HW1: KNN & NB

The UC Irvine machine learning data repository hosted a famous dataset (the Pima Indians dataset) on whether a patient has diabetes. The dataset is no longer hosted on UCI ML data repository, but it can still be found here:

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Part 1 - Build a KNN classifier to classify the dataset.

- Write standard scaler from scratch - do not scale/z-score features using off-the-shelf scaler from sklearn

$$\text{Standardization: } z = \frac{x - \mu}{\sigma}$$

- Scale data using standard scaler
- Split the dataset into training and testing
- Determine the K value, and create a visualization of the accuracy. Report the best K value
- Run 5 fold cross validations - report mean and standard deviation
- Evaluate using confusion matrix
- Use MARKDOWN cell to explain the accuracy of your model

Part 2 - Build a Naive Bayes classifier to classify the dataset

- Train three classifiers using GaussianNB, MultinomialNB, and BernoulliNB
- Split dataset into training and testing
- Run 5 fold cross validations with training set and validation set - report mean and standard deviation. Use test set (holdout set) for final testing.

- Use MARKDOWN cell to explain the accuracy of each. Determine which NB model fits best with the data we have.

Part 3 - Retrain Using Leave-One-Out

- For both classifiers, retrain using leave-one-out cross validation - report mean and standard deviation
- Do you notice any accuracy improvements on our models during run time and testing time?

Part 4 - KNN or NB?

- Explain whether KNN or Naive Bayes works best with our data
- Select model, and retrain your classifier with all the data available