Lab3

Stat665
Bo Song

Model description:
Predictor: Hour, month, hour, arrest, domestic, loc, beat, dist
Loc, beat, dist are encoded as categorical variables.

Work flow
Raw Data -> PCA(129 dimensions) -> Random Forest(50 trees) -> Model

Precise
The misclassification rate on 20% test set is 0.274

• How did you tune the various parameters in your models?

There are 2 variables in my model.

- PCA components
- Number of trees in random forest

For PCA components, I tried to keep 99% variance, which led to 129 components. In my experience tuning this problem, fixed the number of trees to a small value (10) and increase the variance in PCA. Then I discovered 99% is a good number for both efficiency and accuracy.

For Number of trees, I tried 10, 20, 40, and 50. The more trees I used, the better the performance is. However, it takes a long time to train the model as number of trees goes up. So 50 is the result of trade of between efficiency and accuracy. I believe more trees would get a better result.

• How did you use hierarchical modeling (use the prediction in one model as an input to another) or stacking to combine multiple models together?

I use the output of PCA as the input of random forest.

Originally there are more than 500 dimensions in raw data, while PCA shrinks it to 129 dimensions.

• How did you incorporate the categorical 'location' variable? Did it influence the model significantly?

I first code the location variable to numeric index, then I code the numeric index to dummy variables. It influences the model dramatically. The misclassification rate decreases from 0.454 to 0.296 after adding location variable.

• How did you incorporate the categorical variables such as beat, district, ward, and community area (you can use only one if you would like; but don't ignore them entirely)

I choose loc, beat and district as my final choice in this assignment, and encode them as categorical variables. The more location variables I use, the more accuracy I get. However the accuracy is not improved a lot, just increased 5% or so, but it increases the computation time dramatically.

• Describe the confusion matrix. What is the easiest category to differentiate. Why? What categories are hard to tell apart? Does this make sense to you?

y_true\y_predict	1	2	3	4	5	Mis rate
1	4640	629	820	120	636	0.32
2	810	4793	406	423	521	0.31
3	868	371	4224	350	1325	0.41
4	36	185	156	6546	83	0.07
5	330	252	1102	172	5198	0.26

Type 4 NARCOTICS is the easiest category to differentiate. Because this type has a strong pattern that if the arrest of a record is 1, it's highly possible that the type is 4. The correlation is 0.2 or so, which is much higher than that of other predictors. Narcotics is also the most serious crime category among these 5 classes, hence it's also easy to distinguish it.
Type 3 DECEPTIVE PRACTICE is the hardest category to differentiate.

• What you expect your mis-classification rate on the test set will be?
0.274