CPSC436/536 Project1

Exploring Labeled Data Using kNN and Regression Models

Objective: Gain understanding of kNN and Regression Models and enhance proficiency by applying these techniques to a real-world dataset.

Dataset: You'll work on a dataset (attached) extracted from National Health and Nutrition Examination Survey: https://www.cdc.gov/nchs/nhanes/index.htm.

- The dataset contained health records of n NHANES participants.
  - The attribute list includes:
    - Age, Gender, Race, Blood Pressure readings (Systolic and Diastolic), Lab work (levels of total cholesterol (TCHOL), LDL, HDL, triglyceride), and certain medical conditions such as Diabetes. We also know whether he/she is a current smoker (smoker).
    - In addition to the above attributes, medical professionals consider some cross terms are important, such as age* Systolic, age* TCHOL, age*HDL, age* smoker. You might want to consider them.
  - Target variable: MI. (whether the participant had a heart attack (myocardial infarction).

Goal: Predict the probability of a participant suffering a heart attack (MI) in the near future.

Output: Utilize your most optimal model to forecast the likelihood of individuals in the testing dataset experiencing a heart attack (MI) in the near future.

Evaluation: Your project will be graded based on the difference between your predicted probability and the true label. Specifically, I'll be using both the accuracy and Kullback-Leibler (KL) divergence between the predicted probability and the observed target [wiki],

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x)\ \log\left(\frac{P(x)}{Q(x)}\right),$$

where P(x) is the true label, Q(x) is your predicted probability.

Things to consider when you tune your kNN models:

1. How many features/attributes does the dataset have?
2. What is the class distribution? How many instances are in class1 and how many in class2? If it's unbalanced, should you consider balancing the data?
3. What's the best k value in kNN?
4. What distance metric is good to use? Do you need to scale your dataset?

5. Do you need to include all attributes? Consider excluding those that do not contribute to the classification, as their inclusion may introduce unnecessary noise.
6. Try different dataset partitions (training, testing) to understand your model.
7. …

Additional things to consider when you tune your regression models:

1. Should you use just ordinary least square, lasso or ridge?
2. How does logistic regression compare with linear regression?
3. …

What to submit: Your Jupyter notebook and your predictions for the participants in the testing dataset.

(**Graduate students**) A report that summaries your investigation, and your understanding why some models perform better than others (2 pages)

Attributes keys:

Age	Continues
BMI	Continues
CurrentSmoker	1 yes; 2 no
Diabetes	1 yes; 2 no
Diastolic	Continues
Edu	1- Less than 9th grade; 2- 9-11th grade (Includes 12th grade with no diploma); 3- High school graduate/GED or equivalent; 4- Some college or AA degree; 5- College graduate or above
HDL	Continues
Income	Ratio of family income to poverty
isActive	1 yes; 2 no
isInsured	1 yes; 2 no
kidneys_eGFR	Continues
LDL	Continues
Pulse	Continues
Race*	1 Mexican American, 2 Other Hispanic, 3 Non-Hispanic White, 4 Non-Hispanic Black, 5. None, 6. Non-Hispanic Asian, 7. Other Race - Including Multi-Racial
Sex	1 male; 2 female
Systolic	Continues
TCHOL	Continues
Trig	Continues

*Sometime people consider only three race groups: white, black, and others