Homework 1

Automated Learning and Data Analysis
Dr. Thomas Price

Spring 2022

Instructions

Due Date: February, 9 2022 at 11:45 PM
Total Points: 63
Submission checklist:

- Clearly list each team member's names and Unity IDs at the top of your submission.
- Your submission should be a single PDF file containing your answers. **Name your file**: G(homework group number)_HW(homework number), e.g. G1_HW1.
- If a question asks you to explain or justify your answer, **give a brief explanation** using your own ideas, not a reference to the textbook or an online source.
- Submit your PDF through Gradescope under the HW1 assignment (see instructions on Moodle). **Note**: Make sure to add you group members at the end of the upload process.
- Submit the programming portion of the homework *individually* through JupyterHub.

1 Data Properties (13 points) [Chengyuan]

Answer the following questions about attribute types.

1a) (*10 total points*) Classify the following attributes as *nominal, ordinal, interval* or *ratio.* Also classify them as *binary*[1], *discreet* or *continuous.* If necessary, give a few examples of values that might appear for this attribute to justify your answer. If you make any assumptions in your answer, you must state them explicitly.

i) (*1 point*) Population counts in counties of North Carolina
Solution: Discrete, Ratio

ii) (*1 point*) Annual income (in US dollars).
Solution: Continuous, Ratio

iii) (*1 point*) Existence or non-existence of cancerous tumors.
Solution: Binary, Nominal

iv) (*1 point*) Level of Pain during Injury diagnosis (0: absent, 1: mild, 2: moderate, 3: severe, 4: incapacitating)
Solution: Discrete, Ordinal

v) (*1 point*) The pH of water that has been measured using the pH scale[2].
Solution: Continuous, Interval (note that 7 is our 0 value here)

vi) (*1 point*) Categorization of clothing (e.g., hat, shirt, pants, shoes).
Solution: Discrete, Nominal

[1] Binary attributes are a special case of discreet attributes.

[2] **Note**: A pH of 7 is considered neutral, with ¡ 7 being acidic and ¿ 7 being basic.

vii) (*1 point*) Sugar content in juice in grams.
Solution: Continuous, Ratio

viii) (*1 point*) Calendar dates.
Solution: Discrete, Interval

ix) (*1 point*) The response to a yes-no question.
Solution: Binary, Nominal

x) (*1 point*) Product Satisfactory Rating (5 - Excellent, 4 - Good, 3 - Fair, 2 - Poor, 1- Horrible).
Solution: Discrete, Ordinal

1b) (*1 point*) Are all continuous attributes ratio? If so, explain why, and if not, give a counterexample.
Solution: No. Counterexample: Fahrenheit temperature is continuous, but it is also interval.

1c) (*1 point*) Are all ratio attributes continuous? If so, explain why, and if not, give a counterexample.
Solution: No. Counterexample: Count data (e.g. number of people) is ratio (there is a meaningful 0), but it is discrete.

1d) (*1 point*) Are all ordinal attributes discrete? If so, explain why, and if not, give a counterexample.
Solution: Yes. One simple way of thinking about it is: If an ordinal attribute were continuous, we could map it to the real-valued numbers, and therefore calculate its average, but we know we cannot do this with ordinal attributes. Because ordinal values map to integers (i.e. they are discrete), we cannot compute the average (e.g., the average of 1 and 2 is 1.5, which is not an integer).

2 Sampling (10 points) [Benyamin Tabarsi]

2a) (*6 total points*) State the sampling method used in the following scenarios and give a reason for your answer. Choose from the following options: simple random sample with replacement, simple random sample without replacement, stratified sampling, progressive/adaptive sampling.

i) (*2 points*) You are conducting a study to create a predictive model with ultra-high definition (UHD) videos as the dataset. The samples are large in size, and therefore difficult to collect and store. Therefore, your research team has suggested you continue sampling until you reach 85% accuracy on the validation data set.

ii) (*2 points*) You are investigating the novice programmers' interest in the language they are learning. They can be divided into three groups based on the programming language: Java, Python, and R. The number of students in each group is different, but you are required to create three sets of the equivalent population in your sample.

iii) (*2 points*) You are running a study to find the median and mean time spent studying for undergraduate students in the CS department. You plan to have 100 participants in the study, and no one can participate more than once.

