct 27	Rain	Cool
Oct 8	Part-Cloudy	Cool		Oct 18	Cloudy	Cold		Oct 28	Rain	Warm
Oct 9	Rain	Warm		Oct 19	Fair	Cold		Oct 29	Cloudy	Warm
Oct 10	Fair	Cool		Oct 20	Cloudy	Warm		Oct 30	Fair	Cool

1) Using the above table, what are the probabilities P(Warm, Rain) and P(Cool |Cloudy) in October in Boston? Note that Cloudy != Partly-Cloudy

2) A friend is coaching a youth soccer team. Being a wannabe machine learning expert, he has come up with sets of athletic characteristics that predict who would make a good defensive vs. good offensive player. He has formulated a linear regression model with 0 for defense and 1 for offensive. Comment on the approach.

3) The Yale Chronicle reported a study by Yale students comparing themselves to Harvard students appropriately using many one-sided T-tests comparing IQ, EQ, GPA and likability index. The individual t-tests were 0.9, 0.7, 0.04 and 0.8 leading to the Chronicle claiming the superiority of Yale students. Critique the results. Are there any issues with running 4 tests?

4) How can one assess the quality of a regression fit? List a few methods.

5) What are 3 limitations of K-means?

6) Before running PCA, the standard practice is to normalize your features. Why is this important?

7) Before running clustering, a common practice is to scale your features to a similar range. Why could this be important?

8) What advantages does Gaussian Mixture Model have over other K-means?

9) Describe how one method works for assessing clustering.

10) Hierarchical clustering typically uses either an RNN (reciprocal nearest neighbor) approach or a full distance-matrix method (like HW3). How are these two methods different?

11) How does decision tree pruning relate to the bias vs. variance tradeoff?

12) In SOM, similar data points end up in a similar region of the SOM map. How is this achieved?

13) What does a high bias have to do with machine learning?

14) Without going into the mathematical foundations, how are eigenvalues relevant for interpreting Principal Component Analysis?

15) Why is a ROC curve often considered better than a 2x2 table that lists the true positive, false positives, true negative, and false negatives?

16) Why should one cross-validate?

17) Random forests are usually considered better than decision trees by using bagging. How is this done?

18) Why is a multi-layer neural network considered better than a single-layer perceptron? Specifically, what advantages does it offer?

19) What does backpropagation do?

20) What are mini-batch, batch, and stochastic gradient descent? (The exact numbers are not important—just the concept)

21) Name 3 strategies to avoid overfitting? Your answer can be specific to a given algorithm or general across multiple algorithms.

22) The margin is straight-line width between points of each class. What conceptually enables an SVM to find non-linear boundaries?

23) In finding frequent itemsets, the Apriori algorithm uses an efficient pruning method that reduces the number of itemsets at the end of each iteration. How is this achieved?