MTH312: Data Science Lab

Curve point identification

1 Problem Statement

A curve is a bend in a road (described in XY dimension) that is used to change direction or align one road segment with another. Horizontal curves are essential in road design to ensure smooth transitions and maintain safety by allowing vehicles to navigate changes in direction at appropriate speeds. Different salient features of a curve definition is illustrated in Figure 1.

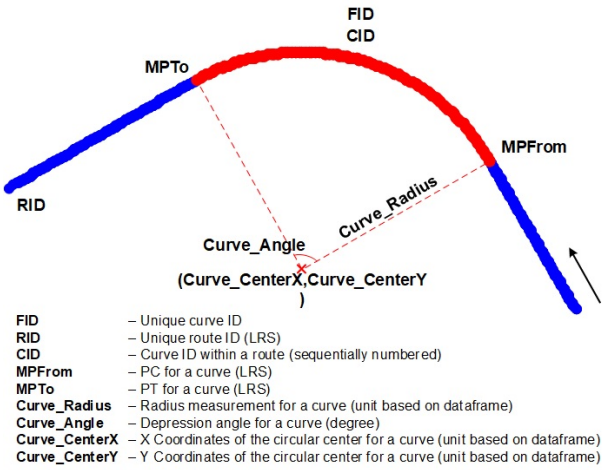

Figure 1: A illustrative example of a curve (image source: https://aichengbo.com/research/national-horizontal-curve-inventory/)

Although there are different types of horizontal curve, the focus of this project is to evaluate a discretized representation of a road stretch as a set of points and classify whether they belong to a curved or a straight section of road. This assessment may be undertaken by iteratively making local approximations of the shape of the road with the help of subset of points (Ai and Tsai, 2015) or by learning a classifier (Bíl et al., 2018). Atif and Sil, 2023 provide a broad summary of different data types and techniques used for horizontal alignment estimation in the literature.

2 Deliverable

Teams are requested to provide a user-defined function that takes as input a data frame containing two columns labeled X and Y. The output should be of the form of a list with two components: (1) data frame containing three columns (X, Y, Curve), with the third column representing a binary (0/1) variable where 1 indicates that the point is a part of a curve and 0 indicates that the point is a part of a tangent/straight section; and (2) other model attributes you'd like to submit (will not be evaluated on).

Although you can use any software to train your model, the requested function must be submitted in `R`. The function must be called `predict_curve` and should look like:

```
predict_curve <- function(dat)
{
.... # possibly load .Rdata objects if needed
....
out <- list(predict, other)
return(out)
}
```

- `predict`: is data frame of three columns (X, Y, Curve).
- `misc`: is itself a list of any other attributes of the model you'd like to share. This can be empty as well if you'd like to not submit anything extra.

The above function should be in a file called `road_prediction.R` in your GitHub repo folder 'Prediction'.

3 Training Data

There are two sets of data sets made available to you as part of the evaluation:

3.1 Indian dataset with a high resolution of points

An Indian highway data set with points available at 10 m intervals is provided which also contain instantaneous radius estimates provided by an instrumented vehicle. The data file provides curve classification based on a threshold radius of 1.5 km (Figure 2).

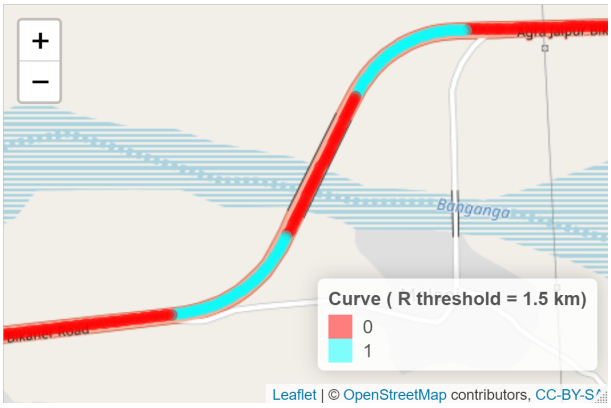

Figure 2: Indian highway stretch

3.2 Czech dataset with variable point resolution

Bíl et al., 2018 provide an annotated data set of points together with their paper (link provided in the paper as well). The data sets contain different sections of roads (indexed by a unique id) with points that belong to both curve and straight sections. However, the underlying threshold/process chosen by the authors to obtain these labels has not been described by the authors (Figure 3).

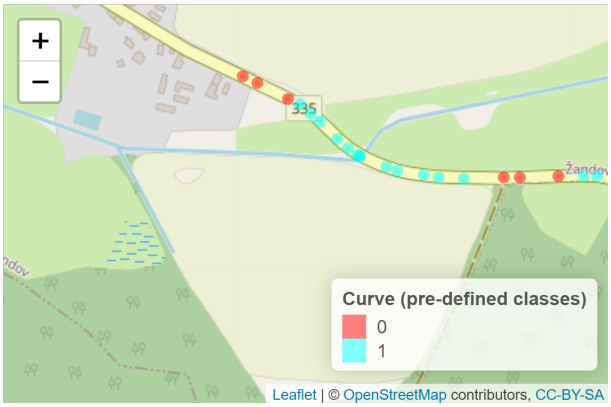

Figure 3: Sample of Czech data

4 Evaluation

There will be two data sets, one containing points at 10 m resolution and another variable point resolution. For each of those datasets, the performance of the submissions will be evaluated using **F1 score** on the test data (which is a harmonic mean of the precision and recall of the data set). To evaluate your score, we will use code in the following form

```
pred_curve <- predict_curve(unlabelled_test_1)
score1 <- calculate_F1(pred_curve$predict, true_labels_1)
```

```
pred_curve <- predict_curve(unlabelled_test_2)
score2 <- calculate_F1(pred_curve$predict, true_labels_2)

score <- score1 + score2
```

As you can see, the performance of the submissions across both datasets will be given equal weightage. You are also required to submit a short (max 3-page) summary of your methodology employed. Your overall marks will be out of 20: 15 for your leaderboard score + 5 for your report. Your leaderboard score marks will be calculated in the following way:

- A lowest possible F1 score will be calculated by us, based on blind guessing of curve or no-curve with equal probability. Any group below this F1 score gets a 0/15.
- If the best F1 score is higher than blind guessing, that group(s) gets 15/15.
- The other groups get marks out of 15, interpolated between F1 score at blind guessing and F1 score by best team.

Your report will be marked on (i) novelty of method and (ii) clarity of presentation.

Your GitHub repositories will be downloaded on: **8 pm Sunday 2nd Feb**.

References

Ai, C., & Tsai, Y. (2015). Automatic horizontal curve identification and measurement method using gps data. *Journal of Transportation Engineering*, *141*(2), 04014078.

Atif, M., & Sil, G. (2023). Investigation of horizontal alignment data extraction methodologies in terms of cost and time. *Proceedings of the Institution of Civil Engineers-Transport*, 1–23.

Bíl, M., Andrášik, R., Sedoník, J., & Cícha, V. (2018). Roca–an arcgis toolbox for road alignment identification and horizontal curve radii computation. *PLoS One*, *13*(12), e0208407.