Name: Gregory Knapp

# CS 472 – Introduction to Machine Learning
# Winter 2022 Midterm Exam
# Take Home
## [C. Giraud-Carrier, 2232 TMCB]

**Open Notes/Book**
**Closed Internet/Neighbors/Friends**

1. (5 points) For each data mining task, circle the approach you would recommend.

(a) From sets of keyword searches (i.e., searches involving 1 or more keywords), discover which keywords tend to occur together in searches.

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

(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

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

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

(d) 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

(e) From records of student class schedules, discover which classes tend to be taken in sequence.

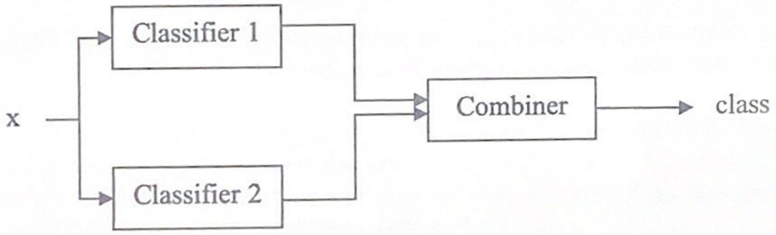
i. Classification
ii. Clustering
[✓ iii.] Association Rule Mining

Name: ________________________________

2. (1 point) Which one of the following is not a success factor in machine learning applications?

i. Domain knowledge
ii. Clear objectives
iii. Relevant data
[✓ iv.] Large volumes of data
v. Good quality data
vi. Clear communication with customer

3. (4 points) One common approach to improving classification accuracy is to train several classifiers with the same training data and then combine their outputs in some way. Assume the following setup:



Assume that 1) Combiner is a Naive Bayes (NB) learner; 2) the target function has only two possible values ($v_1$ and $v_2$); and 3) the output $a_i$ of the $i$th classifier is simply one of these two values. Then, the combiner may treat $a_1$ and $a_2$ as two Boolean attributes that describe $x$.

(a) Which of the following is the formula used by the NB classifier for classification?

$\underset{v_j \in V}{\text{argmax}} \sum_{h_i \in H} P(v_j|h_i) P(h_i|D)$

[✓ i.] Argmax($v$ in $\{v_1,v_2\}$) $P(v)P(a_1|v)P(a_2|v)$
ii. Argmax($v$ in $\{v_1,v_2\}$) $P(v|a_1)\, P(v|a_2)P(a_1)P(a_2)$
iii. 0.5
iv. Argmax($a$ in $\{a_1,a_2\}$) $P(a)P(a|v_1)P(a|v_2)$
v. If $P(a_1) > P(a_2)$ then $v_1$ else $v_2$

(b) What assumption does your system make about the two classifiers?

i. They work well on the data at hand ?
ii. Their outputs are different from each other X
iii. Their biases are different X
[✓ iv.] Their outputs are independent given the target class ?

(c) What would happen if both classifiers were exactly the same?

i. The NB combiner would output a probability that is too high
ii. The NB combiner would output a probability that is just right X
[✓ iii.] The NB combiner would output a probability that is too low X
iv. The NB combiner could not output a probability ?

Name: ______________________________

(d) Suppose you now want to combine *N* classifiers. What would you favor?

i. The number of classifiers
ii. The speed of the classifiers
iii. The accuracy of the classifiers
[✓ iv.] The diversity among the classifiers
v. The precision of the classifiers

4. (1 point) SVM can perform margin maximization efficiently in hyper-dimensional space thanks to the "kernel trick." What is it about kernel functions that make them attractive for SVM learning?

i. They are mathematically elegant
ii. They are the only ones that can be used for margin maximization
[✓ iii.] They allow all computations to be done in the original, low-dimensional space
iv. They allow computations to be done in parallel
v. They can avoid overfitting

5. (2 points) You meet Dr. Clement in the hall. He knows that you have been taking CS 472. He mentions to you that his group has recently become interested in the prediction of the secondary structure of proteins, as such structure seems to have some connection with various diseases. He further tells you that he has access to a relevant data set and invites you to help him with his prediction task.

