Name: Alexander Wilson

CS 494 / Stat 420
Big Data Science and Capstone

Fall 2019 Final

Take-home
Due 19 December 2019 at 10:00pm

[Q. Snell, ES. Tass]

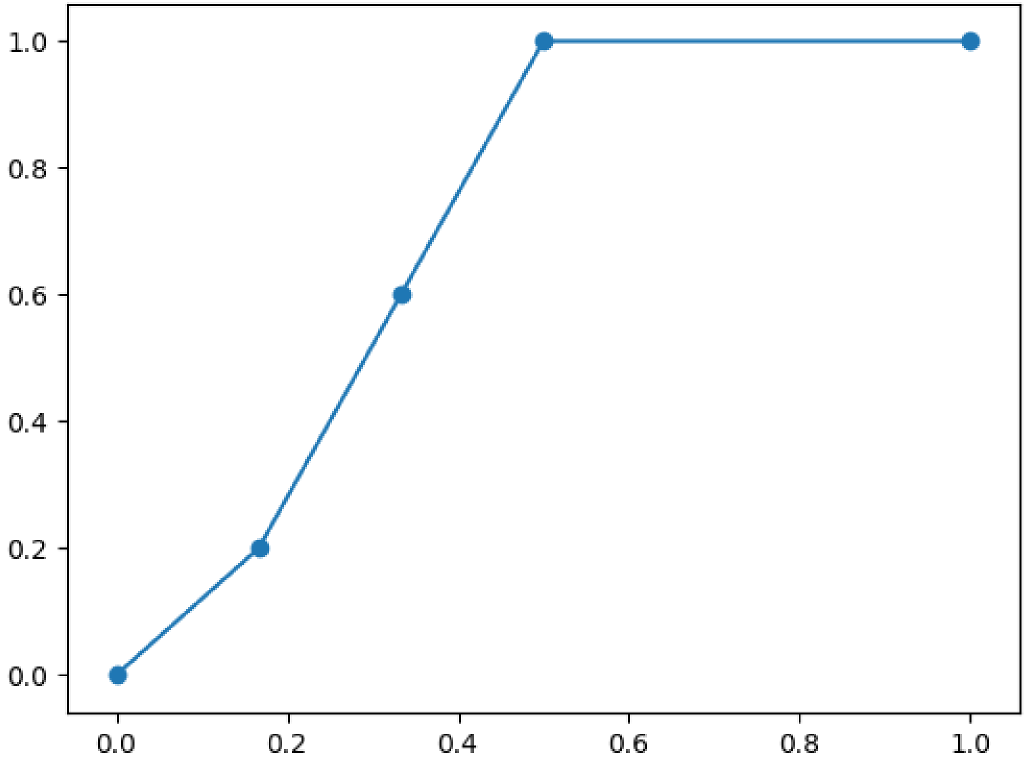
Open Notes/Book
No Discussion Allowed

Name: Alexander Wilson

1. Suppose that you perform a logistic regression. The following table shows eleven points from the test data along with the predicted probabilities that Y=1 estimated by your model:

Truth	0	0	0	0	1	1	1	0	0	1	1
P(Y=1) (esti-mated)	0.21	0.46	0.52	0.89	0.45	0.49	0.86	0.24	0.22	0.70	0.53

Draw the ROC curve for 5 points using cutoffs of 0, 0.25, 0.5, 0.75 and 1.0. In other words, find the values (the "x" and "y" coordinates) of the ROC for those 5 points and make a sketch of the plot.



Name: Alexander Wilson

2. (4 points) Consider the problem of overfitting.

a) What does it mean for a learning algorithm to overfit the training data?

The algorithm begins to discover and rely on associations that are specific to the training data, causing it to come closer to memorizing the data than discovering general associations that can apply to most datasets.

b) What would be a practical way of detecting overfitting?

If the model's loss on the training set continues to decrease while the loss on the testing data is not, then the model has entered a realm of overfitting.

c) Why would it be important to detect, and if needed, combat overfitting?

The goal of training most models is to minimize the loss on novel data after the training period is over. Once the loss of a model on novel data begins to increase as a result of overfitting, then we are being counterproductive by continuing to train it, and we should stop before this point.

d) How does decision tree learning try to avoid overfitting?

Pruning the lowest nodes of a decision tree, where the differences in traits may not be very indicative of the data's classification, and then having the prediction at the new leaf nodes be the majority class of the testing data points in that node can help the decision tree not become overfitted to the specific data points in the training set. This can be one by limiting the maximum depth of the tree, or by limiting how many points can be in a node before no longer splitting to be greater than 1.

