For favour of posting

DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE
THE UNIVERSITY OF HONG KONG

Seminar

Professor HUANG Dawei

Bell Labs Research China

will give a talk

entitled

Critical *k*- Nearest Neighbor Forest for Classification

Abstract

k-Nearest Neighbor (*kNN*) is a widely used machine learning method for classification. However, it may not work for high dimensional data, especially when there are irrelevant features in the data. Also, selecting the neighbor number *k* depends on the data distribution (Hall etc, 2008), we may not know the true distribution in practice. Finally, ensemble learning methods, such as bagging (Bootstrap Aggregation), may not improve the accuracy of *kNN* because *kNN* is relatively stable (Breiman, 1994).

In this talk, we study *kNN* from hypothesis test point of view. We just look at two class case for simple. Assume a feature variable is independent to the class label (H_0 hypothesis), then the probability of the event there are *m* class 1 feature samples in any *k*-sample selection can be worked out easily. The resulting probability density function(pdf) of *m* depends on n_1 and n_2, the sample size for class 1 and 2 respectively, and *k*, the number of neighbors used in *kNN*. Only when *m* falls in the critical region against H_0, the feature might be useful in the classification.

Although this pdf cannot tell us how to obtain the optimal *k*, but it shows us what kinds of *k* might be no good. For example, the current used rule for classification, "majority vote", may lead to errors. Instead, we work out new rule: only when *m* is smaller or larger than corresponding critical values, we classify the sample as class 1 or 2, otherwise we have no conclusion. We call *kNN* with such rule the critical *kNN* classifier.

Using critical *kNN* classifiers, we build up the *kNN* forest. Firstly we select features which are useful for the classification by the hypothesis tests and control the Family Wise Error Rate(FWER, Effron 2010). We use each selected feature as the root for each tree in the forest. Samples, which are not in the critical regions of the root classifiers, will go to next layer of critical *kNN* classifiers. The trees stop if there is no related features or the remaining sample size is too small. Then we have the critical *kNN* forest. When a sample comes, it will go through the trees in the forest, each tree will generate a score for the classification, the final output will be generated by averaging these scores with certain weights.

This system has been tested by real dada.

on

Wednesday, December 2, 2015

(Refreshments will be served from 2:15 p.m. outside Room 301 Run Run Shaw Building)

2:30 p.m. – 3:30 p.m.

at

Room 301, Run Run Shaw Building

Visitors Please Note that the University has limited parking space. If you are driving please call the Department at 3917 2466 for parking arrangement.

All interested are welcome