THE NEXT DIMENSION OF TOUCH

2483 Old Middlefield Way Suite 150
Mountain View, California 94043
info@qeexo.com
Fax +1 408.228.3787

Qeexo Machine Learning Challenge

The goal of the challenge is to write a program that takes input files describing information about a touch on a mobile device and determines whether the touch is from a finger pad or knuckle. Your program will classify approximately 10,000 unlabeled touches. This task is a simplified version of the actual classification task performed by the FingerSense engine.

Dataset

The *data* directory in QeexoMLChallenge.tar.bz2 file contains training and test data sets. There are approximately 20,000 training instances (10,255 *knuckle* and 10,404 *pad*) and 10,000 test instances (5,263 *knuckle* and 5,265 *pad*). The *train* and *test* directories have the same structure as follows:

- *root* (*train* or *test*)
- *user* directories
  - The name of each directory starts from *hand* or *table* and followed by a timestamp, which is unique to each user who provided a set of the data instances (e.g. *hand-20140213_135947*).
- *instance* directories
  - The name of each directory is a timestamp, which is guaranteed to be unique. For training data, the directory name is followed by a class label (*pad* or *knuckle*, e.g. *20140213_194036984-pad*)

Each instance represents data from a single finger tap, which contains *touch.csv* and *audio.wav*. A *touch.csv* contains information about a touch – the location of the touch represented as *x* and *y* coordinates, the size of the touch represented as the major and minor axis of an ellipse, etc. An *audio.wav* contains a sensor response represented as a one-dimensional signal. Each touch may be from a *pad* or *knuckle*. Each instance folder represents a single instance; its class label (*pad*/*knuckle*) is specified in the directory name.

Each user directory contains a collection of instance subdirectories. A user directory represents data collected from a single user. The *hand* and *table* prefix on the user directory specify whether the data for this user was collected when the user was holding a device in his/her hand, or resting the device on the table. This extra information is provided because it may be useful to you but don't read too deeply into this (there is no guarantee that it will improve accuracy for your classifier). The set of users in train and test are disjoint—no user in *train* is also in *test*.

Output

Your program will generate a file *fingersense-test-labels.csv*, which provides the classification of each instance in the *test* directory of the form:

```
timestamp,label
20140213_134921234,pad
20140213_134922345,pad
20140213_134923456,knuckle
…
```

An example file, *fingersense-test-labels-example.csv*, is included in the zip file.

Language/Libraries

Our recommendation is to use Python and SciPy toolkit (or any other machine learning libraries written in Python); however, using another language will not count against you if you are more comfortable in another environment.

The *example* directory contains example scripts for running an experiment either in Python or MATLAB. The scripts contain some utilities, which might be useful for completing the task. We recommend taking a look if you will be working with one of the languages.

Deliverable

Please email us a zip file containing the following:

- *fingersense-test-labels.csv*
- A document describing your algorithm along with any other notes.
- Your source codes with brief instructions on how to build and run the code.

Time Expectation

We expect this task to take about 10 hours for a developer experienced with machine learning in addition to the time it takes for the documentation. Of course, as with most tasks in machine learning, you can always improve your algorithm. Feel free to spend as much time as you wish on the task, as long as you submit within 7 days of your starting date. We also accept the late submission of the documentation at most several days, so please let us know if it is necessarily for you.

Additional Notes

Please let us know if the requirements are unclear or ambiguous.