“Cyber Analytics” Introduction to Machine Learning Lecture 2

John Cavazos

Dept of Computer & Information Sciences

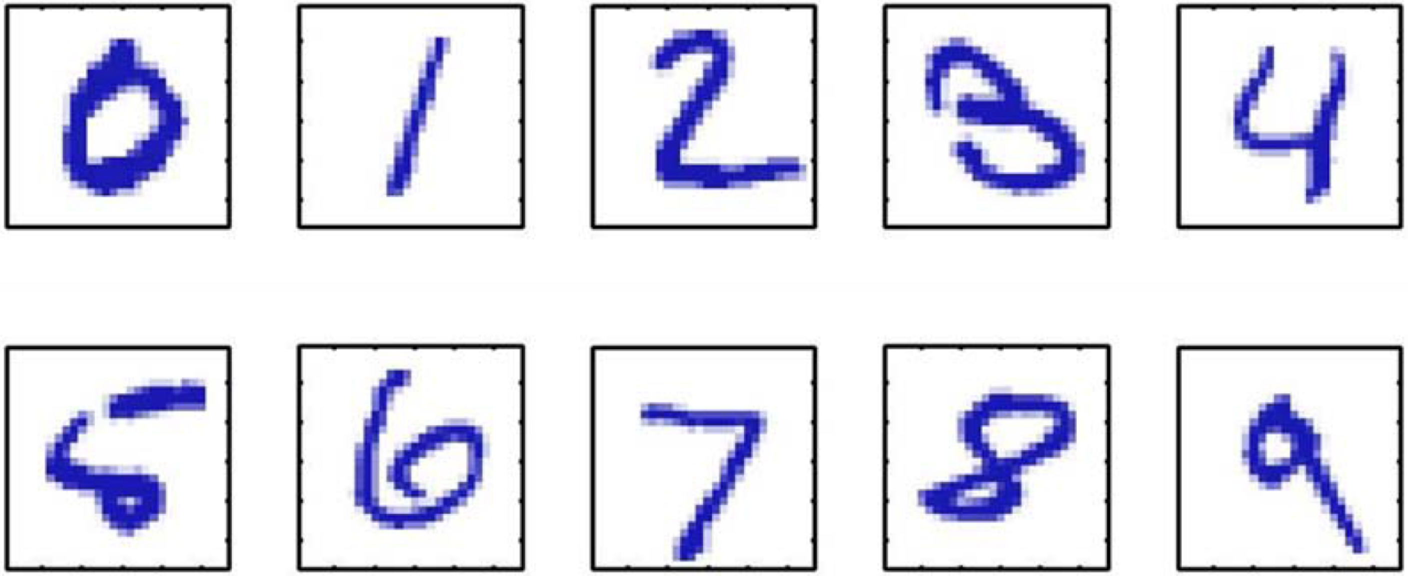
University of Delaware

Derived from: A. Zisserman (www.robots.ox.ac.uk/~az/lectures/ml/lect1.pdf)

Overview

- Supervised classification
  - perceptron, support vector machine, loss functions, kernels, random forests, neural networks, and deep learning
- Supervised regression
  - ridge regression, lasso regression, SVM regression
- Unsupervised learning
  - Nearest Neighbor, PCA

Example 1



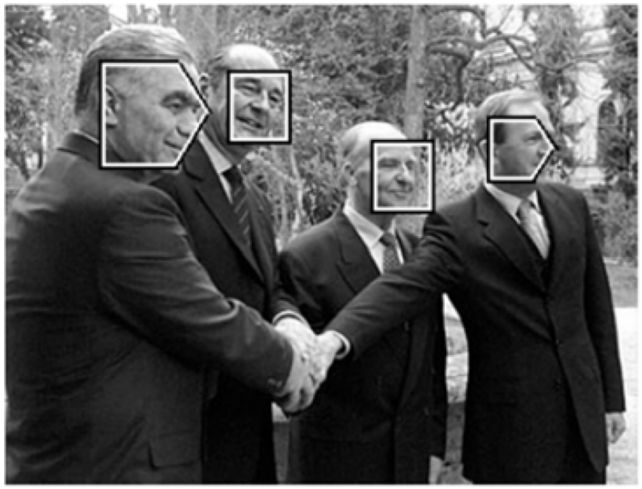
Images are 28 x 28 pixels

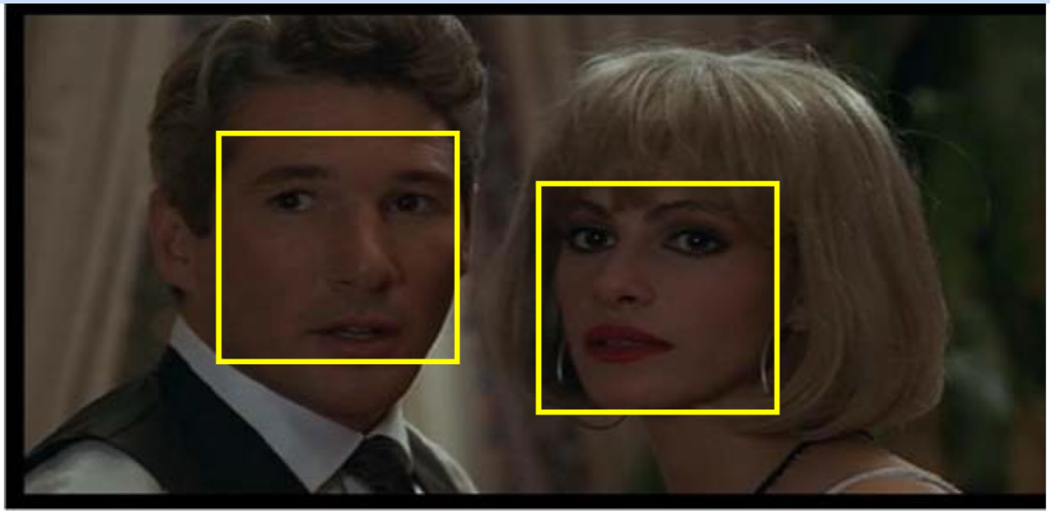
Represent input image as a vector $\mathbf{x} \in \mathbb{R}^{784}$