2b) (*4 total points*) The U.S. Congress is made up of 2 chambers: 1) a Senate of 100 members, with 2 members from each state, and 2) a House of Representatives of 435 members, with members from each state proportional to that state's population. For example, Alaska has 2 Senators and 1 House representative, while California has 2 Senators and 53 House representatives. Both the Senate and the House are conducting surveys of their constituents, which they want to reflect the makeup of each chamber. You suggest that they use stratified sampling for this survey, sending surveys to a certain number of people from each state. Each survey will be sent to 1200 participants.

i) (*1 point*) Why is stratified sampling appropriate here?

ii) (*1 point*) For the Senate survey, how many surveys would you recommend sending to people in Alaska?

iii) (*1 point*) For the House survey, how many surveys would you recommend sending to people in California?

iv) (*1 point*) What are some advantages of the "Senate" approach and the "House" approach to stratified sampling?

Solution:

2a) i) Adaptive or Progressive sampling starts with a small sample and then increases the sample size until a sample of sufficient size gets obtained. The question mentions that we continue sampling until we reach the desired accuracy, so we most likely do not require all available data to meet the expected sensitivity requirement.

ii) Stratified sampling is able to do the sampling by dividing the data into subgroups considering some attribute (e.g. programming language).

iii) Simple random sampling without replacement would be appropriate for this situation since we are told that a population element cannot be selected more than one time.

2b) i) A stratified sampling scheme accommodates differing frequencies for the items of interest. For example, since they want the sample to reflect the makeup of each chamber, then a different number of surveys will be sent to each state. Stratified sampling allows this.

ii) 24 Surveys. Each state gets equal number of Senate surveys

iii) Approximately 146 surveys i.e. $(53/435) * 1200$. We want the number of surveys sent to Florida to be proportional to its representation in the House, which is 53/435. Since there are 1200 surveys, we multiply $(53/435) * 1200$.

iv) The advantage of the "Senate" approach is that each state is equally-well represented, meaning any conclusions from the surveys will apply to people from all states. The advantage of the "House" approach is that the conclusions we gain from the surveys will apply well across the whole of the U.S. Both approaches have the advantage of ensuring that there is some representation from all states, no matter how surveys are mailed out.

3 Discretization (12 points) [Benyamin Tabarsi]

Consider the following dataset:

#	GENDER	AGE	GLUCOSE LEVEL	BMI	STROKE
1	Male	71	153	21	no
2	Female	14	57	31	yes
3	Female	8	110	17	no
4	Female	78	79	19	yes
5	Female	68	247	40	yes
6	Male	57	85	37	yes
7	Male	71	88	35	yes
8	Male	57	197	34	yes
9	Female	70	69	36	no
10	Male	58	88	39	no
11	Female	52	77	17	no
12	Male	68	233	42	yes
13	Female	75	80	29	yes
14	Male	3	95	18	no
15	Male	14	161	19	no

3a) (*3 points*) Discretize the attribute AGE by binning it into 5 equal-width intervals (the range of each interval should be the same, and they should collectively span from the minimum to maximum values). Show your work by writing intervals for each bin.

Solution:
Bin 1: [3, 18) $\rightarrow Data : \{3, 8, 14, 14\}$
$Bin2 : [18, 33) \rightarrow Data : \{\}$
$Bin3 : [33, 48) \rightarrow Data : \{\}$
$Bin4 : [48, 63) \rightarrow Data : \{52, 57, 57, 58\}$
$Bin5 : [63, 78] \rightarrow Data : \{68, 68, 70, 71, 71, 75, 78\}$

3b) (*3 points*) Discretize the attribute GLUCOSE LEVEL by binning it into 5 equal-depth intervals (the number of items in each interval should be the same). Show your work.

