Introduction to Machine Learning Course

Exercise 5
The Elections Challenge – Putting It All Together

You have made it to the finish line and the moment of truth finally arrived … The elections are about to start and you are asked to predict the outcome. To this end, you are given a new sample of voters, represented by their features, but this time without their Vote. Your goals are as follows

1. Predict which party would win the majority of votes
2. Predict the division of votes between the various parties
3. On the Election Day, each party would like to suggest transportation services for its voters. Provide each party with a list of its most probable voters, meaning – Predict the vote of each voter in the new sample
4. What will be a steady coalition – show why this coalition will be more stable than other coalitions
    - Terms for a "steady coalition"
        - Over 51% of the votes, relatively homogeneous with respect to the participating parties, and very much different from the opposition

Please be aware that voters may have changed their mind, which means that party sizes (relative number of voters per party) may have changed. Still, you can assume that voting characteristics (as reflected in the features/labels interactions) remains similar to the ones reflected in the (labeled) data set you have, which means that the trained models (discriminative, generative, clustering, ensemble) are still relevant.

Mandatory Assignment

You should submit an end-to-end process that does it all – Starting from loading and preparing the data, and up to the completion of the tasks. It is permitted to reuse everything from the processes you have submitted in previous exercises

Please note that you now have two data sets –

- The original labeled data set, that can be used for training
- A new, unlabeled data, for which you will need to provide predictions
    - I've added a column named "IdentityCard_Num" (the 1st column) which is merely a running number. It should be used for submitting the per-voter prediction (see below)

Your 1st process should train the models and make the necessary predictions (tasks 1,2,3 above). The process should output a two columns CSV file, where

- "IdentityCard_Num" – Holds the Identity Card number of the voter
- "PredictVote" – Holds your prediction to which party this voter will vote
    - The predict label should be in its original form (Yellows, Blues, …) and NOT an integer representation

Your 2nd process should include all the modelling that you are using in order to handle task 4 (identify a steady coalition).

Please submit

1. The Python script files that implement the 1st and 2nd processes, as described above
2. A CSV file that contain the prediction per voter, as described above
3. A documentation that
    a. Explains your process and any significant decision/insight you would like to share
    b. Predict which party will win the majority of votes
    c. Predict the division of voters between the various parties
        i. Provide a percentage breakdown
    d. Suggest a steady coalition (one that its voters are most similar)
        i. Detail the calculations that made you reach this decision
        ii. Explain why this coalition will be more stable than other coalitions

Comments

- This exercise summarizes the Election Challenge. You may use any possible technique in order to come up with your best results!
- This time, an important criterion to assess your work will be accuracy of your predictions, according to the following scale
    - Fair = 75% - 85% ; Good = 85% - 90% ; Very Good = 90% - 95%
    - Excellent – above 95% – will get a bonus

  A warning note:
    - The Adrenalin level may run high – Beware of overfitting!!!
- A good documentation, which lay out your line of thoughts, is very important for all tasks. It may help me not to consider just the bottom line … For task 4, detailing your calculations and considerations, is actually critical.
- It is highly recommended to consider a hybrid modelling approach. It may not provide you the best results, and you can certainly decide to favor another approach, but in particular, for the Election Challenge, it may provide you some additional insights.
- Just before submitting your results is the right time to use the validation set. It should serve you as a last sanity check to make sure that you did not "mentally" over fitted. More specifically, the validation set is used to test the performance of your final model, for example to get a final assessment of your generalization error.

This exercise can be submitted in pairs!

- **You should submit only one copy but remember to document who are the contributors**
- **“No Couples Swapping” during the semester**
    - **At least not without my formal approval**