Learn a classifier $f(\mathbf{x})$ such that,

$$f : \mathbf{x} \to \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

Example 2





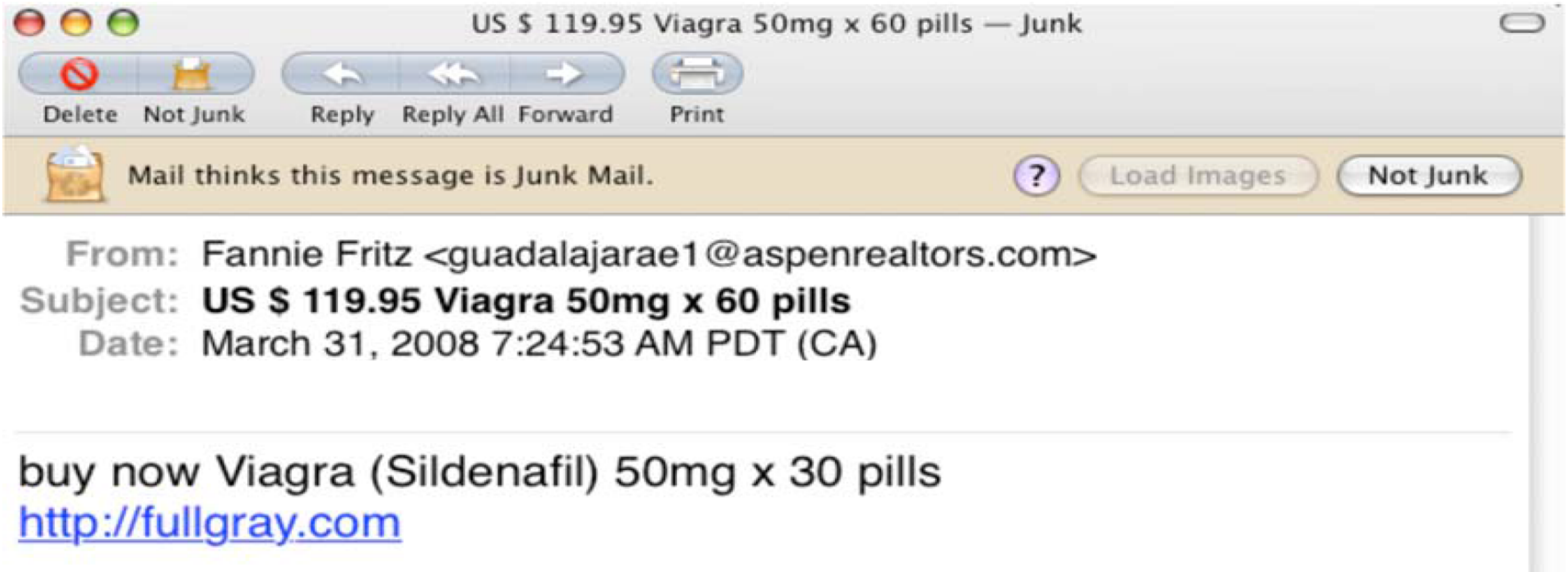
- Again, a supervised classification problem
- Need to classify an image window into three classes:
  - non-face
  - frontal-face
  - profile-face

Classifier Learnt from Data

Training data for frontal faces

- 5000 faces
  - All near frontal
  - Age, race, gender, lighting
- 10^8 non faces
- faces are normalized
  - scale, translation

Example 3

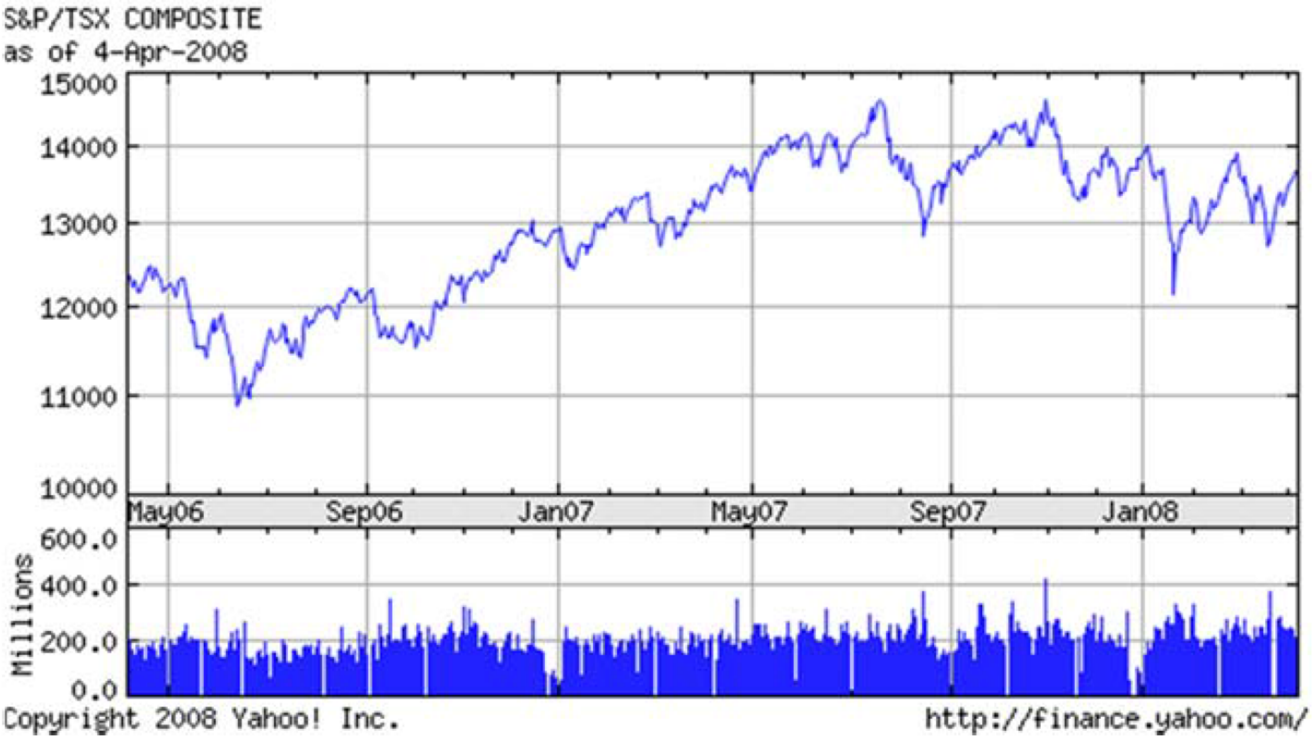

US $ 119.95 Viagra 50mg x 60 pills — Junk
Delete
Not Junk
Reply
Reply All
Forward
Print
Mail thinks this message is Junk Mail.
?
Load Images
Not Junk

