

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, *discrete* 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
 - ii) (1 point) Annual income (in US dollars).
 - iii) (1 point) Existence or non-existence of cancerous tumors.
 - iv) (1 point) Level of Pain during Injury diagnosis (0: absent, 1: mild, 2: moderate, 3: severe, 4: incapacitating)
 - v) (1 point) The pH of water that has been measured using the pH scale².
 - vi) (1 point) Categorization of clothing (e.g., hat, shirt, pants, shoes).
 - vii) (1 point) Sugar content in juice in grams.
 - viii) (1 point) Calendar dates.
 - ix) (1 point) The response to a yes-no question.
 - x) (1 point) Product Satisfactory Rating (5 - Excellent, 4 - Good, 3 - Fair, 2 - Poor, 1 - Horrible).
- 1b) (1 point) Are all continuous attributes ratio? If so, explain why, and if not, give a counterexample.
- 1c) (1 point) Are all ratio attributes continuous? If so, explain why, and if not, give a counterexample.
- 1d) (1 point) Are all ordinal attributes discrete? If so, explain why, and if not, give a counterexample.

¹Binary attributes are a special case of discrete attributes.

²**Note:** A pH of 7 is considered neutral, with < 7 being acidic and > 7 being basic.

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?

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.
- 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.
- 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. $k = \dots - 4, -3, -2, -1, 0, 1, 2 \dots$. 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.
- 3d) (3 points) Give an example of a situation where you would want to use equal-width binning, rather than equal-frequency.

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 L^AT_EX, 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.
- 4b) (*5 points*) Construct the tree manually using Information Gain. The maximum depth of the tree should be 1.
- 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.
- 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?
- 4e) (*1 point*) Which will perform better on a test dataset? Can we know the answer?

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 $i = 1 \dots 6$ and PC_j represents a Principal Component for $j = 1 \dots 6$.

