Admission Predictions Conformal Prediction Analysis

Pooja S, Karla F, Pius B, Soham S

Outline

- Introduction
- Conformal Prediction
- Illustration and Dataset
- Methods
- Testing with Machine Learning Models
- Results

Introduction

- ML tools are used to make a potentially critical decision: self-driving cars, diagnosis, etc.
- While traditional algorithms provide a point estimate for the new data, can we have confidence in these predictions?
- How does the level of uncertainty affect the decision made and can the model be safely deployed?
- The correction of these uncertainties becomes particularly important as the cost of being incorrect becomes high and could lead to devastating losses.

Introduction

- One solution that accounts for such uncertainty is an algorithm that provides not a point estimate, but a range, that contains the true label with given levels of confidence. This algorithm is called **Conformal Prediction**.

- Conformal predictors are based on a Nonconformity Measure, a function that gives some measure of dissimilarity between an example and a set of other examples. The higher the value of this function on an example, the more unlikely it is that this example belongs to the selected group of examples.

Conformal Prediction

- A typical model outputs a label $\hat{y}$ but it does not include any confidence parameter in that result.
- Conformal prediction uses past experience to determine a prediction set with a given measure of accuracy.
- Conformal Prediction can be applied to any Machine Learning model. Exchangeability of the data is the only assumption in the data.
- CPs can operate in a classification as well as in a regression setting.

CP Theory

1. Calculate the nonconformity scores (NC Score) for each point $(x_i, y_i)_{i=1}^{n}$ in the validation set.

   NC Score = $1 - P(y_i|x_i)$

1. Get the proportion of the examples in the collection that have a larger NCM than the example in hand.

$$p = \frac{|\{\alpha_i | \alpha_i \geq \alpha_{n+1}\}|}{n+1}$$

1. Create the prediction set

$$S_\alpha = \{y | p \leq 1 - \alpha\}$$

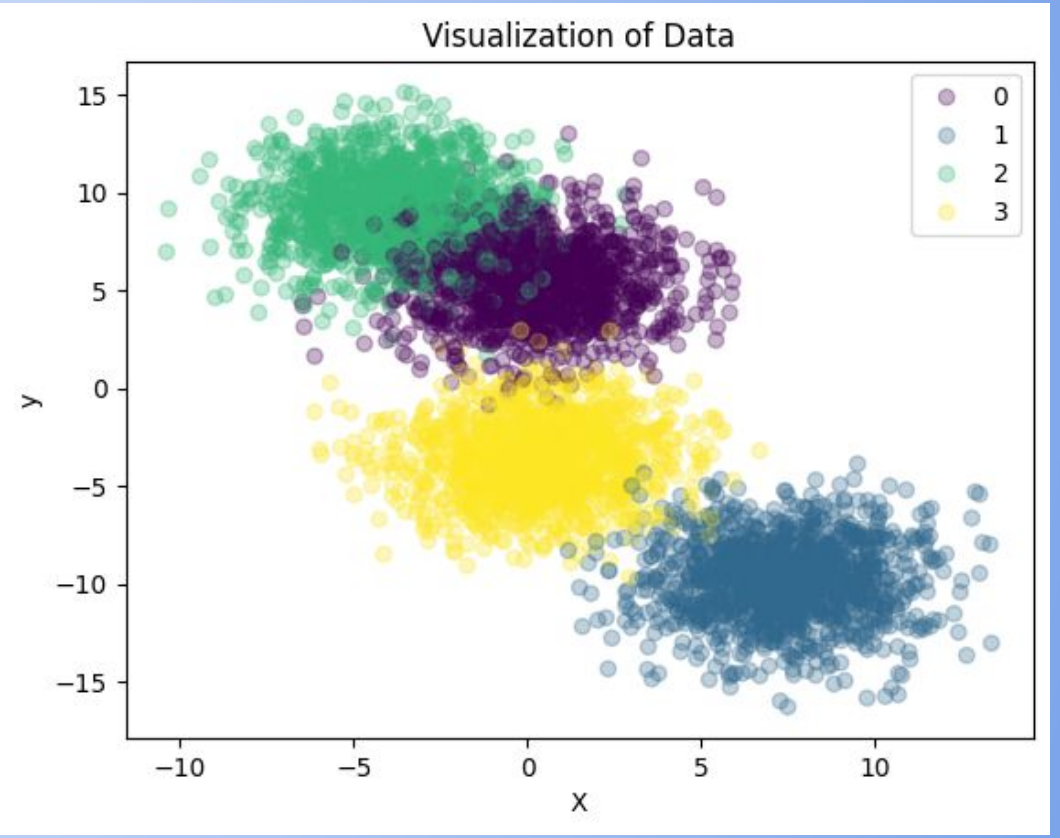
Approach

- Split the data into training, testing and validation set.
- Train the data on one of the supervised algorithms.
- Calculate the non-conformity score for each point in the validation set.
- Create the prediction set for testing data using the NC-Score and p-values.

Blobs

- For an easier understanding approach, we decided to use make blobs from sklearn. The X data is a 2D array showing location of each point on a graph and the Y data is a 1D array consisting of labels for each entry in X.