From: Fannie Fritz <guadalajarae1@aspenrealtors.com>
Subject: **US $ 119.95 Viagra 50mg x 60 pills**
Date: March 31, 2008 7:24:53 AM PDT (CA)

buy now Viagra (Sildenafil) 50mg x 30 pills
http://fullgray.com

- This is a classification problem
- Task is to classify email into spam/non-spam
- Data x_i is word count, e.g. of viagra, outperform, “you may be surprized to be contacted” …
- Requires a learning system as “enemy” keeps innovating

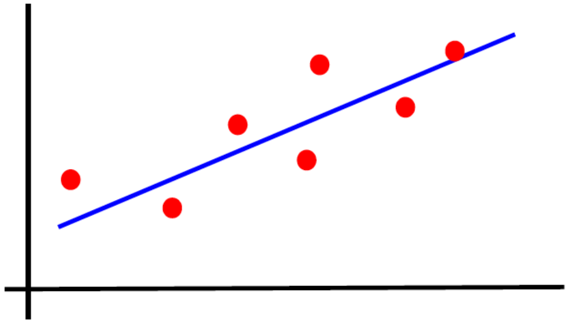
Example 4



- Task is to predict stock price at future date
- This is a regression task, as the output is continuous

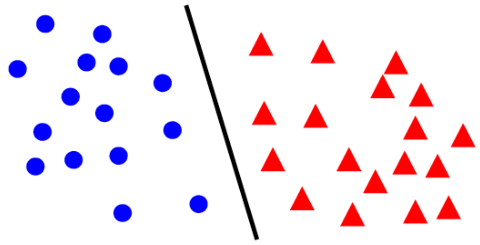
Three Learning Problems

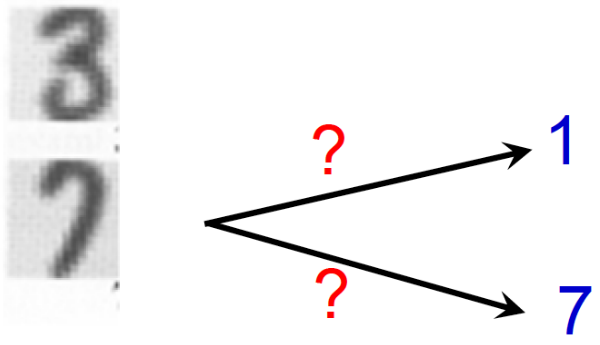
1. Regression - supervised
   - estimate parameters, e.g. of weight vs height



Three Learning Problems

2. Classification - supervised
   - estimate class, e.g. handwritten digit classification

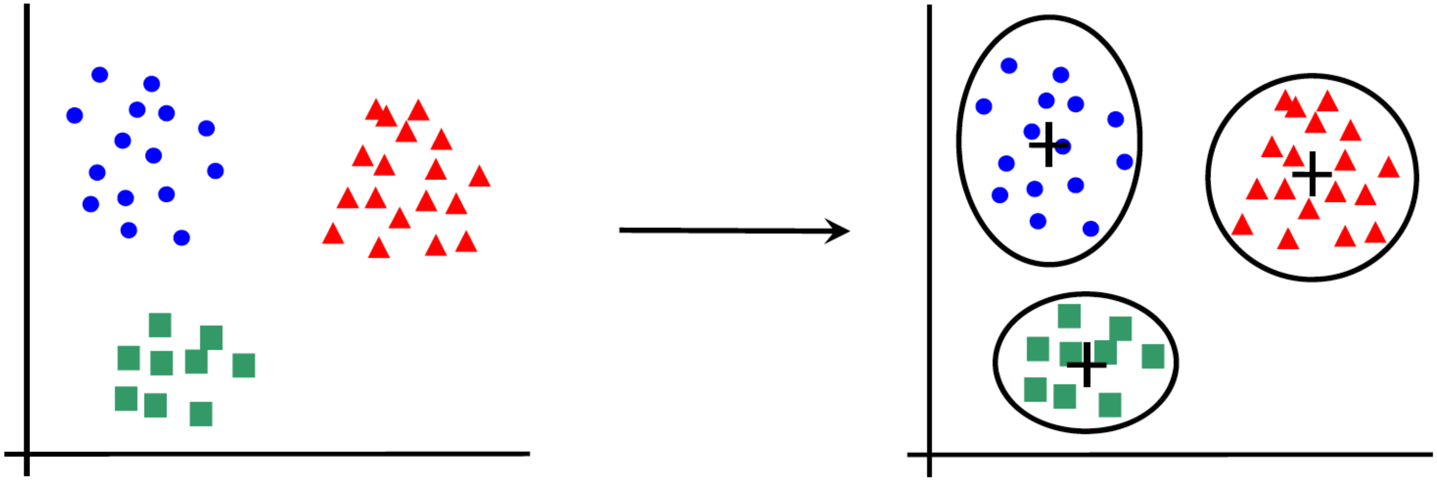




Three Learning Problems

3. Unsupervised learning – model the data

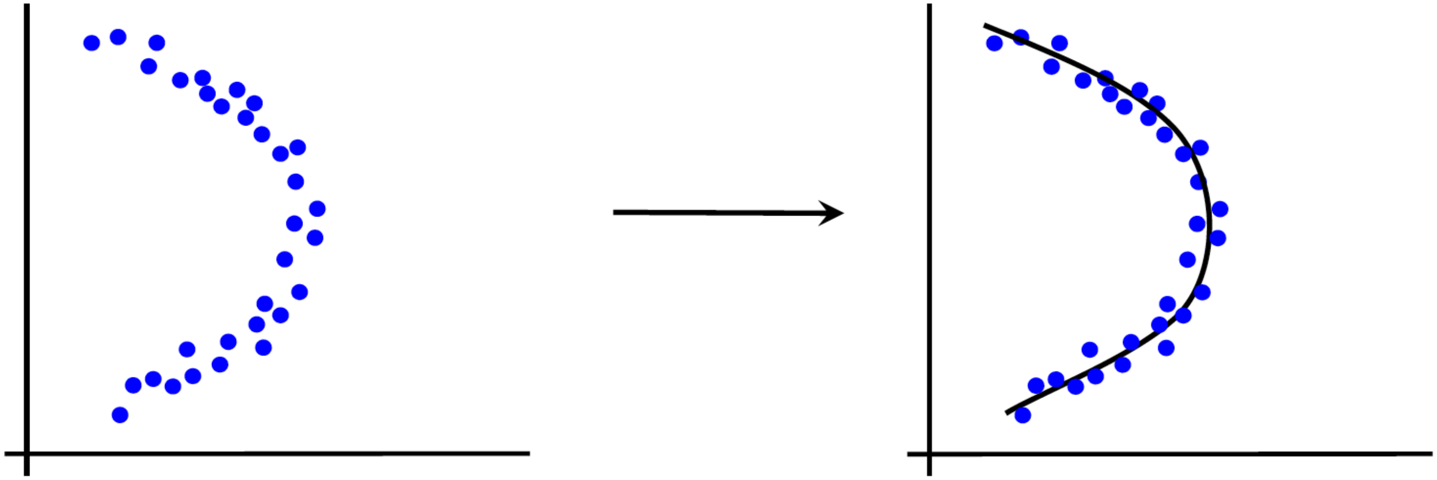
- clustering



Three Learning Problems

3. Unsupervised learning – model the data

- dimensionality reduction



Cybersecurity Examples?

