DAV exam

24 Jan 2020

Intro

Welcome to the exam! You have from 13.30-16.00 to take it. If, for whatever reason, you require extra time, you may use an additional half-hour and finish by 16.30.

This exam is "open book", with limits:

- You may use: your notes, as well as use the internet, to look up existing information.
- You may use your own written course notes.
- You may not: communicate, in any way, with others.
- Please ask Maarten, Ayoub, or Erik-Jan if you are in doubt. In the interest of time and fairness, any decision made by them is non-negotiable.

Procedure:

- Answer the exam questions by including them in an .Rmd file. A template .Rmd file is provided in this directory.
- Brevity is the soul of wit. In other words, our feelings about your answer are inversely proportional to the number of words used. Obviously, do use enough words to explain your answer...
- Go to the DAV course Blackboard page. Under Assignments, you will find "Exam DAV 2019-2020". Upload your exam here, following the instructions. If you have chosen to write on paper, turn in the paper with Maarten, Ayoub, or Erik-Jan.
- Please note: you only have **one try** to submit on Blackboard the above, so please double- and triple- check your submission.
- Any late submissions on Blackboard will not be admitted. Submissions after you have left the room, or made remotely from a different location will invalidate your exam.
- The points given with each question are provided as an indication only. We reserve the right to change the weighting of questions at a later stage.
- Good luck!

EXAM STARTS ON FOLLOWING PAGE

Q1. Visualization

For each of the following two graphics (Figure 1 and Figure 2), answer these questions:

- Q1(a): Name its aesthetics, geoms, scales, and facets. That is, name the variables to which various aesthetics etc. are mapped. Also name any statistical transformations or special coordinate systems. [1 point]
- Q1(b): Give at least three suggestions for improvement of the visualization, and explain your rationale. [1 point]

Figure 1

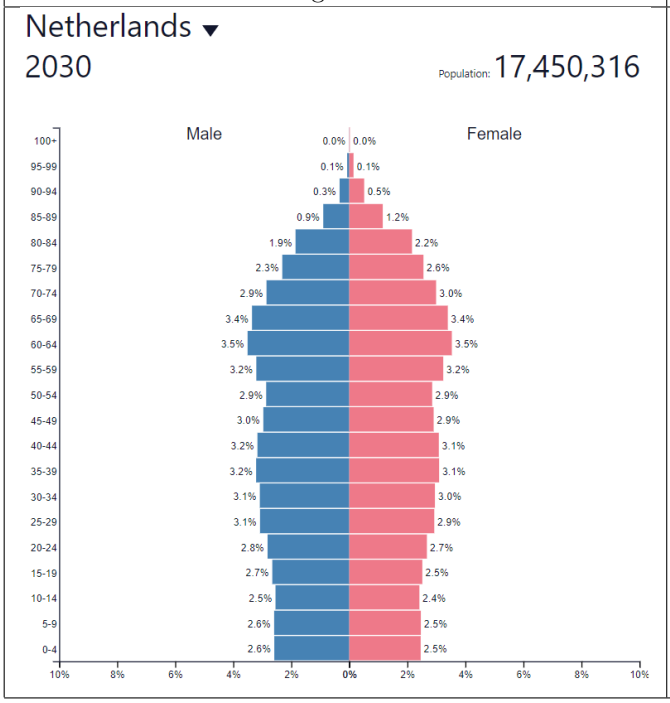
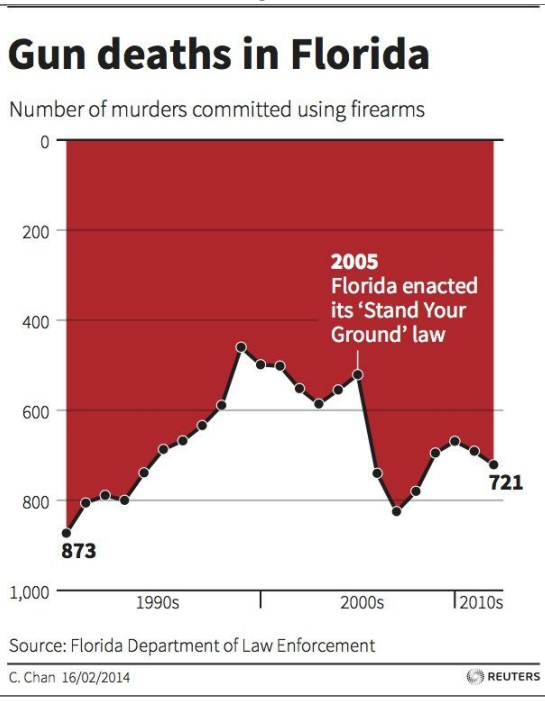


Figure 2



Theoretical questions

Q2. Suppose you are given a **supervised** regression task. There are many possible models that could accomplish such a task. Which one is the best? [1 point]

Q3. Suppose you are given an **unsupervised** task. There are many possible models that could accomplish such a task. Which one is the best? [1 point]

Q4. For each of the following supervised learning situations, indicate which of the following feature selection/regularization techniques are most likely to be useful. Explain your rationale. [3 points]

Several answers may apply. You may not need to mention all four techniques a-d.

Techniques:

a. variance filter;
b. correlation filter;
c. random forest-based wrapper;
d. L2 penalization.

