XBRL validation logs analysis and classification using supervised learning methods

A Research Project Proposal

Eduardo González
e.gonblan@acm.org

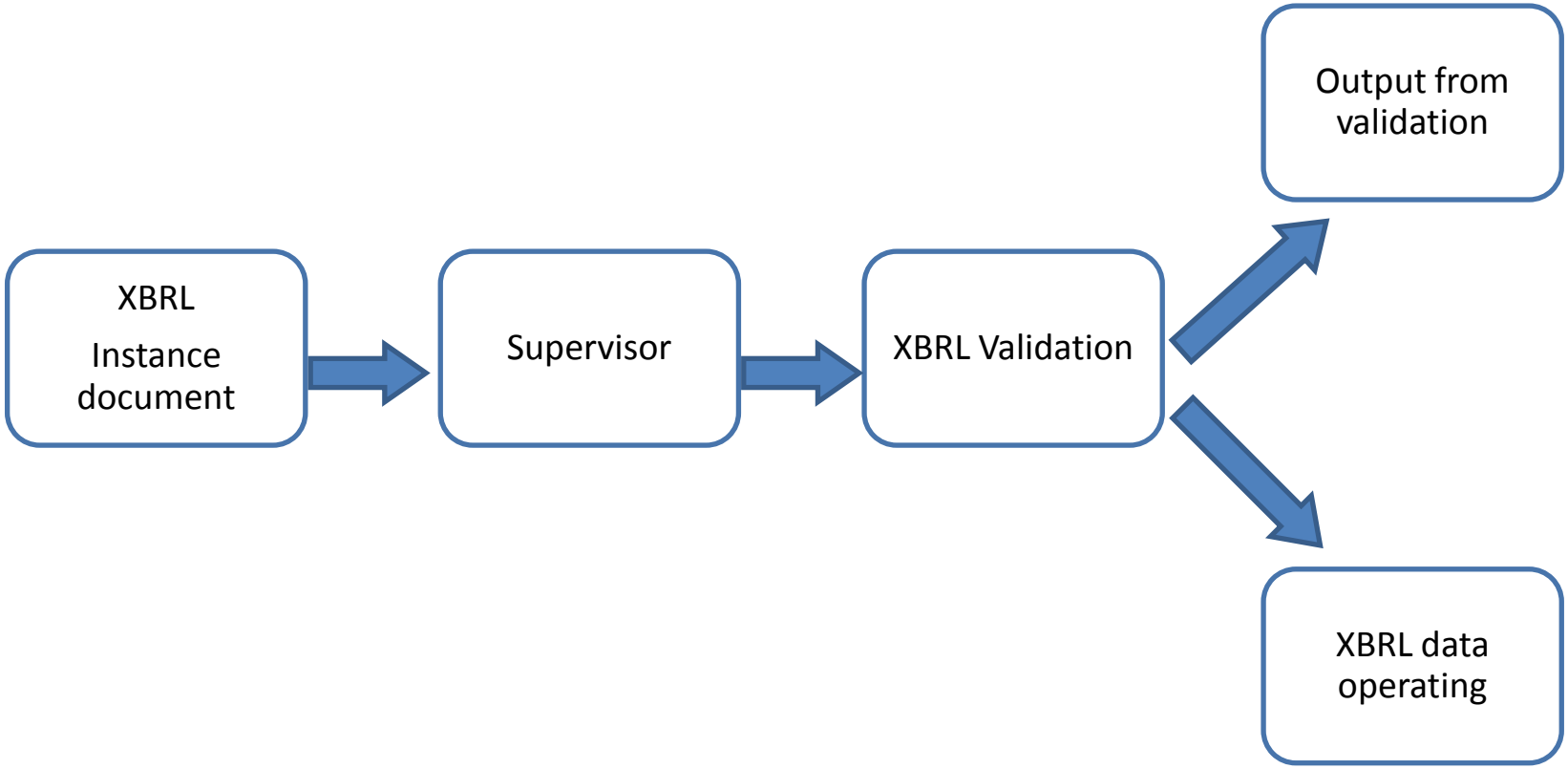
XBRL validation logs analysis and classification using supervised learning methods

- XBRL validation is an essential phase for the acceptation of XBRL Instance Documents. So, execution of this phase should be carefully monitored and controlled.