- What are some examples of cybersecurity problems that can be phrased as a machine learning problem?

Cybersecurity Examples?

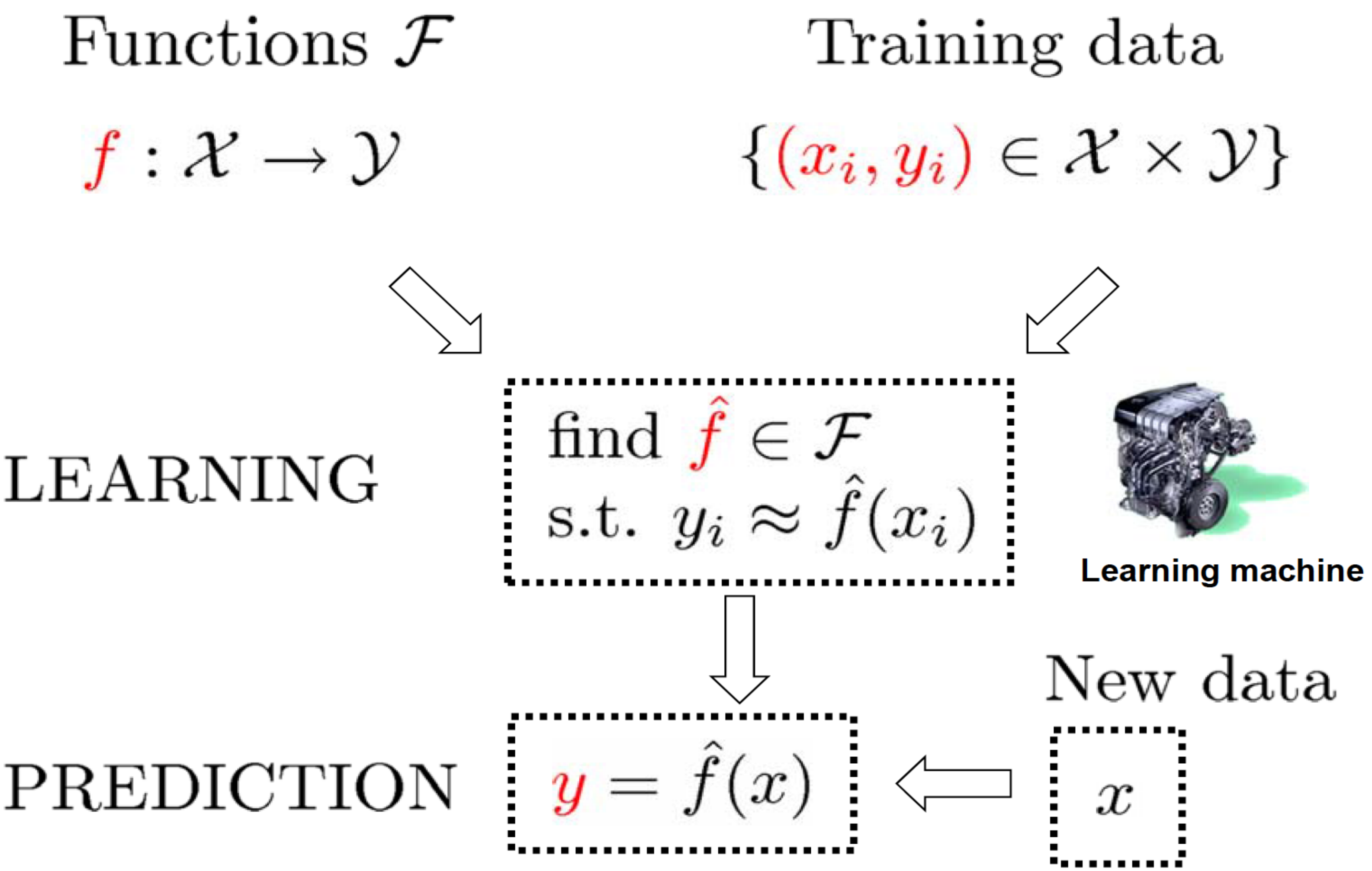
- Is a file being downloaded malware?
- What family does a malware belong to?
- What are the capabilities of a malware?
  - File Encryption? Password stealer?

Cybersecurity Examples?

- What is malware severity?
  - Predict from 1 to 100, where 1 is benign and 100 is really really bad
- What is the risk score of an organization?
  - Depends on vulnerabilities, attacks, protection?

Break

Supervised Learning



Functions $\mathcal{F}$

$f : \mathcal{X} \rightarrow \mathcal{Y}$

Training data

$\{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}$

LEARNING

find $\hat{f} \in \mathcal{F}$
s.t. $y_i \approx \hat{f}(x_i)$

Learning machine

New data

x

PREDICTION

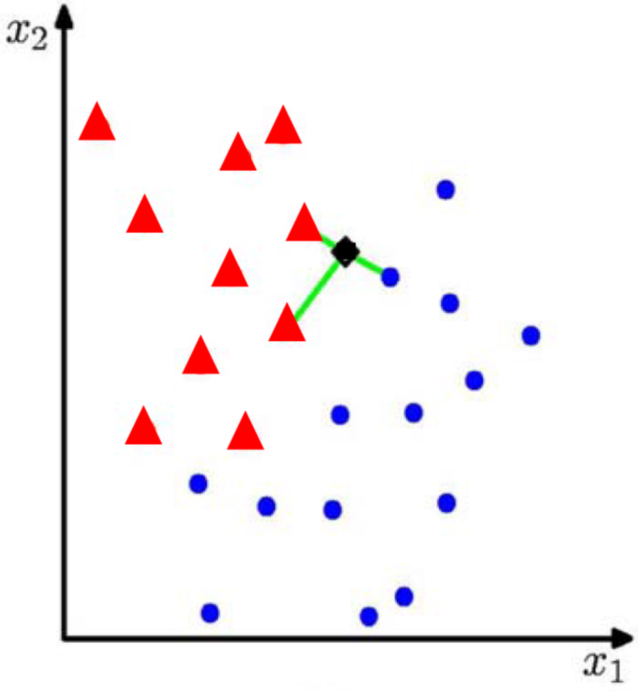
$y = \hat{f}(x)$

K Nearest Neighbor Classifier

Algorithm

- For each test point, x, to be classified, find the K nearest samples in the training data
- Classify the point, x, according to the majority vote of their class labels