Name: Alexander Wilson

3. (3 points) Assume that you run Principal Component Analysis (PCA) on a dataset, and that PCA produces the following eigenvalues: 1.4, 0.45, 2.0, 0.55 and 0.6.

a) How much (proportion) of the overall variance is explained by the first 3 principal components (show your work)?

$$\frac{1.4+0.45+2.0}{1.4+0.45+2.0+0.55+0.6} = 0.77 = 77\%$$

b) What does that tell you about the quality of a potential reconstruction from these 3 components?

Using only these three components would capture most (77% to be exact) of the variation in the dataset. Conversely, dropping the last two components would only lose 23% of the variation in the dataset.

c) What are potential advantages/benefits of using PCA?

When dealing with high-dimensional datasets, there may be some dimensions of the data which do not add enough information to be worth the extra processing they require. PCA allows us to determine which dimensions have the most variation between points of data. The more variation, the more information our models have to work with. This means that dimensions which have little variation in them can be removed to make the data easier to comprehend and process without significantly reducing the amount of information on relationships between different points.

Name: Alexander Wilson

4. (6 points) Design MapReduce algorithms to take a very large file of integers and produce as output:

(a) The largest integer

- read in and parse file into list of integers
- parallelize the list
- call distinct to remove duplicates
- call sortBy with a lambda that just returns the value (they aren't in key-value pairs), specifying that ascending is false
- call take with an argument of 1 to retrieve the largest integer
- print the result of the call to take

(b) The average of all the integers

- read in and parse file into list of integers
- capture the length of the list in a variable
- parallelize the list
- map each integer into a key-value pair of (0, <integer>)
- call reduceByKey on the mappings, specifying the add operation as the combining function)
- collect the 1 mapping
- divide the value of the mapping by the length of the list captured earlier
- print the result of this division

(c) The same set of integers, but with each integer appearing only once

- read in and parse file into list of integers
- parallelize the list
- call distinct to remove duplicates
- collect the integers
- print the integers

Name: Alexander Wilson

5. (6 points) For each data mining task described below, circle the approach you would recommend.

(a) From searches involving 1 or more keywords, discover which keywords tend to occur together.

i. Classification
ii. Clustering
iii. [✓ Association Rule Mining]
iv. Regression

(b) From past underwater sonar data, build a model that allows you to decide whether an approaching object is a fish or a torpedo.

i. [✓ Classification]
ii. Clustering
iii. Association Rule Mining
iv. Regression

(c) From data on incoming BYU students, identify predictors of ACT scores.

i. Classification
ii. Clustering
iii. Association Rule Mining
iv. [✓ Regression]

(d) From descriptions of a number of animals, build a zoological taxonomy (or hierarchy).

v. Classification
vi. [✓ Clustering]
vii. Association Rule Mining
viii. Regression

(e) From past bariatric surgery patient records and outcomes, build a model that predicts what type of surgical procedure to use on new patients.

i. [✓ Classification]
ii. Clustering
iii. Association Rule Mining
iv. Regression

(f) From records of student class schedules, discover which classes tend to be taken concurrently.

i. Classification
ii. Clustering
iii. [✓ Association Rule Mining]
iv. Regression

Name: Alexander Wilson

6. (3 points) Tom Khabaza (a leading UK data mining consultant) once said: “Projects never fail due to lack of patterns.” If this is true, then, in your view, and based on your understanding and experience, what may cause data mining projects to fail?

- Clients may not know what they actually want to learn from the study
- Clients may not communicate clearly what they want to learn from the study
- Clients may not have a clear idea of what they hope to gain from the insights produced by the study, which may result in the project being scrapped before completion with changes in leadership
- Due to a lack of understanding, clients may hire people who are not prepared to handle their particular data and needs.
- Clients may produce data for a study which can’t be used to gather the type of insights they would like
- Clients may not understand the costs of processing big data, and as a result they may not budget enough money for the project in the case of unseen delays or issues.

Name: Alexander Wilson

7. (2 points) You are running a local caucus meeting and wish to invite people from the party you represent. To do that, you want to use a model that predicts people's political affiliation. You want to make sure people of the opposing party are not antagonized by a misdirected invitation to attend your caucus meeting. Two companies, C1 and C2, offer you a predictive model. C1's product has an F-measure of 0.82, while C2's product has an F-measure of 0.91.