Solution:
Bin 1: [57, 78) $\rightarrow Data : \{57, 69, 77\}$
$Bin2 : [78, 86) \rightarrow Data : \{79, 80, 85\}$
$Bin3 : [86, 96) \rightarrow Data : \{88, 88, 95\}$
$Bin4 : [96, 162) \rightarrow Data : \{110, 153, 161\}$
$Bin5 : [162, 247] \rightarrow Data : \{197, 233, 247\}$

3c) (*3 points*) Consider the following new approach to discretizing a numeric attribute: Given the mean ($\bar{x}$) and the standard deviation (σ) of the attribute values, bin the attribute values into the following intervals: $[\bar{x} + (k-1)\sigma,\ \bar{x} + k\sigma)$,
for all integer values k, i.e. $k = \ldots - 4, -3, -2, -1, 0, 1, 2 \ldots$
Assume that the mean of the attribute BMI above is $\bar{x} = 29$ and that the standard deviation $\sigma = 9$. Discretize BMI using this new approach. Show your work.

Solution:
Bin 1: [11, 20) $\rightarrow Data : \{17, 17, 18, 19, 19\}$
$Bin2 : [20, 29) \rightarrow Data : \{21\}$
$Bin3 : [29, 38) \rightarrow Data : \{29, 31, 34, 35, 36, 37\}$
$Bin4 : [38, 47] \rightarrow Data : \{39, 40, 42\}$

3d) (*3 points*) Give an example of a situation where you would want to use equal-width binning, rather than equal-frequency.

Solution: Equal width is better for graphical representations (especially histograms), e.g. If we want to know the population from each age level, we may always make equal-width binning as 20-30, 30-40, 40-50, regardless of whether the data is evenly distributed or not.

4 Decision Tree Construction (18 points) [Chengyuan]

Create decision trees **by hand** for the `hw1_dt.csv` Titanic survival dataset, as explained below, using Hunt's algorithm. Note the following:

- In the given dataset, all of the input attributes are binary except for the "Pclass" which is categorical. "Pclass" should create a 3-way split if used in the tree.
- The output label has two class values: T or F, which represent Survival or Not Survival.
- In the case of ties when selecting an attribute, break ties in favor of the leftmost attribute.

Showing your work: You must fully show your work when calculating Information Gain or Gini Index for **the split at the root node** (but not for later splits). You can do so by either 1) writing out subseps (e.g. conditional entropy for each child node), or including code for a program you used to make your calculations.

You should draw the final tree for each problem. You can use a program (e.g. tikz with LaTeX, Lucidchart, draw.io, etc.) to draw your trees, or draw them by hand on paper and scan your results into the final pdf.

4a) (*8 points*) Construct the decision tree manually, using Gini index to select the best attribute to split on. The maximum depth of your tree should be 2 (count the root node as depth 0), meaning that any node at depth 2 will automatically be a leaf node, even if it has objects with different classes.

Solution: Refer to "hw1_dt_solution.pdf".

4b) (*5 points*) Construct the tree manually using Information Gain. The maximum depth of the tree should be 1.

Solution: Refer to "hw1_dt_solution.pdf".

4c) (*1 point*) Give an example of a data object (either in the training dataset or not) that would be classified differently by the two trees.

Solution: Given a test case {Pclass = Upper, Sex = Male, Embarked = Queenstown}, The IG tree will predict it as T while the Gini tree will predict it as F.

4d) (*3 points*) Use the training dataset (`hw1_dt.csv`) to calculate accuracy for each tree (% of instances correctly classified). Which decision tree will perform better on the training dataset?

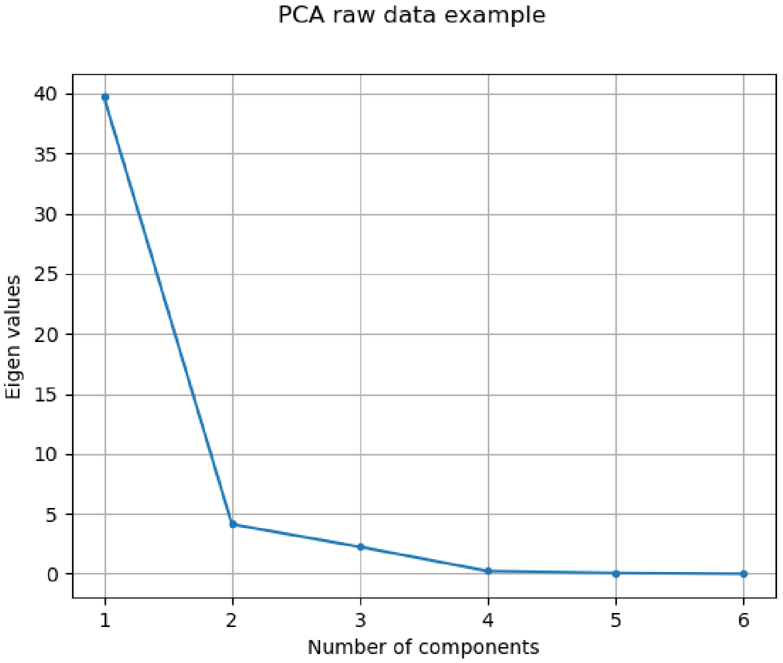
Solution: Accuracy for Gini tree is 13/16=81.25%, for IG tree is 11/16=68.75% (regardless of the tie in the Pclass=Middle branch since there are 3 correctly classified data points for both cases), therefore the Gini tree performs better in the training dataset.