e.g. K = 3



- applicable to multi-class case

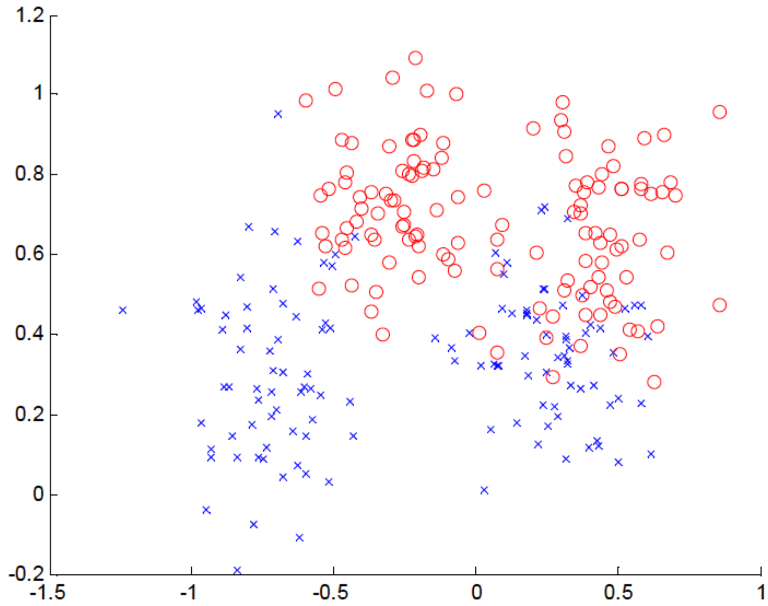
Sampling Assumption

- Assume training examples are drawn independently from set of possible examples

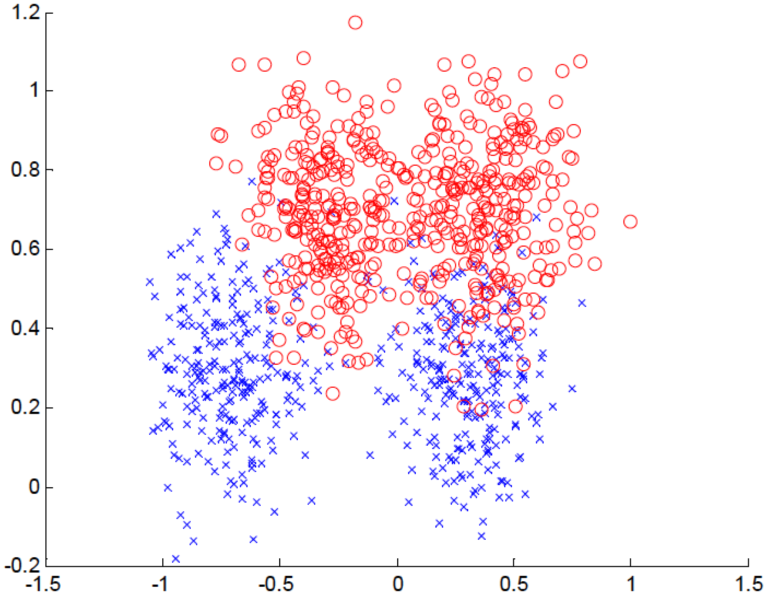
- Makes it unlikely that strong regularity in the training data will be absent in the test data

Measure classification error as $= \frac{1}{N}\sum_{i=1}^{N}\underbrace{[\mathbf{y}_i \neq f(\mathbf{x}_i)]}_{\text{loss function}}$