(a) Would you be able to make a decision on the basis of this information?

i. [✓ Yes]

ii. No

(b) Which of the following metrics would you like to measure and favor in your selection of a predictive model?

i. [✓ Precision]

ii. Recall

iii. Accuracy

iv. F-measure

v. V-measure

vi. Rand index

Name: Alexander Wilson

8. (2 points) Consider the following scenarios/situations.

(a) You run the Apriori algorithm on a medical database containing symptoms about a number of patients. After what seems a long time, the algorithm completes and one of the rules it returns is: <*gastro-esophageal reflex disease*, *diabetes*, *urinary stress*, *back pain*, *depression*, *high cholesterol*, *arthritis*, *short breath*, *morbid obesity*> □□ *sleep apnea*. Based on this output, explain what caused Apriori to take a long time to execute.

i. The computer it ran on was very slow
ii. [✓ The algorithm generated over 1,000 frequent itemsets]
iii. The support threshold was set too low
iv. The algorithm is recursive
v. Obesity is a complex medical problem

(b) What performance/behavior criteria would most likely have you choose decision tree classification?

i. [✓ Comprehensibility]
ii. Incrementality (i.e., ability to update model as new training data becomes available)
iii. Discrete attributes
iv. Speed of prediction
v. Attribute selection
vi. None of the above
vii. ii, iii and v
viii. i, iii, iv and v
ix. i through v

Name: Alexander Wilson

9. (2 points) State at least 2-3 distinct reasons why constant communication with the client is so essential in a data mining project?

- As was mentioned earlier, a client may not understand or clearly communicate exactly what they hope to gain as a result of the project. Over time, as progress in the project moves forward, the client may realize that what they asked you to study isn't what they actually wanted you to study. If they realize this in the earlier phases of the study, then the project's direction can change to fit what the client actually wants. In order for this to happen, however, the client needs to know about the current state of the project and be thinking about it often.
- Studies don't have infinite funds, and if a client feels like inadequate progress is being made towards objectives in a project, they may cut some or all funding to the study. The important point is that the client just needs to *feel* that way. Although they may not understand the technicalities of what you're doing in the study, constant communication will assure them that you are focused on your work and that progress is being made.
- A large portion of a partnership with a client in any industry comes down to if the client feels like they are respected by their contractor. Clients should feel very comfortable asking questions about a project or providing input and insights, They should feel like their time is valued, and that they are kept informed about the processes that bring the two organizations together. Constant communication helps the client trust you, and will make a big difference in whether or not they work with you again.

Name: Alexander Wilson

10. (2 points) Consider the k-medoids clustering algorithm.

(a) Why is the k-medoids algorithm inappropriate in big data situations?

The algorithm creates clusters by selecting some data points to act as centers for clusters (medoids), then classifies the rest of the points into clusters based on minimum distance to the various medoids. However, the initial medoid points are usually not the best points to use as medoids in order to produce good clusters. To find them, the current medoids must be iteratively re-selected by comparisons between the meoids and the other non-medoid points. In big data scenarios, this can be a lot of points to iterate through each time and require many rounds of comparison before ideal medoids are found.

(b) What clustering method would you suggest for big data and why?

CLARA is a good clustering method for datasets that would have nice results from k-medoids if it was smaller. CLARA performs clustering on smaller samples of the dataset, resulting in fewer comparisons between points to find the final medoids.

Name: Alexander Wilson

11. (4 points) Consider the following customer transaction data set.

Cust1: {Milk, Chips, Bread, Honey, Detergent, Lettuce, Ground Beef}
Cust2: {Bread, Milk, Ground Beef}
Cust3: {Detergent, Honey, Milk, Peanut Butter}
Cust4: {Cheese, Crackers, Honey, Bread}
Cust5: {Detergent, Crackers, Pizza}
Cust6: {Peanut Butter, Milk, Cheese, Bread, Lettuce}
Cust7: {Chips, Soap, Apples, Lettuce}
Cust8: {Milk, Lettuce, Bread, Pinto Beans, Ground Beef}
Cust9: {Cheese, Lettuce, Yeast, Apples, Crackers, Milk, Sour Cream, Bread}
Cust10: {Pinto Beans, Detergent, Sour Cream, Sugar, Salt, Milk}

(a) Which of the following are frequent itemsets if minsupport is set to 30%?

i. {Milk}, {Chips}, and {Milk, Bread}
ii. {Bread}, {Bread, Lettuce}, {Ground Beef}, and {Soap, Apples}
iii. {Chips}, {Chips, Soap}, {Chips, Lettuce}
iv. [✓ {Ground Beef, Bread, Milk}, {Ground Beef, Milk}, {Milk, Bread}, {Bread}]

