INSTITUTO SUPERIOR TÉCNICO

Traffic Engineering

Lab Project #6

Automatic classification

Fernando Mira da Silva

NOVEMBER 2017

1 Goal

The goal of this project is to test automatic identification of traffic patterns to identify p2p flows on a given network connection. Traffic identification will be performed using simple machine learning based classifiers. The final goal is implement the learning algorithm and to assess the generalization ability of the Naïve Bayes classifier when applied to this specific problem. Learning and generalization will be assessed in four different network scenarios, each one with different network traffic samples.

2 Data set

The file p2pdata.zip contains 8 data files with (simulated) data extracted from the analysis of network flows. The first 5 columns of all these files have an identic structure:

Column 1: Network address of the source IP.

Column 2: Number of simultaneous connection per IP. 50 means less than 50, 100 more than 50 and less than 100, and so on.

Column 3: Average bandwidth used by each IP in the last 60 seconds, rounded to one of the following nearest bandwidth classes: 1-1Mbit/s, 5-5Mbit/s, 10-10Mbits, 25-25Mbit/s, 50-50Mbit/s.

Column 4: Average packet size of all traffic with source on a given IP, rounded to one of the following values: 100 bytes, 300 bytes, 500 bytes, 700 bytes, 900 bytes, 1100 bytes, 1300 bytes.

Column 5: Time of the day: 0:00-8:00, 8:00-16:00, 16:00-24:00.

The goal is to develop an automatic classifier, based on the Naïve Bayes classifier, which is able to predict "unfair" usage of bandwidth by a given network IP due to an heavy usage of p2p connections.

In order to train the Naïve Bayes classifier, there are 4 labeled data files ([1234]-labeled.dat), where previously collected data under controlled conditions is classified as p2p traffic (marked as *yes*) or not p2p (classified as *no*) in the 6^{th} column of each row. The files [1234]-unlabeled.dat are supplied for blind evaluation purpose. The goal is to develop a Naïve Bayes predictor of unfair bandwidth usage for each one of the 4 data files, and to perform the classification on the corresponding unlabeled data set.

For each data set, perform the following actions:

1. Develop an automatic classifier using the Naïve Bayes algorithm;
2. Estimate the expected generalization behaviour of each algorithm (percentage of expected errors in each class);
3. The point 1 and 2 above only require the [1234]-labeled files. After 1 and 2 completed using only the supplied training set, use the developed classifier to fill the missing output column in the "unlabeled" data file. The resulting file must be called "out-[1234]-labeled.dat" and must have 6 values per line (the 5 input attributes plus the estimated output class).

3 Assignment report

The written report of this assignment must include:

1. Estimated error rates in each classification problem;
2. Discussion of the 4 training problems and the obtained results;
3. The "out-x-labeled.dat" for each problem must be supplied along with the written report.

4 Assignment duration and deadline for report delivery

This assignment must be completed in two lab sessions.

The assignment report is due 10 December, by 8PM. Discussions will take place on the week of 11 December.