(a) What is your first action item?

i. Ask for the data
[✓ ii.] Ask a lot of questions about the problem, the data, etc.
iii. Assure him you can do it
iv. Tell him he is wasting his time

(b) (1 point) Given that proteins are molecules and that molecules are structures best represented by graphs, consisting of atoms and the links between them, which of the following algorithms would you suggest using or adapting to this problem?

i. ID3
ii. Backpropagation
iii. NB
[✓ iv.] FOIL
v. kNN

6. (4 points) Circle the correct answer.

- [✓ T] or F : Margin-maximization finds linear boundaries between classes
- [✓ T] or F : Metalearning is learning about the behavior of learning algorithms
- T or [✓ F] : It is possible to generalize without bias
- [✓ T] or F : FOIL can learn from complex, first-order concepts such as relations and graphs

Name: ________________________________

7. (1 point) What criteria would most likely have you choose decision tree learning?

- [✓ i.] Comprehensibility ✓
- ii. Incrementality
- iii. Discrete attributes ✓
- iv. Speed of prediction ✓
- v. None of the above
- vi. ii, and iii
- [✓ vii.] i, iii and iv
- viii. i, ii, iii and iv

8. (1 point) 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 ↑ causes
- iv. The algorithm is recursive
- v. Obesity is a complex medical problem

9. (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 alone?

- i. Yes
- [✓ ii.] No

Not clear if this means select multiple since most of this test is choose 1 answer:

(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

If we want to avoid antagonizing somebody then precision helps us to know if we are avoiding false positives,

Acc = TP+TN/TP+TN+FP+PN

Pre = TP/TP+FP

Recall = TP/TP+FN

$$\text{F-measure} = \frac{2*(\text{Prec} \times \text{Recall})}{\text{Prec}+\text{Reall}}$$

Vmeasure

Name: ______________________________

10. (4 points) Circle the correct answer.

T or [✓ F] : It is easy to select the best value of *k* for *k*-medoids clustering

T or [✓ F] : The F-measure can be used to assess both classification and clustering performance

T or [✓ F] : Business users have little to do with the success of a data mining project

[✓ T] or F : The bias of decision tree learning is better than the bias of *k*-NN learning ?

11. (6 points) For each classification learning algorithm, circle the properties that apply. (C=comprehensible, L=lazy learning, E=eager learning, F=fast training, B=better than all others)

| | | | | | |
|---|---|---|---|---|---|
| ID3 | [✓ C] | [✓ L] | E | [✓ F] | B |
| *k*-NN | [✓ C] | [✓ L] | E | [✓ F] | B |
| NB | C | [✓ L] | E | [✓ F] | B |
| Backprop | C | L | [✓ E] | F | B |
| SVM | C | L | [✓ E] | F | B |
| CNN | C | L | [✓ E] | F | B |

12. (1 point) Logistic regression (LR) and SVM both rely on input space transformations. For each one, how is data dimensionality affected?

i. LR no change, SVM no change
ii. LR increase, SVM no change
[✓ iii.] LR no change, SVM increase
iv. LR decrease, SVM no change
v. LR no change, SVM decrease
vi. LR increase, SVM increase
vii. LR decrease, SVM increase
viii. LR decrease, SVM decrease
ix. LR increase, SVM decrease

13. (4 points) Consider the following customer transaction data set.