- We ran the conformal prediction on this data and predicted the interval of label for each random point.



Dataset

- For our actual project, we are using the Admissions Data dataset.
  - 500 rows - students
  - 9 attributes - students profile
- We are using this data to apply the concept of conformal prediction, to predict the chance of a student to get admitted with their profile.
- Chance of admit was modified to have values ranging from 0-5.

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
0	1	337	118	4	4.5	4.5	9.65	1	0.92
1	2	324	107	4	4.0	4.5	8.87	1	0.76
2	3	316	104	3	3.0	3.5	8.00	1	0.72
3	4	322	110	3	3.5	2.5	8.67	1	0.80
4	5	314	103	2	2.0	3.0	8.21	0	0.65
...	...	...	...	...	...	...	...	...	...
495	496	332	108	5	4.5	4.0	9.02	1	0.87
496	497	337	117	5	5.0	5.0	9.87	1	0.96
497	498	330	120	5	4.5	5.0	9.56	1	0.93
498	499	312	103	4	4.0	5.0	8.43	0	0.73
499	500	327	113	4	4.5	4.5	9.04	0	0.84

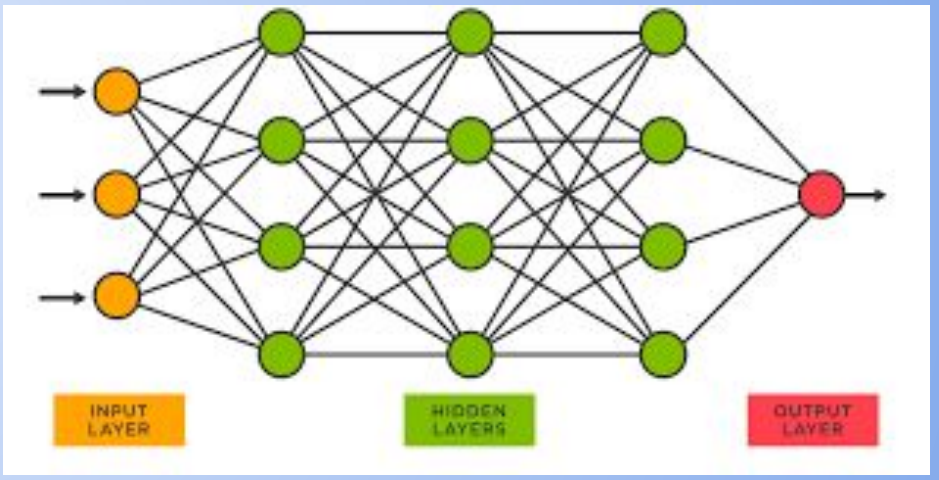
500 rows × 9 columns

Supervised Learning Algorithms

- The KNN algorithm works by finding the k-nearest neighbors to a given input point in the training data and using their labels to make a prediction for the output of the new input.
- Random Forest works by creating individual decision trees by randomly selecting a subset of features as well as a subset of train data for each tree. The final prediction is made by taking the majority vote of output of each decision tree.

Supervised Learning Algorithms

- Neural Networks consist of interconnected nodes, called neurons, that process information and transmit signals to other neurons. Neural Networks can learn to recognize patterns and relationships in data by adjusting the weight and biases of connections between neurons and these connections are adjusted using an optimization algorithm to minimize the difference between the predicted output and the actual output for a set of labeled training data.



Results

Results for blob data:

X_data	Y_pred	Conformal Pred	nc_score
[1.65, 2.862]	0	[0]	0.01
[-3.80 , 7.54]	0	[2]	0.98
[-3.94, 8.07]	0	[3]	0.82

Accuracy:

KNN	Random Forest	Neural Network	Conformal Prediction
0.942	0.94	0.944	0.948

Results

Results for Admissions Data:

X_data	Y_pred	Conformal Pred	nc_score
[9, 8, 4, 5, 8, 8]	1	[2]	0.78
[9, 9, 6, 7, 7, 8]	2	[0, 1, 2]	0.04
[9, 8, 8, 8, 9, 8]	2	[1, 2]	0.01

Accuracy:

KNN	Random Forest	Neural Network	Conformal Prediction
0.62	0.68	0.70	0.96

Conclusions

- Conformal prediction offer a framework for obtaining set-valued predictions exhibiting a chosen error rate.
- In some contexts it may be extremely useful to have a prediction set with a level of confidence (e.g, 95%):
  - If CP exhibits a large prediction set it may be concluded that there may be high uncertainty in the prediction.
  - Otherwise if we have a prediction set with one element, decisions out of it can be taken more safely.

References

1. A Tutorial on Conformal Prediction, Journal of Machine Learning Research 9
   G. Shafer, V. Vovk
   https://jmlr.csail.mit.edu/papers/volume9/shafer08a/shafer08a.pdf

1. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification
   Anastasios N. Angelopoulos, Stephen Bates
   https://arxiv.org/abs/2107.07511

1. Nonconformity Score
   Science Direct
   https://www.sciencedirect.com/topics/computer-science/nonconformity-score