SML Assignment 1: Directed Edge Prediction in a Twitter Social Network

Laurence Davies, Steven Xu, Herman Johannes

Tuesday 1st September 2015

1 Introduction

Edge prediction in a social network is a well-visited problem in machine learning. This particular example is the prediciton of directed edges that describe subscription to information (”following”) from one party to another. The criteria for evaluation is the successful classfication of a test sample comprised of 1000 edges that has been extracted from a larger training sample without replacement, and augmented with 1000 fabricated edges.

2 Learning Method

Herman, talk about the feature-based method we used. Common neighbours, Adamic-Adar, Katz.

3 Sampling Method

3.1 Naive Sampling Method

3.2 Sample Selection Bias

After several runs on Kaggle where the AUD was significantly lower than that estimated by a test sample extracted from the training data using the above sampling method, it became apparent that there was some sampling selection bias which was skewing the success rate of the predictor.

An analysis of the node degree and common neighbours features extracted on the test-public.txt data and samples from the training data using the naive method described above revealed that although the distributions appeared to be close, two particular differences were noted:

1. The distribution of the to-node in-degree in the naive sample contained far more low-degree nodes than those contained in the test-public set.

2. The test-public set edges generally had more common subscriptions between the from and to nodes than the edges in the naive sample.

4 Results

The final AUC for the team was xxxx