## ICML-09 dataset

**Dataset files**

* [URL Data Set (Matlab)](http://www.sysnet.ucsd.edu/projects/url/url.mat) (470 MB)
* [URL Data Set (SVM-light)](http://www.sysnet.ucsd.edu/projects/url/url_svmlight.tar.gz) (234 MB)

We checked the SVM one.

**Dataset entry description**

Each entry is a feature set of a URL, description:

Attribute 1 (Label) : +1 corresponds to a malicious URL

-1 corresponds to a benign

Attribute 2 to Attribute n : Attributes derived for the URL

**Example entry**

-1 4:0.0788382 5:0.124138 6:0.117647 11:0.428571 16:0.1 17:0.749633 18:0.843029 19:0.197344 21:0.142856 22:0.142857 23:0.142857 28:1 33:0.0555556 41:0.1 54:1 56:1 64:1 70:1 72:1 74:1 76:1 82:1 84:1 86:1 88:1 90:1 92:1 94:1 96:1 102:1 104:1 106:1 108:1 110:1 112:1 155:1 190:1 204:1 359:1 360:1 361:1 1306:1 1309:1 1310:1 1311:1 2408:1 2921:1 2923:1 7000:1 7001:1 7002:1 7005:1 7006:1 7007:1 7009:1 7010:1 7759:1 7762:1 155153:1 155154:1 155155:1 155156:1 155157:1 155158:1 155159:1 155160:1 155161:1 155163:1 155164:1 155165:1 155166:1 155168:1 155169:1 155170:1 155172:1 155173:1 155174:1 155175:1 155176:1 155177:1 155178:1 155179:1 155180:1 155181:1 155182:1 155183:1 155194:1 155195:1 155196:1 155197:1 155198:1 155199:1 155200:1 155201:1 155202:1 155203:1 155204:1 155205:1 155206:1 155207:1 155208:1 155209:1 155210:1 155211:1 155212:1 155213:1 945789:1 1988571:1 2139257:1 2987739:1 3224681:1

**Labelling field**

Attribute 1 (Label)

Possible values

* +1 corresponds to a malicious URL
* -1 corresponds to a benign

**Questions to answer using ML**

1. URL is benign or malicious?
2. What percentage of URLs were malicious?

**Reference**

* http://www.sysnet.ucsd.edu/projects/url/