Artificial Intelligence

Team members:

Naveen Gaddam [111344916]

Shalaka Sidmul [111367731]

### Clickstream Mining with Decision Trees

The project is based on a task posed in KDD Cup 2000. It involves mining click-stream data collected from Gazelle.com, which sells legware products. Your task is to determine: Given a set of page views, will the visitor view another page on the site or will he leave?

Following are our observations:

1. For p\_value = 0.01

No. of internal nodes: 26

No. of leaf nodes: 105

Accuracy reported by autograder:

Tree prediction accuracy: 0.75216

Output file prediction accuracy: 0.75216

Tree prediction matches output file

1. For p\_value = 0.05

No. of internal nodes: 35

No. of leaf nodes: 141

Accuracy reported by autograder:

Tree prediction accuracy: 0.75324

Output file prediction accuracy: 0.75324

Tree prediction matches output file

1. For p\_value = 1.0

No. of internal nodes: 265

No. of leaf nodes: 1061

Accuracy reported by autograder:

Tree prediction accuracy: 0.74832

Output file prediction accuracy: 0.74832

Tree prediction matches output file

Our observations: Number of nodes increases as p\_value increases.

Standard decision trees have no learning bias and in such situation the training error may become zero. Nevertheless, this results into over-fitting. This over-fitting can be prevented if we build simpler tree by performing pruning.

For p\_value = 1.0, the whole tree is expanded and may result in over-fitting.

Whereas, in case of p\_value = 0.01, excessive pruning takes place and it may result in under-fitting.

Thus, we conclude from our observations that for p\_value = 0.05, we get balanced tree and better accuracy.