Sampling Assumption



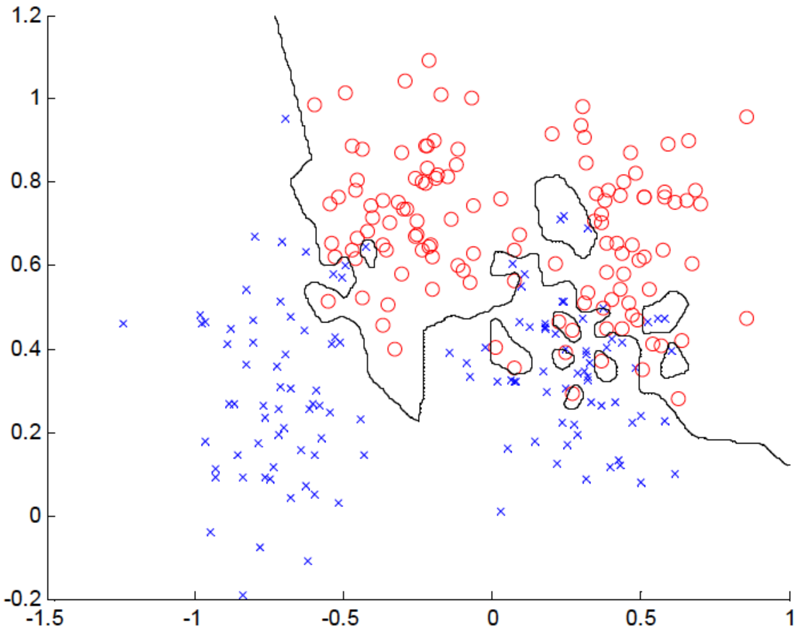
Training data



Testing data

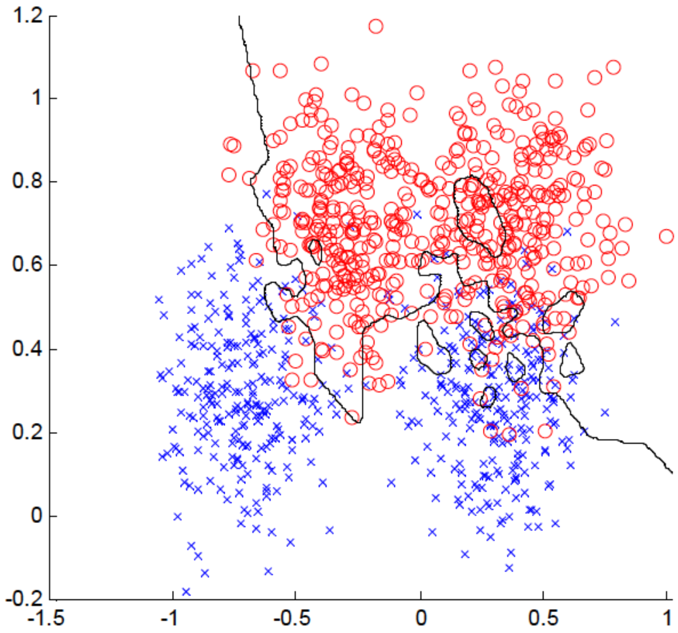
K=1

Training data



error = 0.0

Testing data

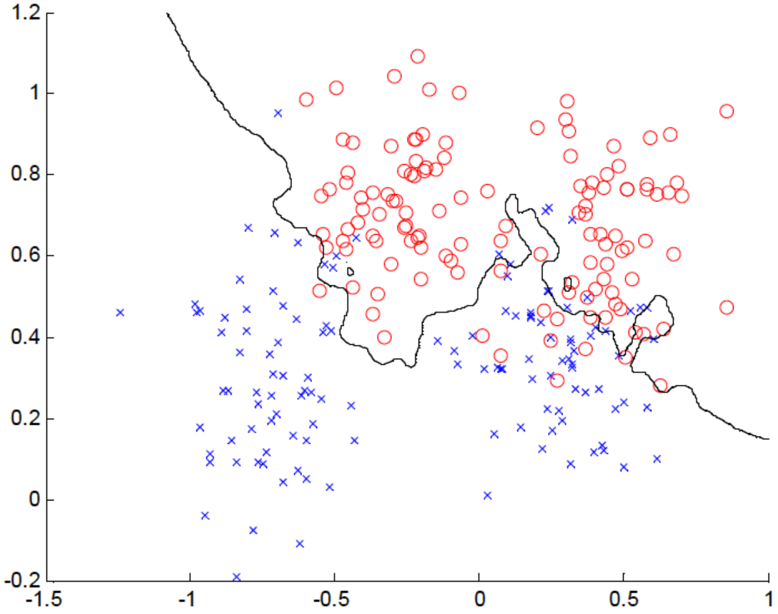


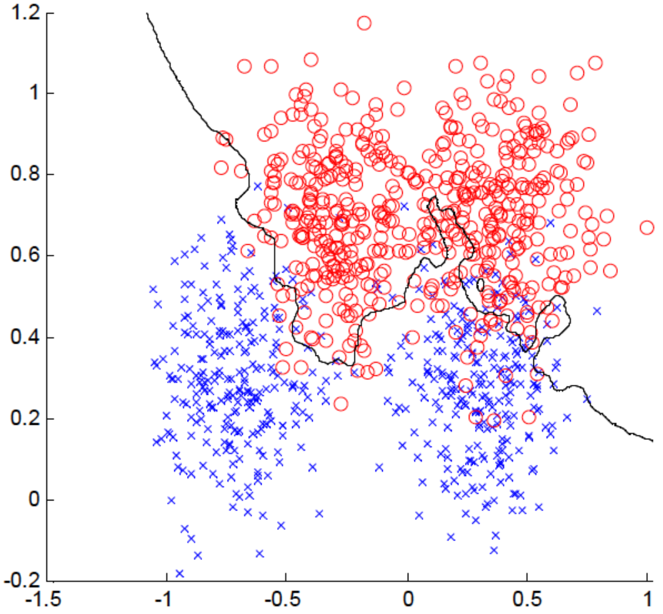
error = 0.15

K=7

Training data

Testing data





error = 0.1320

error = 0.1110

Properties and Training

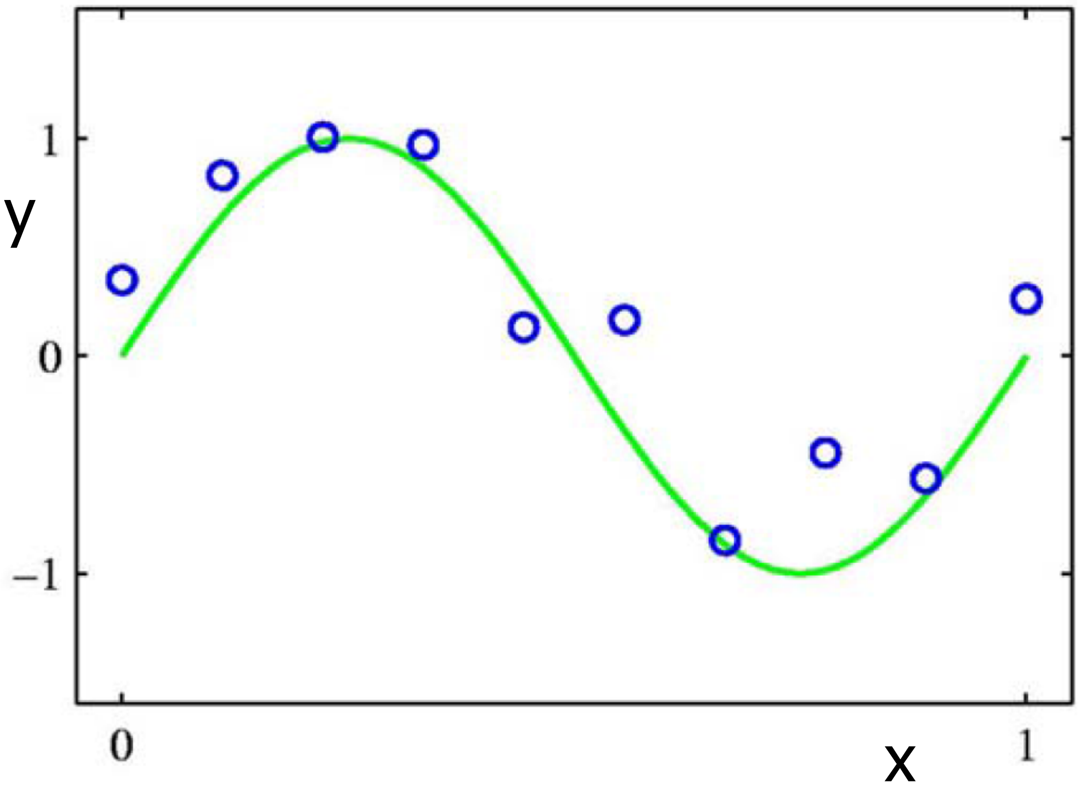
As K increases:

- Classification boundary becomes smoother
- Training error can increase

Choose (learn) K by cross-validation

- Split training data into training and validation
- Hold out validation data and measure error on this

Regression



Regression

- Suppose we are given a training set of N observations

$$(x_1, \ldots, x_N) \text{ and } (y_1, \ldots, y_N), x_i, y_i \in \mathbb{R}$$

- Regression problem is to estimate y(x) from this data

How to set parameters?