(b) Assuming minsupport = 35% and minconfidence = 75%, is the rule “Bread -> Lettuce” a valid association rule?

i. Yes
ii. [✓ No]

(c) Justify your answer to question (b).

i. The rule’s support is above minsupport and its confidence is above minconfidence
ii. [✓ The rule’s support is above minsupport but its confidence is below minconfidence]
iii. The rule’s support is below minsupport but its confidence is above minconfidence
iv. The rule’s support is below minsupport and its confidence is below minconfidence

(d) Assuming the corresponding itemset has support above minsupport, what is the confidence of the association rule: “Bread, Milk -> Ground Beef”?

i. [✓ 60%]
ii. 40%
iii. 50%
iv. 80%

Name: Alexander Wilson

12. (2 points) Consider the following distance matrix for a small data set of 7 points. (Only the upper part of the matrix is shown, since it is symmetric).

	P_1	**P_2**	**P_3**	**P_4**	**P_5**	**P_6**	**P_7**
P_1	0	9	100	102	25	25	75
P_2		0	105	107	26	30	90
P_3			0	4	101	99	7
P_4				0	105	103	39
P_5					0	5	56
P_6						0	54
P_7							0

(a) Which pair of points is merged first by HAC (Hierarchical Agglomerative Clustering)?

i. P_1 and P_2
ii. P_1 and P_3
iii. [✓ P_3 and P_4]
iv. P_5 and P_6
v. P_3 and P_7

(b) Which of the following is the clustering obtained by HAC after 4 merge operations, using single-link?

i. [✓ $\{P_1, P_2\}, \{P_3, P_4, P_7\}, \{P_5, P_6\}$]
ii. $\{P_1, P_2, P_7\}, \{P_3, P_4\}, \{P_5, P_6\}$
iii. $\{P_1, P_2, P_3, P_4\}, \{P_5, P_6, P_7\}$
iv. $\{P_1, P_2\}, \{P_3, P_4, P_5, P_6\}, \{P_7\}$

Name: Alexander Wilson

13. (4 points) Your soccer coach has been keeping track of team statistics for the past six months or so, including when practices are held, what games are won, etc. One day he shows up to practice and announces that, although you have not done that in the past, the team will now start practicing on Sundays. You ask why this sudden change. The coach then proceeds to explain that the data shows that your team wins all of its Wednesday games but loses most of its Monday games. There are practices on Tuesdays but none on Sundays, so he has determined that the lack of practice on the day prior to the game must explain the poor performance at Monday games.

a) Would you immediately agree with your coach? Why or why not?

I would not, because even though there is a (colloquial) positive correlation between practice the day before and performance at the game, it doesn't mean that practice the day before causes better performance at a game the next day. There could be other factors affecting the performance of the team that haven't been explored by the coach, and those other factors may be more important to focus on.

Your coach seems unwilling to budge and requires Sunday practices as of this coming Sunday. As you are about to announce your decision to withdraw from the team due to your commitment not to participate in sports on Sundays, you see your teammate, Carlos, coming across the field, late for practice. As you see him, it dawns on you that over the past two months, Carlos has been at most Wednesday games but has not been able to make Monday games because he has had to help his sick grandmother on Mondays after school. With that realization, you challenge the coach's finding and offer that a better explanation for the poor performance on Mondays may actually be due to the absence of one of the team's most talented players, Carlos.

b) How can you test your assumption, based on available data from prior games?

Create a decision tree from the data and observe what features the tree splits on. The features will be split on in descending order of how much they affect the outcome of the game. If Carlos' presence is the biggest factor in the game's outcome, then the tree will split on it first, and before the occurrence of practice the day before is split on if it's at least a bigger factor.

Assume that your test proves that you were correct and that the coach's conclusion was ill-founded.

c) What do you call the variable "Carlos is at the game" or the effect of Carlos' presence at the game in the coach's analysis?

A lurking variable.

d) Although it was somewhat easy for you to detect the problem in the case of your team's performance, what makes such effects difficult to detect in more complex situations?

There can be many factors in play, and some of them may be very hard to discover without exhaustive searching or wildly speculative testing. Some of them may be related to things nobody knows about yet (an example would be attempting to analyze causes of the plague before germs and how they spread were discovered). In causes where novel phenomena are being analyzed, most of the factors affecting the observed behavior can be lurking variables.