Sample Midterm Questions

CSCI E-82 Fall 2017

We are giving an exam so that you will have a chance to learn, revisit, and achieve an understanding of the material presented in this course. The exam is open-book, open-computer, open-note but strictly not open-contact with anyone regarding the test questions during or after your test until Monday. Sharing questions or answers during the 48 hour window of the exam is a violation of Harvard policy.

Although the exam allows open materials, the intent is for you to show us what you have learned. It is not an exercise in information retrieval. Consequently, unacceptable answers include:

1) Copying text from lecture or section notes
2) Taking a screenshot of a lecture slide or section code
3) Internet searching and finding some random answer that we never covered
4) Internet searching and plagiarizing a discovered answer as your own even with a citation. This is a test of your knowledge and not your internet search prowess

Problem 1: The following output for a multiple linear regression provides the model. Praise or critique the model based on the findings.

```
Coefficients:
            Estimate  StdError t-value  Pr(>|t|)
(Intercept)  7.000    0.045    <2e-16    ***
x             2.300    0.028    <2e-16    ***
x:y           1.500    2.598    0.433
y             2.940    1.598    0.033     *
x:z          -1.770    0.053    0.007     ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.935e-15 on 98 degrees of freedom
Multiple R-squared:      0.953,       Adjusted R-squared:      0.912
F-statistic: 7.836e+31 on 1 and 98 DF,  p-value: < 2.2e-16
```

Problem 2. A gradient descent problem for multiple linear regression sometimes converges to [x,y,z] = [0.34, 0-.71, 1.35] but other times converges to two other points. Explain what might be happening and suggest an approach to fix it.

Problem 3. Discuss the issue of heteroscedasticity or non-constant variance in terms of how you would detect it, why it's an issue, and what would you do about it?

Problem 4. Machine learning often recommends having a training set, validation set, and a test set. Why?

Problem 5. For time series modeling, discuss the appropriate usage of differencing, log differencing, and seasonal differencing.

Problem 6. The standard dynamic time warping goes from the bottom left to the top right corner of the distance matrix. What would it imply if it started in the middle and proceeded to the top right?

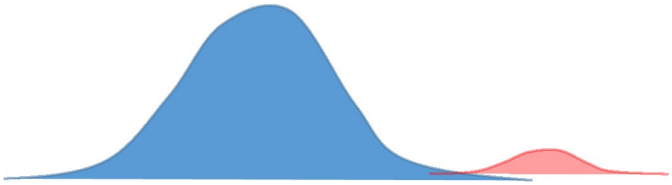
Problem 7. The knee in the curve plot for k-means (distance to cluster center on the y-axis vs. k on the x-axis) is often used to determine good values of k. When should it be used or when should it not be used?

Problem 8. How could the Jarvis Patrick algorithm relate to network models?

Problem 9. How does the centroid method differ from the average pairwise difference linkage method?

Problem 10. Compare/contrast the Lance-Williams hierarchical clustering with the BIRCH method.

Problem 11. In lecture, we discussed how the Expectation-Maximization algorithm could be used to separate two Gaussian distributions for edge pixels and non-edge pixels after using an image processing edge detector. What variables would be modeled in the E and M steps?

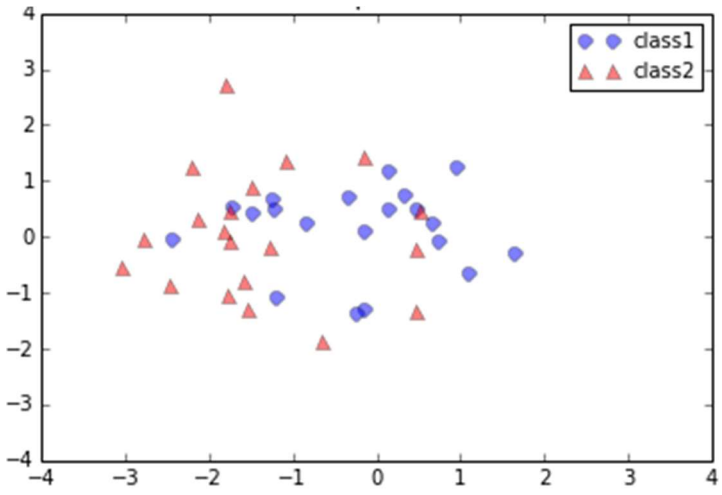


Problem 12. The B-cubed metric is a useful extrinsic clustering measure yet few students used it to assess the stock clustering problem. How could such an extrinsic measure be used in HW2?

Problem 13. When would you use a treeMap versus a dendrogram?

Problem 14. In regression and most machine learning algorithms, we aim to minimize the variance. PCA searches for the solutions that maximize the variation. Why is this?

Problem 15 Interpret the following figure that shows the PC1 (x-axis) and PC2 (y-axis). What questions should you be asking before making an interpretation?



Problem 16. Compare/contrast SOM and K-Means.

Problem 17 How would you construct a bagged KNN algorithm? What benefits would it offer?

Problem 18. My buddy Joe has written a super-fast low-memory binary decision tree classifier. How would you use it to distinguish 4 classes?

Problem 19. As an experiment, you vary the regularization parameter for a logistic regression model on your training set and your non-blind test set that you've set aside. The training data result gets continually worse by increasing the regularization parameter. The test set improves initially improves but later decreases. Explain what is happening.

Problem 20. How does a genetic algorithm optimization improve rather than just randomly propagating with mutations and cross-overs?

Problem 21. What are the challenges of using a supervised approach for outlier analysis and how could you get around them?

Problem 22. Can regression be used to identify outliers and what are its limitations?

Problem 23. Isolation trees generally continue until there is one node per leaf. Decision trees are usually built and pruned back so that leaves are no longer not a single class. Random forests continue until a set depth but generally are never pruned. Explain the differences.

Problem 24. Kaggle has a competition to identify spooky words. Hallowe'en is over. How would you go about identifying pro- or anti-racist articles on the web using machine learning techniques?

Problem 25. Some CNN architectures can include up to 20 layers of nodes. Apart from faster and parallel computers and better GPU cards, what enables CNNs to function with such large networks?

Problem 26. Bias is used to ensure the right level of activation in neural networks. If dropout removes input values to a node, how does a network achieve sufficient activation and why you even bother using it?

Problem 27. What are the similarities/differences between PCA and LSI?

Problem 28. Explain the issue of fat tails and random Erdos-Renyi networks

Problem 29 Given the following item set, which 2-item combinations are complete?

Item sets: X,Y X,Y,Z, Y,Z, W, X, Y, Z, X,Z W,Y,Z

The complete ones are listed in red. The answer would be XY, XZ, and YZ since none of these have a 3-combo superset with the same count yet the other 3 do.

Problem 30. APRIORI algorithm with the following itemset. Assume 2 is the frequent support .

Item sets: X,Y X,Y,Z, Y,Z, W, X, Y, Z, X,Z W,Y,Z

1-itemsets: W=2 X=4 Y=5 Z=5

2-itemsets: W,X = 1 W,Y=2 W,Z=2 X,Y = 3 X,Z=3 Y,Z=4

3-itemsets: W,Y,Z=2 X,Y,Z=2

Why were only these two 3-itemsets considered?