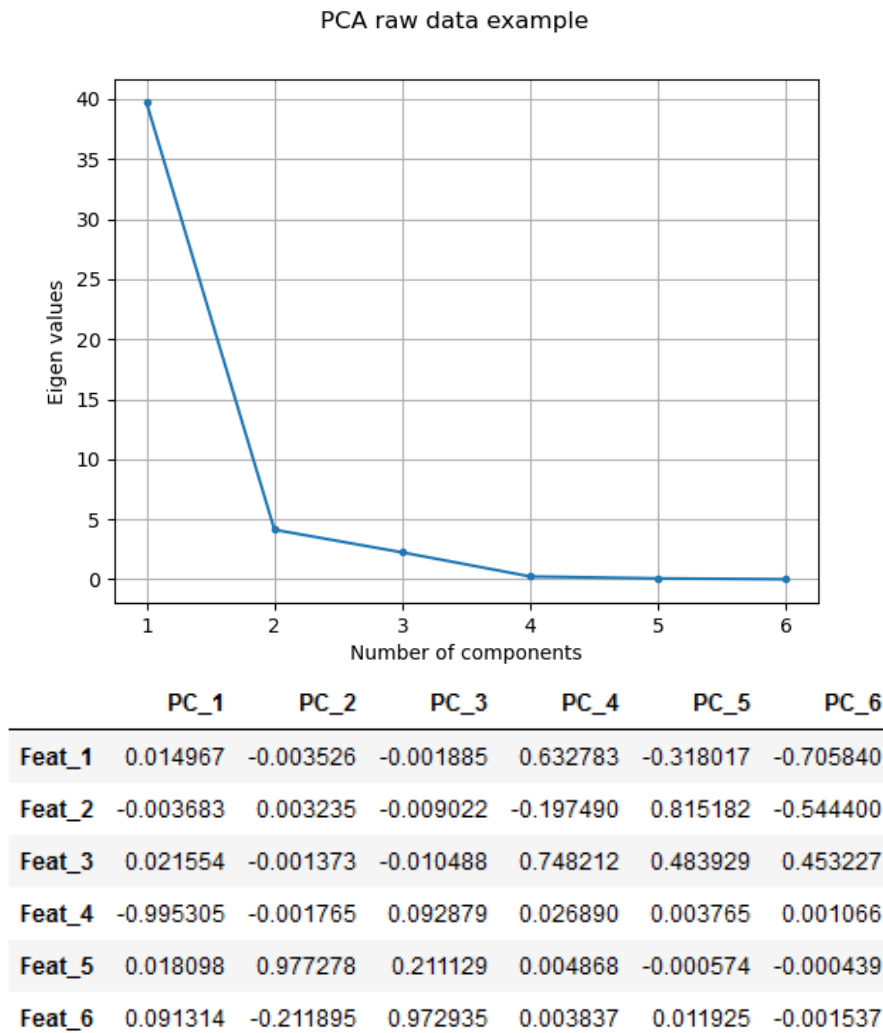


Figure 1: PCA1 on Raw Dataset

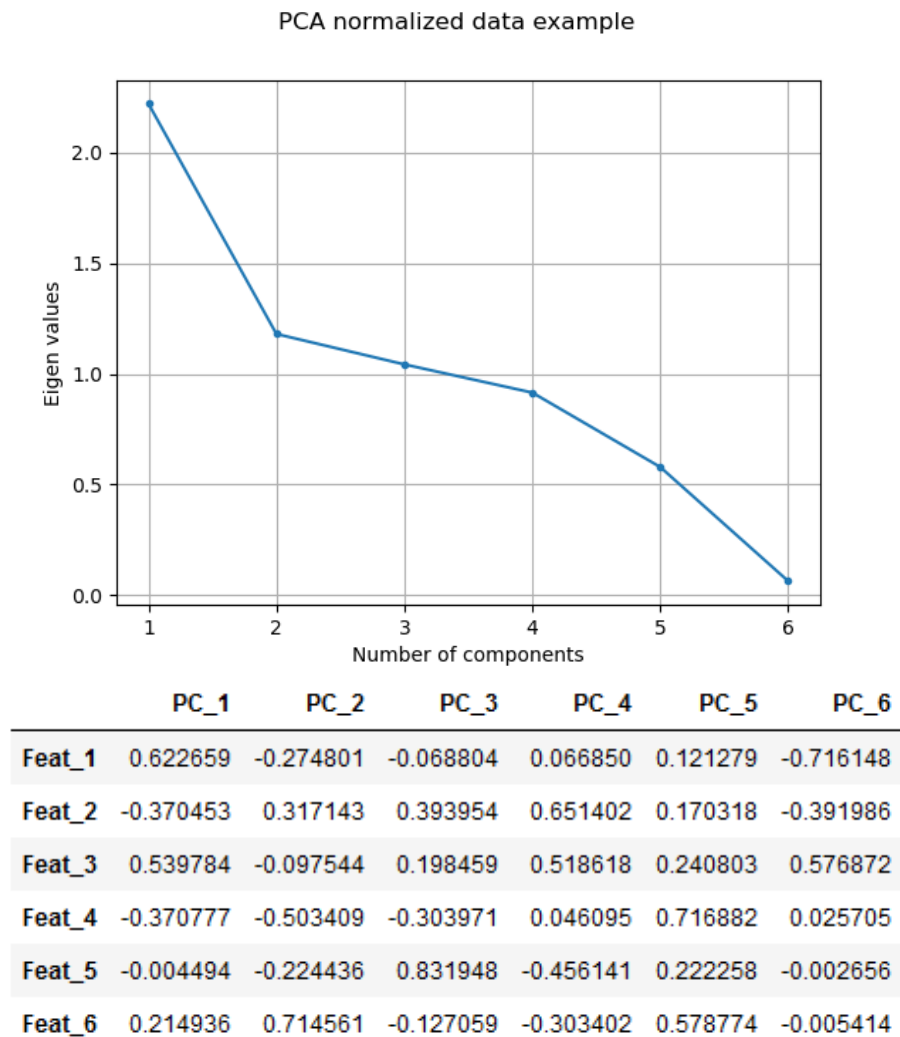


Figure 2: PCA2 on Z-score Normalized Dataset

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?
- 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.
- 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.