CPSC 392– Introduction to Data Science

Spring 2020

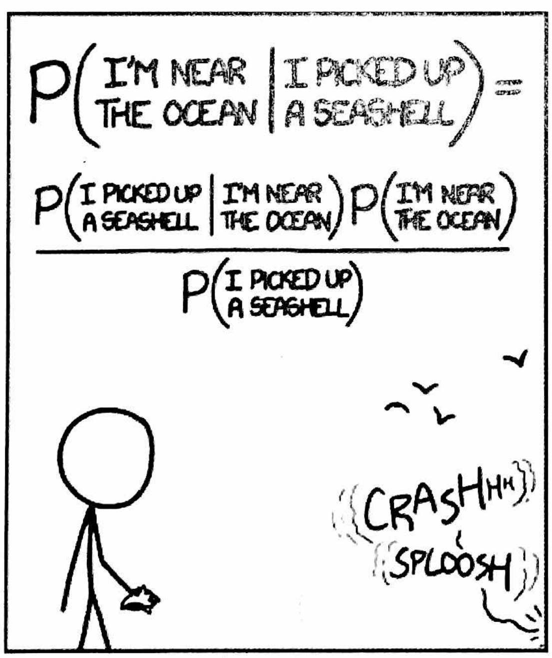
Exam II

This examination is closed book and notes. There are 6 problems and 40 points possible. You have 75 minutes to earn as many points as you can. You also get 15 extra minutes to upload your answers to Blackboard. Good luck!

In the questions below, whenever an explanation is required ("why?"), full credit will not be given if the explanation is not provided.

Name :

Toby Chappell



STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND *DON'T* HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

Academic Integrity Agreement

*failure to sign will result in a 0 for this exam

I certify that I have read and understand Chapman University's policy on academic integrity (https://www.chapman.edu/academics/academic-integrity/_files/academic-integrity-policy.pdf).

In addition to the examples listed in the policy document, I am aware that the following actions also constitute an academic integrity violation:

- Copying source code from another individual or the Internet without attribution
- Modifying someone else's code, without attribution, with the intention of claiming it as one's own work
- Referencing solutions to exams or assignments from previous course offerings that have not been made publicly available by the instructor

Furthermore, I understand that any instances of academic misconduct (regardless of circumstances or severity) in computer science or software engineering courses will result in a report to the university Academic Integrity Council with the recommended sanction being one of the following:

a) A grade of "F" in the course
b) A one semester suspension
c) Expulsion

Toby Chappell

(Print Name)

(Signature)

4/22/20

(Date)

1. So far, we have learned about unsupervised machine learning algorithms that perform clustering on a data. Suppose you are tasked with predicting a numerical value of some attribute, but that attribute is missing from historical data. Is it possible to use unsupervised techniques to predict a numerical value? Why? Support your argument either ways. (5)

It is possible to predict a numerical value for some attribute but there must be certain criteria that need to be met. For one, values to predict must be discrete. If this is not the case unsupervised clustering will predict each data point as its own cluster (assuming the model is accurate) or possibly many small clusters (both of which would provide ~~m~~ meaningless results). In addition, data points should have numerical values that are repeated frequently throughout the data. While it may difficult to distinguish this piece of criteria (since we don't have access to the numerical values), we should be able to determine this based on what we're predicting. For example, if we are trying to predict survey results from an int 1-5, we could use unsupervised clustering.

2. In your own words, explain how PCA works. What happens when you don't perform standard scaling and mean normalization on the data prior to running PCA? (5)

allows data scientists to

PCA ✓ reduces the number of features in ~~a piece~~ some data while maintaining the integrity of the data. If standard scaling and normalization are not performed prior, PCA will ~~the~~ ~~are affected by data of~~ not be able to determine correctly which features are the most relevant (gives bias to extreme values since units are not taken into account).

PCA Process:

data: x

covariance matrix: $C = x * x^T$

eigen decomposition on C: W (ordered by eigenvalues: λ)

Multiply x by $W * k$ where k is number of features: T

$$K \left\{ \begin{array}{l} P(\text{class } 1 \mid \text{attribute } 1, \ldots \text{attribute } N) = \underbrace{P(\text{att}1 \mid c1) * \ldots P(\text{att}N \mid c1) * P(c1)}_{N+1} \\ \vdots \\ P(\text{class } k \mid \ldots) = \ldots \end{array} \right.$$

3. Assume you build a Naïve Bayes classifier on a data set consisting of N attributes, each of which can take on D values. Further assume that the target (class) you are trying to predict can take on K values. Approximately how many probability values do you need to compute for the model? How many would you need if we removed the conditional independence assumption and modeled the joint distributions. (Hint: Your answers can be expressed in terms of N, D, and K alone). Why is this important in practice? (10)

Conditional Independence:

\# of values: $D^N * (N+1) * K$

Reason: There are D^N combinations of attributes, K classes to compute, and $(N+1)$ probabilities to find for each class (+1 from P(class))

No Independence:

\# of values: $D^N * K * (D^N + 1)$

Reason: Same reason for D^N and K, but now $(D^N + 1)$ probabilities for a each class and combination (joint probabilities)

It's important in practice because if data does not have independence, accuracy of model will be impacted. However, Naive Bayes allows for greater speed which may be optimal if data set is large even if data is not conditionally independent.

4. A k-Nearest Neighbors algorithm takes a k value and predicts the class of a data point, using the labels of k closest data points. You are tasked with coming up with a version of this algorithm, which predicts a continuous value of some attribute instead of a class. How will you go about making this algorithm? (10)

In order to implement this algorithm, we would need to assign a continuous value to a test point based both on the distance from and values of the k data points. Distance away would allow us to predict the value of a test point relative to the k data points (acting as a "slope" function as for a line). The actual values of the data would provide a ~~base~~ value to model off of (acts as a y-intercept as for a line). For instance, value of test point can equal average of k distances plus average of k values.

Algorithm:

1) Set value of k

~~2) Initialize centroids randomly~~

2) ~~For test points~~ find distance between test point and all data points

3) Sort data points by distance

4) Assign continous value to test point using k data points (use average ~~on~~ value + average distance)

5) Repeat for all test points

5. Write the steps for performing kMeans on a data. Given that kMeans and Hierarchical Clustering both use distance between data points as a main feature, when will you use one over the other? (5)

1) Set value of k (can use elbow method)

2) Initialize k centroids randomly

3) Assign data points to each centroid based on minimum of some distance metric

4) Set centroid position to average of cluster

5) Repeats steps 3-4 until centroids reach convergence (don't move)

Hierarchical clustering cannot be used when data becomes too large. However, it can detect patterns within the data even if the data is "noisy".

The opposite is true about kMeans in that it scales to large data well but is difficult to find clusters when they are not obvious (ie. something like

vs)

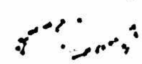
doesn't work well

works well

6. List all linkage functions for agglomerative hierarchical clustering. What type of data will each function work best with? (5)

Ward: Finds the average of each cluster (centroid) and links clusters with a min distance between centroid. Works best with average set of data (no abnormalities) or data with varying densities

Complete: Uses max distance of data points between 2 clusters.
Works best with average set of data

Single: Uses min distance of data points between clusters.
Works best with spherical () or oddly shaped () data

Average: Uses average distance of all data points between clusters
Works best with average set of data