# Assignment 2

STAT30270 – Statistical Machine Learning

**Deadline - Tuesday 23rd April at 5pm**

## Exercise 1

Consider some two-dimensional input observations $\mathbf{x}_i = (x_{i1}, x_{i2})$ and the associated target variable $y_i \in \{-1, +1\}$. A soft-margin support vector classifier has been trained on these data. The associated separating hyperplane is given below:

$$\{x_1, x_2 \,:\, 0.5 - 3x_1 + x_2 = 0\}$$

Consider the following subset of data points $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, 5$:

| $\mathbf{x}_i$ | $x_{i1}$ | $x_{i2}$ | $y_i$ |
|---|---|---|---|
| $\mathbf{x}_1$ | 1/6 | 3/2 | +1 |
| $\mathbf{x}_2$ | 1/3 | -1/2 | -1 |
| $\mathbf{x}_3$ | 1/2 | 2 | +1 |
| $\mathbf{x}_4$ | 1/3 | 1 | +1 |
| $\mathbf{x}_5$ | 1/4 | 3/4 | -1 |

Which data points identify the support vectors? Justify your answer. *(10 marks)*

## Exercise 2

A k-means algorithm with $K = 2$ is employed to cluster the following five 3-dimensional observations stored in the data matrix below:

$$\mathbf{X} = \begin{bmatrix} 1.76 & -1.30 & 2.39 \\ 2.41 & 0.66 & 0.41 \\ -2.63 & 1.18 & -0.36 \\ 0.82 & 0.64 & 1.66 \\ -0.73 & 0.83 & 0.85 \end{bmatrix}$$

At the current iteration $t$, the corresponding estimates of the cluster centroids and the binary indicator matrix of cluster allocations are given below:

$$\hat{\mu}_1^{(t)} = (0.62,\ 0.06,\ 1.63) \qquad \hat{\mu}_2^{(t)} = (-0.11,\ 0.92,\ 0.02)$$

$$\hat{\mathbf{Z}}^{(t)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$

1. Without using software, implement iteration $t + 1$ of the k-means algorithm. Show all calculations and steps.
2. Using calculations and the data in this example, show that the algorithm decreases the value of the within cluster sum of squares moving from iteration $t$ to iteration $t + 1$.

*(20 marks)*

## Exercise 3 – Data analysis

### 2020 US presidential elections

This 2024 is going to be marked by one of the most controversial and divisive US presidential elections, which will see the incumbent president Joe Biden (Democrats) against the opponent Donald Trump (Republican). This will be the first rematch election in decades, since the two have already faced each other in the 2020 presidential elections, with the victory of Biden.

For the curious, some good resources covering the US elections from a quantitative and a critical point of view are below:

- 538 – https://abcnews.go.com/538
- Vox – https://www.vox.com/2024-elections

The dataset `elections` in the file `data_assignment_2.RData` includes voting results data of the 2020 US presidential elections for almost all the US counties. The data also include information regarding the numerical voting results of the 2012 and the 2016 US presidential elections, as well as a number of numerical variables measuring demographic, social, and economic aspects of a county. The data were collected from multiple sources:

- MIT Election Lab – https://electionlab.mit.edu/data
- USA Census Bureau – https://www.census.gov/
- IPUMS – https://usa.ipums.org/usa/

A detailed description of the features in the data is reported below. The number indicates the year in which the record was measured or estimated.

- `state`: State.
- `county`: County name.
- `romney_2012`: Proportion of votes for Mitt Romney in 2012 presidential election.
- `obama_2012`: Proportion of votes for Barack Obama in 2012 presidential election.
- `trump_2016`: Proportion of votes for Donald Trump in 2016 presidential election.
- `clinton_2016`: Proportion of votes for Hillary Clinton in 2016 presidential election.
- `pop_2019`: Population.
- `white_2019`: Percent of white population.
- `black_2019`: Percent of black population.
- `asian_2019`: Percent of Asian population.
- `hispanic_2019`: Percent of Hispanic population.
- `median_age_2019`: Median age.
- `age_over_18_2019`: Percent of population over 18 years of age.
- `age_over_65_2019`: Percent of population over 65 years of age.
- `women_16_to_50_birth_rate_2017`: Birth rate in women 16-50 years of age
- `veterans_2019`: Percent of population who are veterans.
- `bachelors_2019`: Percent of population who earned a Bachelor's degree or higher qualification.
- `uninsured_2019`: Percent of population who are insured.
- `unemployment_rate_2016`, `unemployment_rate_2017`, `unemployment_rate_2019`: Unemployment rates for 2016, 2017, and 2019.
- `median_household_income_2016`, `median_household_income_2017`, `median_household_income_2019`: Median household income for 2016, 2017, and 2019.
- `median_individual_income_2019`: Median individual income.
- `poverty_2019`: Percent of population below poverty line.
- `avg_family_size_2019`: Average family size.
- `household_has_smartphone_2019`: Percent of households that have a smartphone.
- `household_has_computer_2019`: Percent of households that have a computer or laptop.
- `winner`: Indicator for the candidate who obtained the majority of votes for that county, `biden` or `trump`.

## Who will win?

The goal is to build a supervised learning model that can predict whether the winner of the majority for a county will be `biden` or `trump`, using the available numerical features that record previous election results and that summarize economic, demographic, and social aspects of a county.

1. Implement at least three different supervised learning methods to predict the winner of the election in a county based on the numerical input features. Utilize an appropriate framework to compare and tune the different methods considered, evaluating and discussing their relative merits. Select the best model for predicting the winner from the available numerical input features. *(50 marks)*
2. Use appropriately some test data in order to evaluate the generalized predictive performance of the best selected classifier. Provide a discussion about the ability of the selected model at predicting if the winner will be `biden` or `trump`. *(20 marks)*

*Guidelines:*

- **Discuss and motivate clearly the various decisions taken in all stages of the analysis**.
- If you wish, you can use only a subset of the features of the data in the model building stage. However, for full marks you **must clearly motivate your choice** and explain why some features are discarded.
- You will not be evaluated on the basis the predictive performance of your classifiers, but you would need to show that attempts have been considered to build a classifier with reasonable performance.

## Submission rules and instructions

- Write a short report and submit it as a single pdf file (approximately max 10-12 pages, code excluded).
- Include the R code used for the data analysis in the report. The report can be produced using R Markdown, with the code included in the main text or as an appendix. **The code must be working and the analysis must be reproducible in all parts**.
- Multiple submissions before deadline are allowed and only the latest one will be considered for marking.
- Submission after deadline will incur in penalization as UCD rules (see "Module details" document).
- In general, for full marks you **must explain** concisely and clearly **all reasoning**, as well as **show all steps** and **computations** in your answers. Correct answers alone will not achieve full marks.
- For the data analysis task, submitting only code does not provide any marks.
- **Plagiarism is strictly prohibited** and it will incur in severe penalties (see "Module details" document and "Information materials" tab).