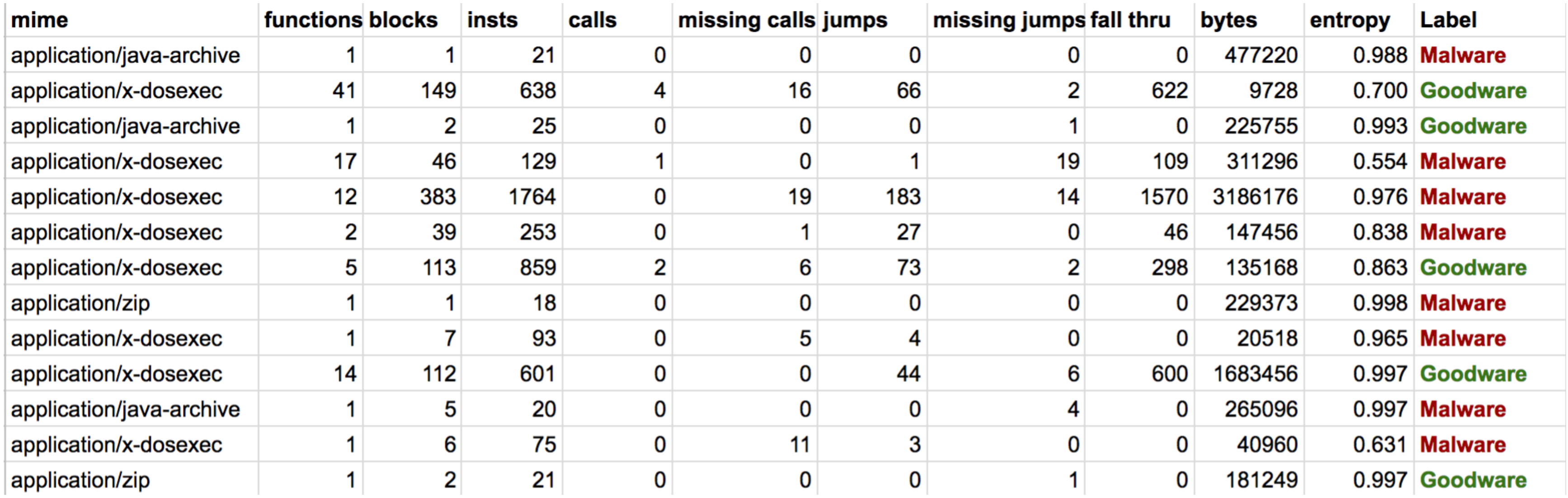
Use a validation set:

Divide the total dataset into three subsets:

- Training data is used for learning the parameters of the model.
- Validation data is not used for learning but is used for deciding what type of model and what amount of regularization works best.
- Test data is used to get a final, unbiased estimate of how well the learning machine works. We expect this estimate to be worse than on the validation data.

We could then re-divide the total dataset to get another unbiased estimate of the true error rate.

