

CS 374 Assignment #12 Students' Choice

Due the week of May 5, 2008

This semester you have learned about many different aspects of machine learning – fundamental algorithms, theoretical foundations, evaluation methodology, etc. This week you will have the opportunity to explore some aspect of machine learning that is of personal interest to you.

During the week of May 5, we will meet as a class, rather than in tutorial groups. At the class meeting, each of you will give a 10-15 minute project presentation. Ideally, you should aim for 10-12 minutes, leaving a little bit of time for questions. You may use PowerPoint slides, transparencies, a blackboard, or any other audiovisual source. The most important thing, however, is that your presentation be *well-prepared and rehearsed*.

I will suggest two project ideas below, but you are free to propose one of your own. Because our focus thus far has *not* been on applications of machine learning, the projects I suggest below are applications-oriented.

1 Exploring a Fielded Application of Machine Learning

From SPAM filters to voice recognition to video games, machine learning has played an important role in many fielded systems. This week you might choose to explore a real problem to which machine learning has been successfully applied. Your job, then, will be to write a short (3-5 page) paper describing the specifics of machine learning technology's role in the application. As described above, you will also present this to the class.

There are many sources of information on machine learning applications. These include (but are certainly not limited to) the following:

- A website maintained by AAAI (the Association for the Advancement of AI), which gives some general information and starting points:

<http://www.aaai.org/AITopics/pmwiki/pmwiki.php/AITopics/MachineLearning>

- Various issues of IEEE Intelligent Systems. For example, I have one on my desk (Nov/Dec 2007) with an article on mapping Martian land forms.
- Various issues of IEEE Computer. For example, the January 2008 issue has an article on using decision trees to study the social fabric of archaic urban centers.
- Occasional issues of the journal Artificial Intelligence. Vol 172, No 2-3 (February 2008) has an article on using neural nets (and latent semantic analysis) to score essays.
- Special Issue on Applications of Data Mining to Electronic Commerce, Data Mining and Knowledge Discovery 5, (1/2), 2001.
- Handbook of Data Mining and Knowledge Discovery, Oxford University Press.
- Special Issue on Applications and the Knowledge Discovery Process, Machine Learning 30 (2/3), 1998.

I have all of the above in my office, if you can't find them in our library or online.

In addition to considering the sources above, you might find it useful to do a search in Schow. Jodi Psoter is the librarian who serves as our department's library liaison. She will be happy to assist you in doing a search for relevant material.

2 Applied Research: Networks

Tom Murtagh has come to me with a real machine learning problem. As you probably know, his research is in the area of computer networks. One problem he's looking at now involves developing good protocols for deciding which packets a router should drop or keep when it becomes overloaded. Tom has provided me with data in several forms. You can find all of it in the usual place (i.e., my "shared" directory).

2.1 Keep or Drop Based on Packet Size and Between-Packet Time

The first set of data are of the form

```
Keep 1512 86 1512 185 1512 -1 0 0
Keep 64 -1 0 0 0 0 0 0
Keep 64 172444 64 -1 0 0 0 0
Keep 1512 185 1512 -1 0 0 0 0
Keep 64 104512 64 198423 64 72451 64 -1
```

Each of the four lines above describes characteristics of four consecutive packets from a flow (an exchange of packets between a fixed pair of endpoints in the network). The first number on each line is the size in bytes of the data in the first packet. This is followed by the time in microseconds that elapsed between the receipt of the first and second packet. Next comes the size of the second packet and then the time between the second and third packets and so on. Tom tells me that it will be essential to be able to classify the flow with which a packet is associated using only a small number of consecutive packets and also using only packets that arrive within a short time-frame. As currently written, his code considers at most four consecutive packets and only considers packets that arrive within a window of 1/2 second. So, in some cases, there will be information about less than four packets available. The inter-packet elapsed time values of -1 indicate that no additional packet information is available. Thus the 0s that follow in the lines above indicate that there is no information.

The data in the form above are in `KeepDropData.zip`. I've created two "arff" files from these data, which are called `KeepDrop.arff` and `KeepDropMissing.arff`. The former is an exact translation of Tom's data into arff format. That is, missing values are represented with the "-1 scheme" described above. The latter substitutes "?" appropriately where values are missing.

You might be wondering why I'm providing you with raw data. The reason is that you might want to format the data differently, and you might find it easier to work with the raw data than with the arff files, which contain lots of information in addition to the actual instances. For example, you'll notice that about 94% of the data are in one class. You might decide to create a new data set with more balanced data.

2.2 Higher-Level Features and Other Information

Tom has given me two additional files of data that are variations of the data described above. The two files have the same format (details below). They differ in how he determined the correct classification.

In the file whose name ends with “halfWindow” the classification was based on looking at all the packets associated with a given flow that arrived within a half second of the first packet considered in determining the attributes that describe the flow. In the file whose name ends in “fullWindow”, a full second of packets were considered.

Each of the lines in these files is divided into three sections (by | symbols). The first two values on each line are correct classifications. Each line starts with either “Keep” or “Drop” indicating whether the flow’s total data rate for the second or 1/2 second fell below or above the threshold for dropping one of its packets. This binary classification is followed by the actual number of bytes processed for the flow during the observation period.

Next (after a |) come the raw attributes that were included in the file described above (i.e., `KeepDropData.zip`). There are eight such attribute values that should be interpreted as four pairs describing four consecutive packets processed for the flow. The first number in each pair is the size of the packet in bytes. The second is the amount of time (in microseconds) that elapsed between the packet and the next packet processed for the flow.

Next (after a second |) you will find three derived attributes:

1. The rate (in bits per second) at which packets were processed for the flow during the period in which the four (or fewer) sample packets described by the raw attributes were received;
2. the minimum elapsed time between any of the four (or fewer) sample packets; and
3. the maximum elapsed time between any of the sample packets.

In some cases, these derived attribute values had to be computed somewhat artificially. For example, if only one sample packet was available for a flow, the minimum and maximum elapsed times were both set to 1000000.

Although Tom included attributes for four packets, since being able to classify flows quickly is essential to making the algorithm practical (by minimizing the number of flows that have to be tracked at any given time), it would be interesting to see whether (or by how much) the accuracy with which flows can be classified decreases if one only uses the data for three (or even two) of the four packets.

2.3 Talk to Tom

I’ve obviously given you only a very high level description of this application and the data. Tom is the real expert here. If you’re interested in working on this project, talk to him. He welcomes the opportunity to discuss this application and the prospect of getting real results.

2.4 Deliverables

If you choose this project, you will put all of your machine learning expertise and algorithms to use in order to find a good classifier of data into the classes “Keep” and “Drop”. Ideally the classifier will be simple and comprehensible so that Tom can make real use of it in his work. The product of your work will be a paper describing the experiments you conducted, including their results. Of course, you will also present your explorations to the class.

3 “Student Choice”

While you obviously have the option of working on either of the projects above, you might have some other machine learning topic in mind that you would be excited to explore. If so, please drop by to

discuss it with me. You have only one week to complete this assignment. While I'm enthusiastic about having you select a project that is most interesting to you, I want to be sure the scope is reasonable for a one-week project.