Final exam

Started: Nov 17 at 11:33pm

Quiz Instructions

The time limit will be 6 hours but it will not take that much! Most likely it will probably not take more than 1 or 2 hours.

Still, be sure to allocate at least couple of hours of distraction-free environment before taking the exam. There will be only one attempt, so answer the questions carefully!

Question 1 — 1 pts

ISOMAP is designed for unfolding complex nonlinear manifold. Therefore, it cannot produce anything useful when the data is neatly distributed on a linear subspace.

- (●) True
- ( ) False

Question 2 — 1 pts

You need to run PCA many times and compare the results to obtain full picture.

- ( ) True
- (●) False

Question 3 — 1 pts

According to the Gestalt principles, grouping by 'connections' or 'enclosures' is a stronger way to group objects than by 'proximity' or 'similarity'.

- (●) True
- ( ) False

Question 4

1 pts

Here is a list of properties of trees. Check "true" if ALL descriptions are correct.

- Trees do not have any loops.
- Trees are automatically 'planar' graphs.
- In a tree, there is only a single shortest path between two nodes.

- (●) True
- ( ) False

Question 5

1 pts

A good, carefully chosen 2D projection can preserve all of the important properties such as shapes, areas, and distances, of the actual globe.

- ( ) True
- (●) False

Question 6

1 pts

When you pick colors to represent categories, it is much better to just vary in one of the dimensions (e.g., brightness or hue) to avoid confusion.

- [x] True
- [ ] False

Question 7

1 pts

The bag of words model represents a document with by the counts of each word that appear in the document. In doing so, it destroys any information contained in the 'sequence' in which the words apprear.

- [x] True
- [ ] False

Question 8

1 pts

In a histogram, the quantity that you want to show is always represented by the **height** of each bar.

- [ ] True
- [x] False

Question 9

1 pts

Empirical CDF (and CCDF) is an information-dense visualization method because it makes use of every single data point in the dataset.

- [x] True
- [ ] False

Question 10 — 2 pts

Which one of the following is NOT the rationale for avoiding colors?

- ○ Color perception is not as accurate as other visual encodings.
- ○ Using color can create artifacts, often emphasizing somewhat arbitrary range of values.
- ◉ Using color does not help conveying useful information.
- ○ Using color may limit accessibility (e.g. color blindness)

Question 11 — 2 pts

To pick categorical colors, we would like to have a large variance of several color dimensions. Which one is the LEAST CRITICAL or NOT RELEVANT?

- ○ varying lightness
- ○ varying hues
- ◉ varying the red-green content

Question 12 — 2 pts

Which of the following colormaps do you want to use for your quantitative data (in most cases)?

- ○ Jet
- ◉ Viridis
- ○ Rainbow
- ○ Parula

Question 13

2 pts

Which of the following is a BAD reason for the usage of colors for quantitative data?

- (●) If you use colors, it will not print well.
- ( ) Color is less accurate than other visual encodings such as length or position.
- ( ) Color perception may largely depend on surrounding colors.
- ( ) Colors, especially variance in hue, can create visual artifacts.

Question 14

2 pts

What is the correct explanation of the Mercator projection?

- ( ) It exaggerates the area of countries near equators compared with those near the poles.
- ( ) It preserves the area of every region.
- ( ) It preserves the distance between every pair of locations.
- (●) It shows every constant-bearing path as a straight line.

Question 15

2 pts

In KDE, we put a "kernel" onto each data point. Imagine a rectangular kernel with (band)width of 0.5. If you have 10 data points, what would be the height of each kernel that we are adding to obtain the KDE?

- ( ) 0.1
- ( ) 2

- ○ 1
- ◉ 0.2

Question 16 — 2 pts

Stevens' power-law states that

- ○ The just-noticeable difference (JND) is proportional to the original intensity
- ○ We must always use certain visual encodings
- ○ Length perception is highly inaccurate.
- ◉ For each sense, there is a power-law relationship between the intensity of the sense and the perception of it.

Question 17 — 2 pts

Which of the following is NOT the primary purpose of data tidying process?

- ○ To reuse data cleaning and manipulation tools
- ○ To facilitate data cleaning
- ○ To facilitate initial data exploration
- ◉ To reduce the size of the dataset to perform analysis more efficiently

Question 18 — 2 pts

When choosing categorical colors, which of the following would be the best practice?

- ◉ Varying the brightness across categories.

- ○ Use colors along the red-green axis to have the most contrast.
- ○ Keeping the brightness similar across categories
- ○ Stick to a single hue as much as possible.

Question 19

2 pts

What is the key assumption of dimensionality reduction process?

- ○ We can always find a good low-dimensional representation of the dataset, regardless of how the data is distributed.
- ○ The data is more or less uniformly distributed across the whole space.
- ◉ The data distribution can usually be approximated using a low-dimensional subspace.

Question 20

2 pts

Which of the following visual encodings (for quantitative information) is known to be most accurate?

- ◉ Length
- ○ Angle
- ○ Volume
- ○ Area

Question 21

2 pts

Which of the following data can be most useful to visualize by a Voronoi diagram?

- ○ Classroom buildings on IU campus

- [x] Gas stations in Bloomington
- [ ] Historical tax rate
- [ ] Earthquake risk in California's counties

Question 22

2 pts

Which one of the followings is **worse** than other criteria of designing a good colormap?

- [ ] Accessible to color blind viewers
- [ ] Perceptually uniform (accurately representing the data)
- [x] Using the full spectrum of hues
- [ ] Accurate representation when printed

Question 23

2 pts

In estimating density of a distribution, the parametric approach is likely to be better than non-parametric approach if ...

- [x] You already have a pretty good idea and a model (parametric distribution) that explains your dataset.
- [ ] The data is distributed across many orders of magnitude.
- [ ] You have no idea about how the distribution look like
- [ ] The dataset has many outliers.

Question 24

2 pts

In log-scale, the apparent distance between 1 and 100 is same as the distance between 50 and

- (●) 50000
- ( ) 149
- ( ) 0.5
- ( ) 500

Question 25

2 pts

What is NOT the reason against using colors for quantitative data?

- ( ) Color is less accurate than other visual encodings such as length or position.
- ( ) Color perception may largely depend on surrounding colors.
- ( ) Colors, especially variance in hue, can create visual artifacts.
- (●) If you use colors, it will not print well.

Question 26

4 pts

You are using the Kernel Density Estimation method with a rectangular kernel to estimate the underlying distribution of your data (X = [5, 1, 3, 1, 2, 3, 5, 5, 6, 5]). The width of the rectangular kernel is 1.2. If we examine the resulting distribution, what would be the area under the distribution/curve (or the estimated probability mass) that is within the range [0, 4]?

6.94

Question 27

4 pts

You received a dataset about the number of items sold for each of your inventories. It's essentially a sorted list of numbers like this: X = [1, 1, 1, 2, 2, 5, 10, 14, 101, 252]. You decided to use empirical CDF to visualize the data distribution. You normalize your CDF so that the CDF accumulates to 1.0 (the largest value becomes 1.0), then what would be the value of this CDF at x = 100?

0.8

Question 28

4 pts

"Values are organised in two ways. Every value belongs to a _____ and an observation. A _____ contains all values that measure the same underlying attribute (like height, temperature, duration) across units."

variable

Quiz saved at 12:44am

Submit Quiz