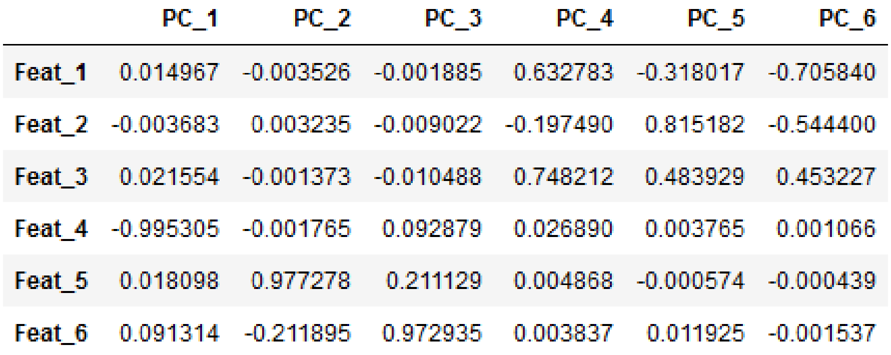
4e) (*1 point*) Which will perform better on a test dataset? Can we know the answer?

Solution: We cannot know the answer because the test dataset may be different than the training dataset.

5 Dimensionality Reduction (10 points) [Jianxun WANG]

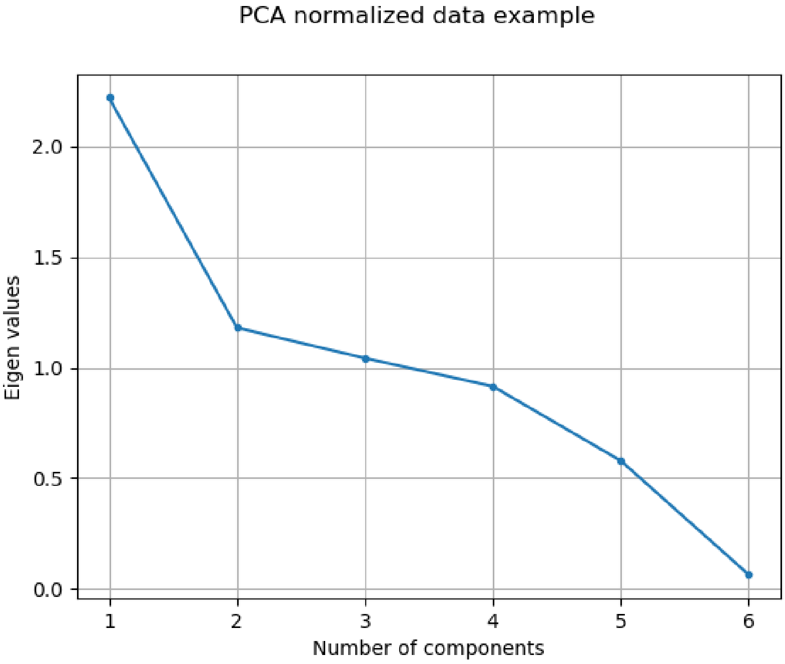
In this problem, you will analyze the PCA results on a subset of a *Wearable Sensor* dataset. Figure 1 shows the Eigenvalue Scree plot and the eigenvectors of the principal components of PCA analysis on the raw dataset. The dataset was then normalized using z-scores, and Figure 2 shows the Eigenvalue Scree plot and the eigenvectors of the principal components of PCA analysis on dataset *after* normalization. Note that $Feat\_i$ represents a feature for $i = 1 \ldots 6$ and PC_j represents a Principal Component for $j = 1 \ldots 6$.