Malware Detection ML Data

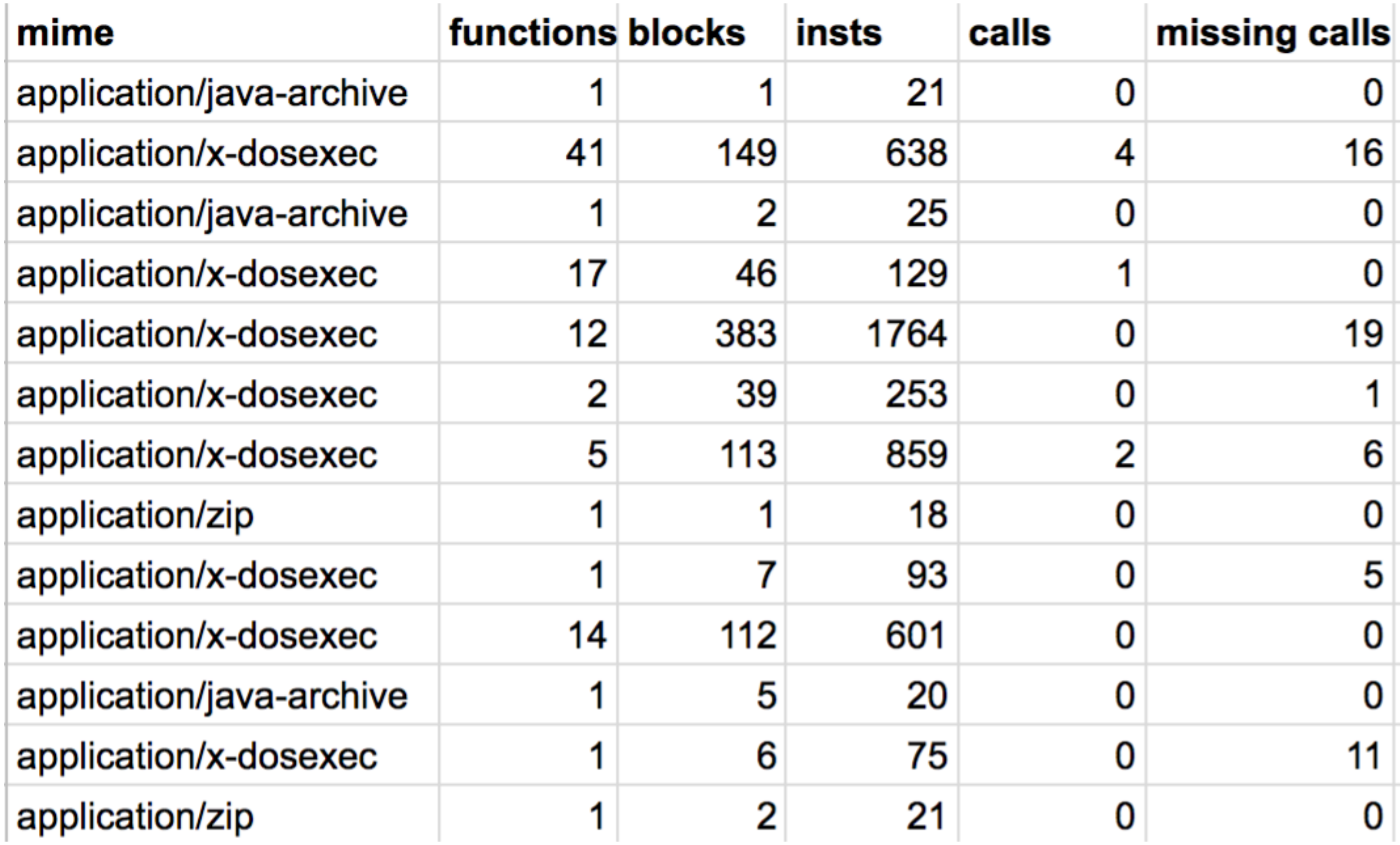


mime	functions	blocks	insts	calls	missing calls	jumps	missing jumps	fall thru	bytes	entropy	Label
application/java-archive	1	1	21	0	0	0	0	0	477220	0.988	Malware
application/x-dosexec	41	149	638	4	16	66	2	622	9728	0.700	Goodware
application/java-archive	1	2	25	0	0	0	1	0	225755	0.993	Goodware
application/x-dosexec	17	46	129	1	0	1	19	109	311296	0.554	Malware
application/x-dosexec	12	383	1764	0	19	183	14	1570	3186176	0.976	Malware
application/x-dosexec	2	39	253	0	1	27	0	46	147456	0.838	Malware
application/x-dosexec	5	113	859	2	6	73	2	298	135168	0.863	Goodware
application/zip	1	1	18	0	0	0	0	0	229373	0.998	Malware
application/x-dosexec	1	7	93	0	5	4	0	0	20518	0.965	Malware
application/x-dosexec	14	112	601	0	0	44	6	600	1683456	0.997	Goodware
application/java-archive	1	5	20	0	0	0	4	0	265096	0.997	Malware
application/x-dosexec	1	6	75	0	11	3	0	0	40960	0.631	Malware
application/zip	1	2	21	0	0	0	1	0	181249	0.997	Goodware

⋮

...

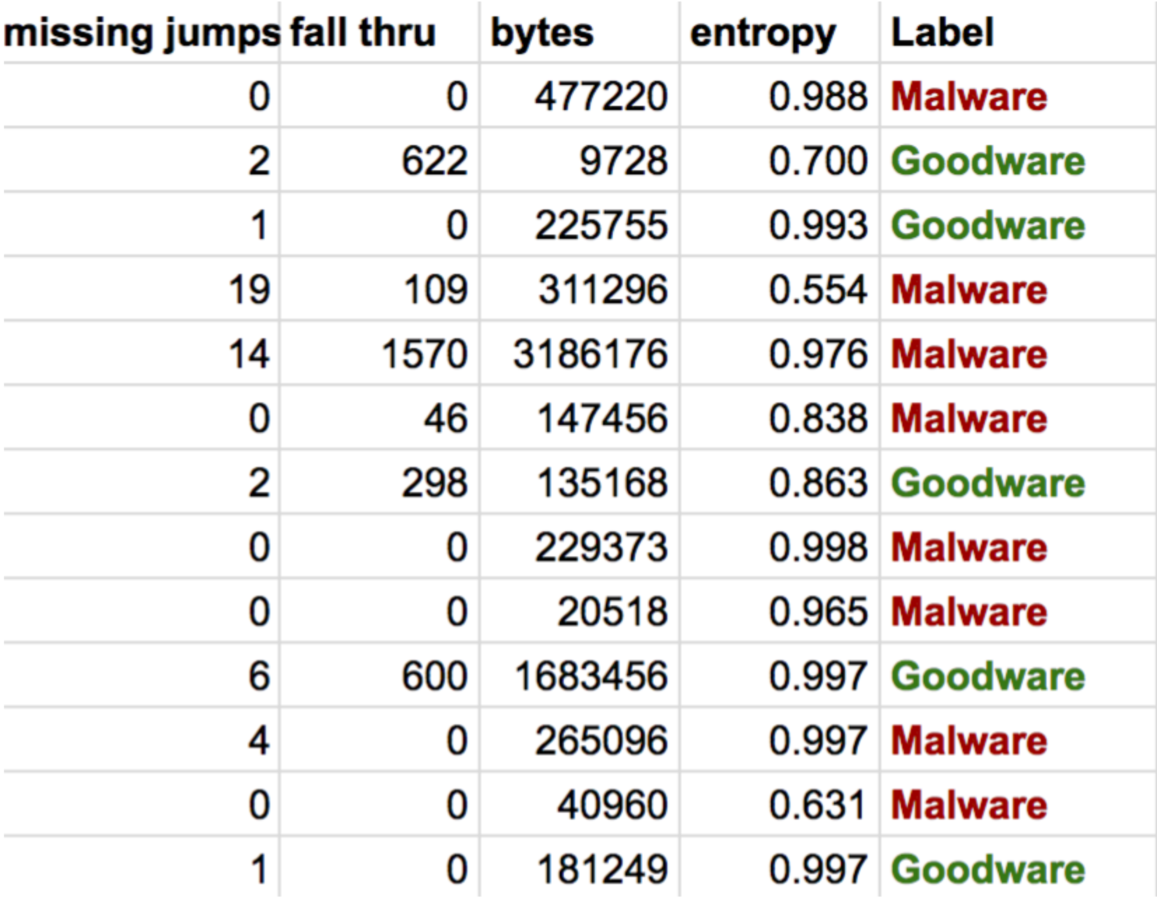
Malware Detection ML Data



mime	functions	blocks	insts	calls	missing calls
application/java-archive	1	1	21	0	0
application/x-dosexec	41	149	638	4	16
application/java-archive	1	2	25	0	0
application/x-dosexec	17	46	129	1	0
application/x-dosexec	12	383	1764	0	19
application/x-dosexec	2	39	253	0	1
application/x-dosexec	5	113	859	2	6
application/zip	1	1	18	0	0
application/x-dosexec	1	7	93	0	5
application/x-dosexec	14	112	601	0	0
application/java-archive	1	5	20	0	0
application/x-dosexec	1	6	75	0	11
application/zip	1	2	21	0	0

...

Malware Detection ML Data



missing jumps	fall thru	bytes	entropy	Label
0	0	477220	0.988	Malware
2	622	9728	0.700	Goodware
1	0	225755	0.993	Goodware
19	109	311296	0.554	Malware
14	1570	3186176	0.976	Malware
0	46	147456	0.838	Malware
2	298	135168	0.863	Goodware
0	0	229373	0.998	Malware
0	0	20518	0.965	Malware
6	600	1683456	0.997	Goodware
4	0	265096	0.997	Malware
0	0	40960	0.631	Malware
1	0	181249	0.997	Goodware

...