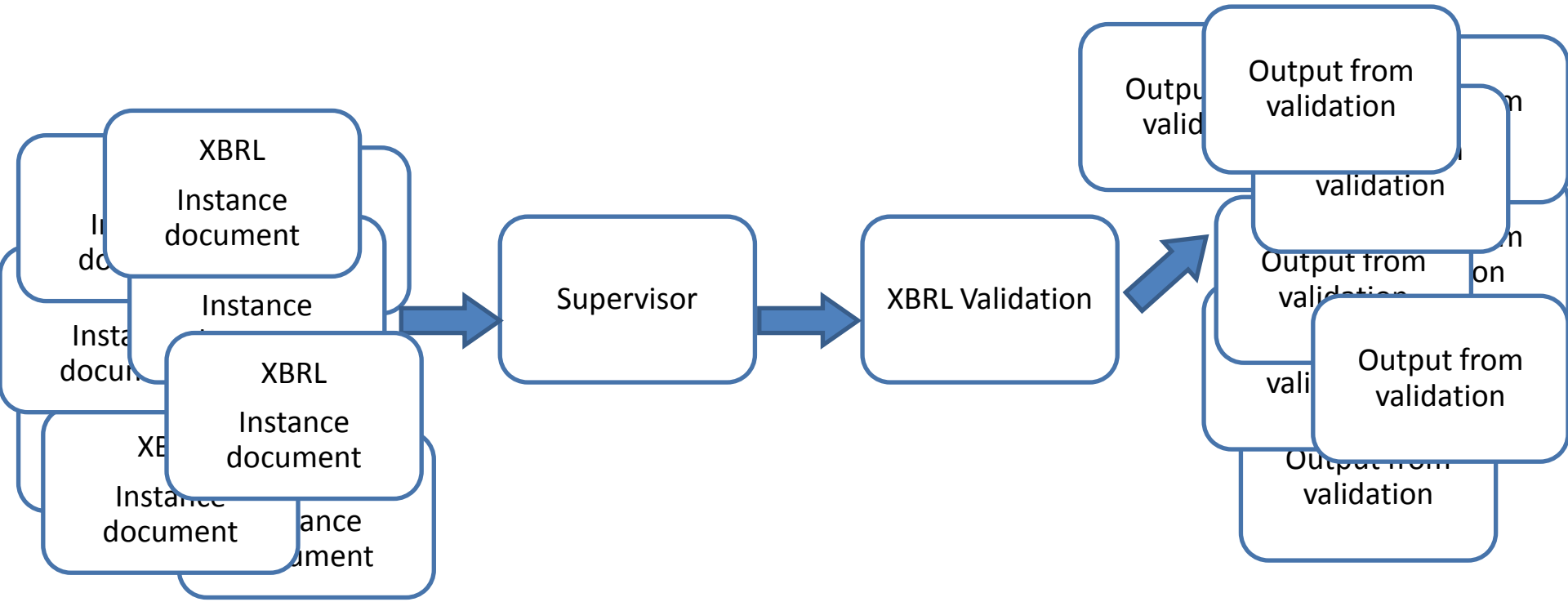
- Logs produced by this validation phase need an analysis.

- This project aims to look into the possibilities of how machine learning analysis such as support vector machines or deep neural nets perform on these results of XBRL instance documents validation, with the aim to improve classification of the logs.

XBRL validation logs analysis and classification using supervised learning methods



XBRL validation logs analysis and classification using supervised learning methods



Validation Output (Fujitsu XWand)

```
...
26/09/2014 07:41:37 589 : VALIDATOR - 2014-09-26 07:32:23.670 - |__EFR Rule Failed: [EFR-MUST: 1.5] Defined period precedes taxonomy publication
26/09/2014 07:41:37 590 : VALIDATOR - 2014-09-26 07:32:23.670 - |__EFR Rule Failed: [EFR-MUST: 2.16] Duplicated facts found
(Context_511:mi119;Context_554:mi116;Context_433:mi119;Context_717:mi119;Context_619:mi116;Context_671:mi119;
    Context_555:mi119;Context_529:mi116;Context_935:mi116;Context_434:mi119;Context_716:mi116;Context_532:mi119;
    Context_432:mi116;Context_508:mi116;Context_1066:mi119;Context_530:mi116;Context_469:mi119;)
26/09/2014 07:41:37 590 : VALIDATOR - 2014-09-26 07:32:40.961 - |__Assertion Failed: eba_v0010_h
26/09/2014 07:41:37 590 : VALIDATOR - 2014-09-26 07:32:40.961 - |__Assertion Failed: eba_v0012_h
26/09/2014 07:41:37 590 : VALIDATOR - 2014-09-26 07:32:40.961 - |__Assertion Failed: eba_v0108_h
26/09/2014 07:41:37 591 : VALIDATOR - 2014-09-26 07:32:40.961 - |__Assertion Failed: eba_v0128_h
26/09/2014 07:41:37 591 : VALIDATOR - 2014-09-26 07:32:40.961 - |__Assertion Failed: eba_v0172_m
...
26/09/2014 07:41:37 591 : VALIDATOR - 2014-09-26 07:32:40.961 - |__Assertion Failed: eba_v0173_m
26/09/2014 07:41:37 592 : VALIDATOR - 2014-09-26 07:32:40.962 - |__Assertion Failed: eba_v0209_m
26/09/2014 07:41:37 592 : VALIDATOR - 2014-09-26 07:32:40.962 - |__Assertion Failed: eba_v0211_m
26/09/2014 07:41:37 592 : VALIDATOR - 2014-09-26 07:32:40.962 - |__Assertion Failed: eba_v0224_m
26/09/2014 07:41:37 592 : VALIDATOR - 2014-09-26 07:32:40.963 - |__Assertion Failed: eba_v0225_m
26/09/2014 07:41:37 592 : VALIDATOR - 2014-09-26 07:32:40.963 - |__Assertion Failed: eba_v0226_m
...
```

