PS3192 Coursework Information

Matteo Lisi

Table of contents

General information

Create a portfolio by answering any three of the five questions below. The total word limit is 5,000 words, which is an upper limit rather than a target. Figures, tables, and other visual elements are often just as informative as text, so you are not expected to reach the full word count if you can adequately convey your analyses more succinctly.

- Appendices (R scripts, Matlab code, etc.) do not count toward your word limit.
- All datasets will be made available on the PS3192 Moodle page (some are also part of specific R packages as indicated).
- You must submit a single Word document containing:
  - Your main report (with text, figures, and tables).
  - Appendices (containing all of your code).
- All code must be included in your appendices.

Question 1: Data Visualization

Dataset

Use the dataset `college_recent_grads` from the fivethirtyeight R package.This dataset contains data on recent college graduates, including median salaries, unmeployment rates, major categories, and gender composition, among other variables. You can load the dataset in R as follow:

```
# install.packages("fivethirtyeight") # install the package if necessary
library(fivethirtyeight)
data(college_recent_grads)
```

Use `?college_recent_grads` for variable descriptions, or refer to the CSV file and data dictionary provided on Moodle if you prefer working outside of R packages.

Task

- Create at least one well-designed visualization (e.g., via `ggplot2`) to illustrate an interesting relationship or trend within the data.
- Write a short descriptive caption that:
  - Explains the variables shown in the plot and how they relate.
  - Highlights any patterns, trends, or outliers.
  - Offers a brief interpretation of the pattern seen in the data and briefly explain why the insights gained from the figure might be important.

Note on best practices: When possible, use visualizations that more transparently convey the data's distribution, such as boxplots, violin plots, or raincloud plots.

What to include

- **Main Text:**
  - Present and describe the plot within your main report.
  - Include the caption as part of your text or as a figure caption.
- **Appendix:**
  - Provide all relevant R code used to clean, prepare, and visualize the data.

Question 2: Machine learning concepts

Write a short essay on core ML concepts discussed in class, including:

- Overfitting
- Out-of-sample prediction
- Cross-validation

In particular you should describe why overfitting is a concern in machine learning and how cross-validation (CV) is used to estimate true predictive performance on unseen data. Discuss potential pitfalls where one might still overfit even when using CV (e.g., “feature hacking”). Reflect on practical steps to mitigate these pitfalls.

What to include

- **Main Text:**
  - Your short essay addressing the above points.
- **Appendix (if needed):**
  - If you choose to provide small code snippets or references to illustrate your points (optional), you may include them here.

Question 3: Supervised Learning

Dataset

Select one of the following classification datasets (each has a binary or multiclass outcome):

- `titanic`
- `heartdisease`
- `oiltype`
- `forest_mapping`
- `autism`
- `banknote_authentication`

(See the dataset information document on Moodle for details on variables and any required preprocessing.)

Task

Predict a discrete outcome (e.g. survived vs. not survived) using a technique such as logistic regression, or decision tree (e.g., via `rpart`). Fit at least one classification model (although you may fit more than one classification model if you wish to compare their performance.). Evaluate performance using a confusion matrix and at least one additional metric (e.g., accuracy, etc.). Provide at least one plot. Describe and what steps you have taken to minimise the risk of overfitting and assess or improve the model's out of sample predictive performance.

What to include

- **Main Text:**
  - Data description: what is the outcome variable being predicted, what variables are used as predictors.
  - Summarize your model-building process, including any data cleaning.
  - Evaluate performance and discuss the model's strengths and weaknesses.
- **Appendix:**
  - Include all R code (data cleaning, modeling, and plotting).

Question 4: Unsupervised Learning

Dataset

Use a dataset suitable for clustering with a Gaussian Mixture Model (GMM). For instance:

- `Iris` (ignoring the `species` column, thus treating the data as unlabeled)
- `faithful`
- `gazedata`
- `wine`

Task

Perform a Gaussian Mixture Model analysis (e.g., using `mclust` in R). Determine a suitable number of clusters (e.g. using BIC or another criterion). Plot your clusters (in a 2D or pairwise plot) to visualize the grouping.

What to include

- **Main Text:**
  - Briefly describe the dataset and the concept of a GMM.
  - Summarize how many clusters were chosen and why.
  - Provide interpretive text: what do the clusters may represent about the data?
- **Appendix:**
  - Include R code for your GMM fitting and any plotting routines.

Question 5: MVPA & Matlab

1. Explain what is meant by MVPA and how it can be used to analyse fMRI data
2. MVPA and classification are known as "supervised learning" algorithms. Explain what is meant by supervised in this context.
3. Linear discriminant analysis and support vector machines are two common algorithms for MVPA. Explain one of the main differences between these two algorithms and under what circumstances one or the other would be preferable?

Use data set 2 (in the `matlab_dataset` folder in Moodle) to answer the questions below. Details about the data set are provided in a separate document. Load the data into the MVPA viewer first the same way as you did in class.

4. Run the classification with default parameters (using linear discriminant analysis as classification algorithm). Report the results, explaining the approach of the algorithm and how you arrived at your conclusions.
5. Re-run the classification but use SVM as the classification algorithm instead. Briefly summarise how the results differ from the default algorithm (linear discriminant analysis).
6. Re-run the classification with different R2 voxel inclusion thresholds and make note of the result for each threshold (try the following thresholds: 0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40). Make a graph showing how classifier performance varies as a function of R2 threshold. Explain how the choice of voxel threshold affects the result.
7. Set the voxel R2 threshold to 0 and re-run the classification with different voxel number thresholds (try 10, 25, 50, 100, 200, 300, 400). Make a note of the result for each voxel number and make a graph showing how classifier performance varies as a function of the number of voxels. Explain how the number of voxels can affect the result. Include the graphs and the commands used to generate them in your report.