Cust1: {Milk, Chips, Bread, Honey, Detergent, Lettuce, Ground Beef} ✓ B BGM BM
Cust2: {Bread, Milk, Ground Beef} B BGM BM
Cust3: {Detergent, Honey, Milk, Peanut Butter}
Cust4: {Cheese, Crackers, Honey, Bread} B
Cust5: {Detergent, Crackers, Pizza}
Cust6: {Peanut Butter, Milk, Cheese, Bread, Lettuce} ✓ B
Cust7: {Chips, Soap, Apples, Lettuce} BM
Cust8: {Milk, Lettuce, Bread, Pinto Beans, Ground Beef} ✓ B BGM BM

Name: ______________________________________

Cust9: {Cheese, Lettuce, Yeast, Apples, Crackers, Milk, Sour Cream, Bread} ✓ B BM
Cust10: {Pinto Beans, Detergent, Sour Cream, Sugar, Salt, Milk}

Sup = 40%

A = Bread → lettuce

(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}

✓ ✓ subsets must also be frequent

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

Support = 40% 4/10

Confidence = P(Lettuce|Bread) = 4/6 = 66.7% < 75%

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"?

BGM = 3 instances

BM = 5 instance

Conf(BM → GB) = 3/5 = 60%

[✓ i.] 60%
ii. 40%
iii. 50%
iv. 80%

14. (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 |

Name: ________________________________

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

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 $P_7$

|  | P1 | D2 | {P3,P4} | P5 | P6 | P7 |
|---|---|---|---|---|---|---|
| P1 | 0 | 9 | 100 | 25 | 78 | 75 |
| P2 | 0 | 0 | 105 | 26 | 30 | 90 |
| P3P4 |  |  | 0 | 101 | 99 |  |
| P5 |  |  |  | 0 | 5 |  |
| P6 |  |  |  |  | 0 |  |
| P7 |  |  |  |  |  | 0 |

min

(b) Which of the following is the clustering obtained after 4 merge operations (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. $\{P_1, P_2\}, \{P_3, P_4, P_5, P_6\}, \{P_7\}$

15. (1 point) Richard Feynman stated: "If I cannot create it, I do not understand it." How does this apply to CS 472?

i. When we build programs, we are being creative
ii. This statement has nothing to do with CS 472
[✓ iii.] We can check our understanding of a concept/idea by how well we can implement it
iv. CS 472 is all about understanding the universe around us

16. (3 points) You are given 3 algorithms: a decision tree learner, a backpropagation learner and a *k*-NN learner. In each of the following situations, state what algorithm, or combination, you would use and why.

(a) Bank of America wants you to assist them in detecting potential fraudulent activities on their customers' credit cards. Once trained, the system must make predictions on-the-fly.

I would use k-NN since it would be easier to treat a mix of continuous data & nominal data that would likely be used in detecting credit card fraud. The prediction speed would also be rather quick & new data points are easilly added to the System

(b) Central Utah Clinic wants you to build a screening system to allocate patients to specialists based on symptoms and other health-related data. Doctors may question why certain patients are assigned to them.

Decision Tree - It will use every attribute to create a decision & is the most easilly decopherable, so a doctor can check what symptoms led to the decision

(c) Dr. Martinez wants to check up on me and see how much smarter you are after my class than before. He gives you a classification task of his choice and tells you to show off your skills. Accuracy is the sole indicator of success.

I would use backprop first since it is not clear how linearly separable the data is. If that is still not satisfactery I would use an ensemble of the 3 to vote on a decision to hopefully produce more reliable results.

Name: ______________________________

17. (2 points) Max Tegmark a Professor of Physics at MIT has said of AI/ML: "The future is ours to shape. I feel we are in a race that we need to win. It's a race between the growing power of the technology and the growing wisdom we need to manage it." Briefly comment.

I agree with Professor Tegmark that it feels as if the advances in AI/ML are coming faster than the regulations & restrictions that appear inevitable with the technology. Its a hard balance & a scary thought; if restrictions get too overbearing too quickly then it could impede possible advances & developments from ever happening. At the same time, we don't want to have to retrospectively add regulations after an accident happens, hence the race between the technology & the wisdom to use it properly

In an ideal world, I would say if our wisdom & regulations could stay just a step ahead of the technology it would be best. That way we don't have to halt or kill promising lines of research due to unneeded rules in place for 5 years but that people still pay attention to new projects & openly disuss what uses are appopriate, how could it be misused, etc.

