ECML/PKDD 15: Taxi Trajectory Prediction

July 22, 2015

1 Preprocessing

- Generate a set of balls covering the map (radiuses and centers being chosen to avoid having too many features in the end)
- Remove the trajectories with lightspeed jumps
- For the training, cut the trajectories according to a $\min(U[0,1], U[0,1])$ law.
- It provided a good matching between cross validation and leaderboard score.
- Replace the (truncated) trajectories by the set of balls they cross
- Keep all the other features

2 Learning

- For each feature (boolean: have this trajectory crossed Ball k, is it id_207?) generate a cloud of points that are the final points sharing this feature Actually, the cloud itself is never stored in memory (it would not fit on most computers I guess). Only its barycentre and variance are (they are then updated as mean and variance would be).
- Features and their interactions were considered (without interactions the performance is really low)

3 Predicting

- Given the features, gather all the barycenters and variances.
- Return an average of the barycenters, weighted by the inverse of the standard deviation (raised to a certain power - CV showed that 7 was the best) :

$$\hat{f}(p_1 ... p_n) = \sum_{k, p_k \; is \; true} \frac{(\#C\,(p_k))^{\alpha}\, \mathrm{bar}(C(p_k))}{\mathrm{sd}\,(C(p_k))^{\beta}}$$

Where :

- p_k stands for a boolean feature,
- $C(p_k)$ stands for the cloud associated to the feature,
- $\#C(p_k)$, sd and bar stands for the number of points, the variance and the barycenter of the cloud.