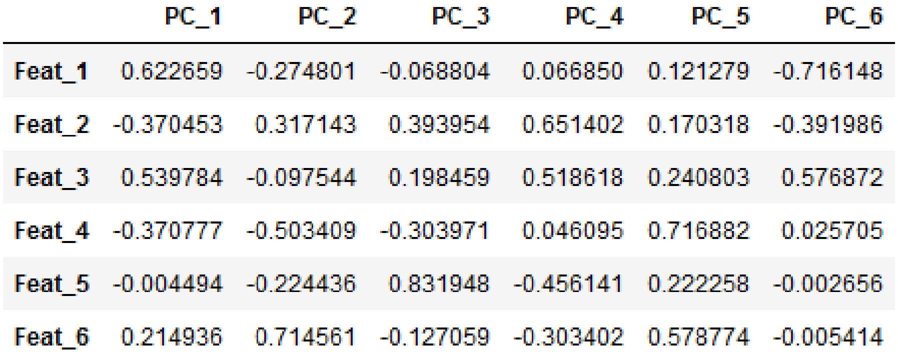




	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
Feat_1	0.014967	-0.003526	-0.001885	0.632783	-0.318017	-0.705840
Feat_2	-0.003683	0.003235	-0.009022	-0.197490	0.815182	-0.544400
Feat_3	0.021554	-0.001373	-0.010488	0.748212	0.483929	0.453227
Feat_4	-0.995305	-0.001765	0.092879	0.026890	0.003765	0.001066
Feat_5	0.018098	0.977278	0.211129	0.004868	-0.000574	-0.000439
Feat_6	0.091314	-0.211895	0.972935	0.003837	0.011925	-0.001537

Figure 1: PCA1 on Raw Dataset





	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
Feat_1	0.622659	-0.274801	-0.068804	0.066850	0.121279	-0.716148
Feat_2	-0.370453	0.317143	0.393954	0.651402	0.170318	-0.391986
Feat_3	0.539784	-0.097544	0.198459	0.518618	0.240803	0.576872
Feat_4	-0.370777	-0.503409	-0.303971	0.046095	0.716882	0.025705
Feat_5	-0.004494	-0.224436	0.831948	-0.456141	0.222258	-0.002656
Feat_6	0.214936	0.714561	-0.127059	-0.303402	0.578774	-0.005414

Figure 2: PCA2 on Z-score Normalized Dataset

Please answer the following questions:

5a) (*4 points*) Based on the table and figure in **Figure 1**, do you think that performing PCA was useful? Why or why not? If not, what properties of the dataset caused PCA to be less useful?

Solution: No, since within each PC there is an unique attribute which has a very high portion of the PC's variance. Under such circumstance, PCA might result in discarding the very important information. The high eigen values also indicate that the data is not normalized in comparison to the eigen values after normalization in Figure 2.

5b) (*4 points*) In **Figure 2**, what is the most reasonable number of principal components to retain for dimensionality reduction? Briefly justify your choice. *Hint*: There may be more than one reasonable answer. **Solution**: Possible answer 1: PC1 only, since it explains considerably more variance than PC2 (but PCs 2, 3 and 4 all explain a similar, small amount of variance); Possible answer 2: PC1, PC2, and PC3, since they are with eigenvalue larger than 1;

5c) (*2 points*) Consider 1 instance/row in the dataset, called A, and the PCs in Figure 2 (after normalization). If we increased A's value for Feat_4 by 2 and decreased its value for Feat_5 by 2, *after normalization*, how would you expect A's value for PC2 to change? Would it increase, decrease, or stay the same? Briefly justify your answer.

Solution: **Decrease**: $PC_2 = \ldots + (-0.5034) * Feat_4 + (-0.2244) * Feat_5 + \ldots$ If $Feat_4$ is increased by 2 and $Feat_5$ decreased by 2, the change is $\Delta PC_2 = -0.5034 * 2 + -0.2244 * (-2) = -0.558$. Therefore, PC_2 will decrease.