18. (6 points) Circle the correct answer.

[✓ T] or F : Boosting can help when the class distribution is skewed
[✓ T] or F : The kernel trick is an automatic way of preprocessing corn data in SVM
[✓ T] or F : It is often the case that a learning ensemble outperforms its constituent algorithms
T or [✓ F] : Deep learning works well on images because it has no bias
[✓ T] or F : Multiple layers of linear perceptrons can solve the XOR problem
T or [✓ F] : Metalearning is doomed to failure because of NFL

19. (1 point) Fill in the blanks:

Managing expectations is key to the successful deployment of ML projects.

Name: ______________________________

20. (2 points) (2 points) Show that the transformation from $\mathbb{R}^2$ to $\mathbb{R}^5$ defined by: $(x_1, x_2) \rightarrow (2x_1^3, \sqrt{12}x_1^2x_2, \sqrt{12}x_1x_2^2, 2x_2^3, 1)$ is a reasonable kernel function for SVM.

Transformed data $2x_1^3, \sqrt{12}x_1^2x_2, \sqrt{12}x_1x_2^2, 2x_2^3, 1$ contains product of each combination of input features & can be factored down to the kernel form

$K(x_1, x_2) = (1 + x_1^T x_2)^3$ which results in the 5 dimensional transformation when expanded out.

21. (1 point) Assume we use gradient boosting with the cross-entropy loss function, i.e., $L(y, F(x)) = -(y \log F(x) + (1 - y) \log(1 - F(x))$. What would be the gradient to which the regression model would be fit? (Note: $\frac{\partial \log x}{\partial x} = \frac{1}{x}$)

i. $-\frac{y}{F(x)} + \frac{1-y}{1-F(x)}$

ii. $-\frac{y}{F(x)} - \frac{1-y}{1-F(x)}$

[✓ iii.] $\frac{y}{F(x)} - \frac{1-y}{1-F(x)}$

iv. $\frac{y}{F(x)} + \frac{1-y}{1-F(x)}$

22. (2 points) What are the roles of crossover and mutation in genetic algorithms? Why should you generally have a relatively low mutation rate?

They are both methods of finding solutions from the current "gene pool". Crossover makes more large dramatic changes to an offspring by combining parts of parent genes while mutation often makes small changes like flipping a bit. Both serve to allow the algorith to search the space for a more optimal solution.

Mutation rate should be kept low because it is essentially a random local search. If every gene mutates, it is no better than doing a random search in exponential time. By having mutation rate low, you simply introduce the possibility of a random better solution appearing then influencing the gene pool, but doesn't side track the overall algorithm progress with constant random changes.

Name: ________________________________

23. (1 point) What is the main difference between metalearning and AutoML?

i. AutoML systems do not learn while metalearning systems do
ii. AutoML systems do not rank algorithms while metalearning systems do
iii. AutoML systems rank algorithms while metalearning systems do not
[✓ iv.] AutoML systems learn while metalearning systems do not

24. (2 points) List 2 or 3 advantages of systems like H2O AutoML and pycaret.

- Compatible and easily implemented with avariety of learning models & tasks.
- Can provide model selection or model ranking outputs for a given task.

25. (2 points) Consider the following snippet of code

```
import numpy as np
from sklearn.neural_network import MLPRegressor
x = np.load('dataset.npy')
reg = MLPRegressor(hidden_layer_sizes = (200, 100, 7, 100, 200),
     activation = 'tanh', solver = 'adam', learning_rate_init = 0.0001, max_iter = 20)
reg.fit(x, x)
```

(a) What kind of neural network model is reg?

Compression auto-encoder

(b) What is happening in the hidden layer with 7 nodes and how can it be used?

The layer with 7 nodes is the latent layer that has the most important features of an image that will be decoded. The hidden layer values could be used to make a generalized compression of the inputs based on its learned values

26. (1 point) Circle the correct answer.

[✓ T] or F : Professor Giraud-Carrier is my favorite teacher ever!