EE 559
Jenkins

Course Project Datasets

3/27/2015

You may choose from these suggested datasets, or find your own:

1. Adult

   http://archive.ics.uci.edu/ml/datasets/Adult

   From a set of demographic data on each person, predict whether the person's annual salary is ≥$50K or <$50K. Input features are provided. Most features are categorical; some are numeric. Plenty of data.

2. Bank Marketing Data Set

   http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

   From a set of data on potential clients, predict whether a marketing phone call will result in a new client. Most of the features are categorical. Plenty of data. **Please note:** Feature 11 is highly predictive and unlikely to be known in advance of the phone call, so should be discarded to make the problem realistic.

3. 20 newsgroups.

   http://qwone.com/~jason/20Newsgroups/

   From data on newsgroups text feeds, classify which newsgroup topic the feeds belong to. May be posed as a 20-class problem or a smaller number of classes. Features are number of occurrences of a given word and have already been extracted.

4. Weight lifting exercise

   http://groupware.les.inf.puc-rio.br/har

   Given data from motion sensors on the body, classify whether a given weight-lifting exercise was performed with proper form or with any of 4 types of incorrect motions. 5 class problem. Has missing data; but you'll find plenty of data left after removing incomplete samples and features.

5. Wearable Computing: Classification of Body Postures and Movements (PUC-Rio) Data Set

   https://archive.ics.uci.edu/ml/datasets/Wearable+Computing%3A+Classification+of+Body+Postures+and+Movements+%28PUC-Rio%29

   Mostly numeric features, but some categorical. 5 classes (sitting-down, standing-up, standing, walking, and sitting). There are lots of data; you may find it best to (randomly) draw a subsample of the data to use.

Suggestions for finding your own dataset

1. UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/index.html

2. If you launch out on your own to find a dataset elsewhere, keep in mind the "tips" given in the project assignment. In particular, output label should be a category, and a dataset with features already extracted will typically save you significant time; also check on amount of data and missing data, and keep in mind categorical vs. numeric features.