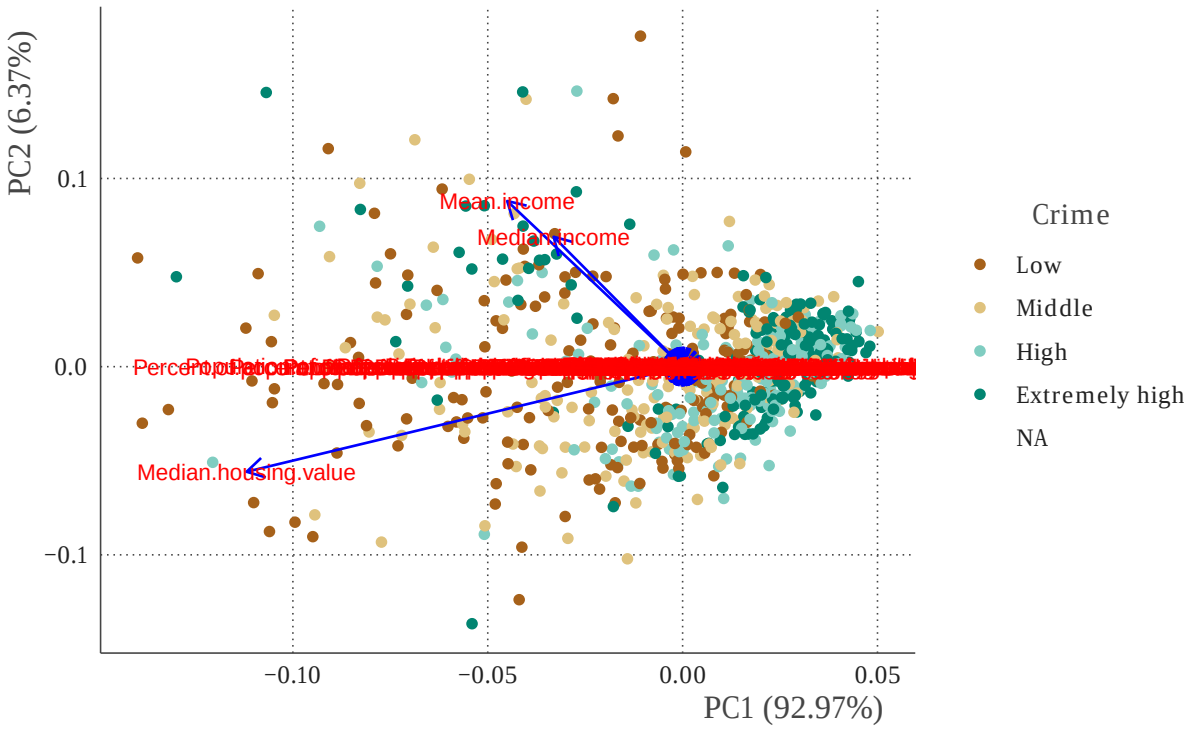
Situations:

1. A raw, uncleaned dataset with many noisy features, many of which are mostly zero;
2. A dataset in which many features are almost the same (e.g. age and birth year);
3. The client wants you to run a linear regression and is interested in unbiased estimates of the coefficients.

Q5.

See the below PCA biplot of Chicago neighborhoods. (The center contains some illegible variable names; you can ignore these.) [3 points]

- Q5a. How much of the total variation in the neighborhood characteristics do the first two principal components explain?
- Q5b. Is income positively or negatively associated with crime? How can you tell from the biplot?
- Q5c. Which explains more crime variance: median income or median housing value? How can you tell?



Practical questions: bee health


Honeybees live in colonies. In each colony, a percentage of the individual bees in the colony die of natural causes each winter. This percentage is called the "winter mortality". A normal winter mortality is usually considered to be around 10 percent. In recent years in some places in the world, the winter mortality appeared to be exceptionally high. It may also sometimes happen that bees do not return to their hive at all, in which case the mortality is 100 percent. This phenomenon is especially worrying to beekeepers and is known as "colony collapse disorder".

Beekeepers keep colonies together in a group, called an *apiary* (see above). Here you are presented with a dataset, `bees.csv`, on 4758 apiaries across Europe. Among the 38 features are winter mortality (`Winter_mortality`), as well as features related to the beekeeper (for example, whether they are professional or amateur, `Activity`), the apiary (for example, the apiary size in five categories, `Apiary_Size`, or whether the colonies were merged to create new ones, `Merger`), and the bees (for example, whether they were infected with the Varroa mite, `VarroosisV1`, or with the larvae of the bacterium Paenibacillus, `AmericanFoulbroodV1`).

1. Load the dataset `bees.csv` and split it into a training (80%) and test set (20%). Use the training set for EDA and model development. [1 point]
2. Which country suffered the highest winter mortality this year? [1 point]
3. Which country experienced the highest rate of CCD? (HINT: use `Winter_mortality` to work out if CCD occured) [1 point]
4. Use logistic regression to predict CCD from the other features in the dataset. (HINT: add the previously calculated feature to the dataset) [2 points]
5. Show the confusion matrix of predictions from your model, for the test data. [2 points]
6. What are the accuracy, recall, and precision of your model? [1 points]
7. COMPLETELY *OPTIONAL* HYPERBONUS: compare the test performance of the following models: L2-penalized logistic regression, random forest, and gradient boosting. Performance should be evaluated in terms of both AUC and calibration.