Seunghyuk Baek

CS5644 Fall2018

sbaek44@vt.edu

HW#3

1. Here are residual sum of squares and variance score for each class (casual, registered and total count of riders)

Daily (732 instances)				
		Casual	Registered	Total Count
Linear Regression	Residual Sum of Squares	157005.08	1615735.75	2084264.28
	Variance Score	0.42	0.4	0.38
Knn Regression	Residual Sum of Squares	98628.35	1990779.94	2441980.81
	Variance Score	0.63	0.26	0.27

Hourly (17380 instances)				
		Casual	Registered	Total Count
Linear Regression	Residual Sum of Squares	1281.33	20585.75	25401.2
	Variance Score	0.36	0.24	0.28
Knn Regression	Residual Sum of Squares	474.9	6318.31	8057.94
	Variance Score	0.76	0.77	0.77

For the regression, it seems that choosing kNN regression is better for smaller data set (daily data). However, with larger instances (hourly data) kNN regression shows much higher in variance score. Thus, in case of large data set, choosing linear regression is correct choice.

2. Here are clusters and ground truth

Clusters (result of k-means):

Counter ({**0**: 77, **1**: 72, **2**: 61})

Ground truth:

Counter ({**1**: 70, **2**: 70, **3**: 70})

The real data was evenly distributed with 70 instances in each class. However, only one cluster was close to this number. There is 13% maximum difference between real class and cluster.