Validation output (Arelle)

...
[info] loaded in 138,93 secs at 2014-11-21T13:03:10 - c:\temp\209220\FILENAME.xbrl
[xbrl.3.5.4:hrefIdNotFound] Href http://www.bde.es/es/fr/xbrl/ext/model.xsd#disable not located - http://www.bde.es/es/fr/xbrl/fws/ebacrr_corep/its-2013-02/2013-12-01/val/vr-v4018_a-lab-codes.xml 5
[] Formula xpath2 grammar initialized in 2,85 secs -
[info:profileActivity] ... custom function checks and compilation 7.76 secs -
[info:profileActivity] ... assertion and formula checks and compilation 12.42 secs -
[err:XPST0017] Variable set es_b1005_m
Exception: Function named fext:SolicitarAtributoString does not have a custom or built-in implementation. -http://www.bde.es/es/fr/xbrl/fws/ebacrr_corep/its-2013-02/2013-12-01/val/vr-b1005_m.xml 9
[err:FORG0001] Variable set eba_v1677_m
Exception: invalid cast from str to xs:QName -
http://www.eba.europa.eu/eu/fr/xbrl/crr/fws/corep/its-2013-02/2013-12-01/val/vr-v1677_m.xml 10
...
[info] validated in 28,19 secs - c:\temp\209220\232_solv_ggee.xbrl

Error processing

Error: Assertion Failed: eba_v0187_m

→ Assign an ID: Message_ID

→ Count the errors

Create this vector:

$$\langle Message\_ID | \, count \rangle$$

Error processing

Convert logs messages into numeric numbers

Use n numbers to represent an n-category attribute:

$$\left.\begin{matrix} \boldsymbol{Cat_1} \\ \boldsymbol{Cat_2} \\ \boldsymbol{Cat_3} \end{matrix}\right\} \begin{matrix} \langle \mathbf{0}|\mathbf{0}|\mathbf{1} \rangle \\ \langle \mathbf{0}|\mathbf{1}|\mathbf{0} \rangle \\ \langle \mathbf{1}|\mathbf{0}|\mathbf{0} \rangle \end{matrix}$$

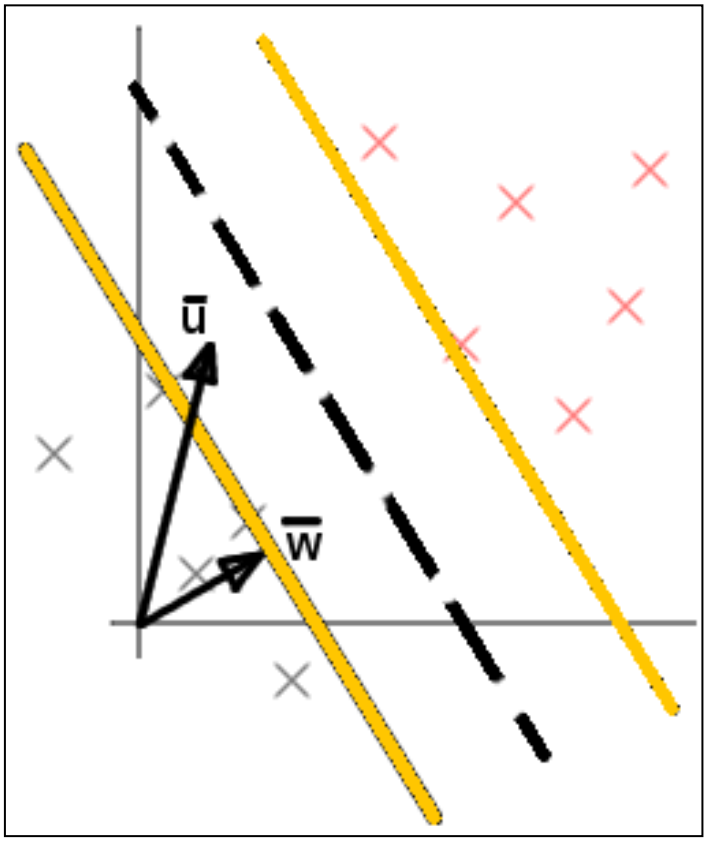
Error processing

Pending data analysis, differentiate between 4 dimensions:

- XML validation
- XBRL validation
- EFR validation
- Formula validation

Support Vector Machine SVM

- Binary classifier
- Supervised
- Find optimal hyperplane that separates training data into two classes.
- After training, classification of unknown pattern is predicted.



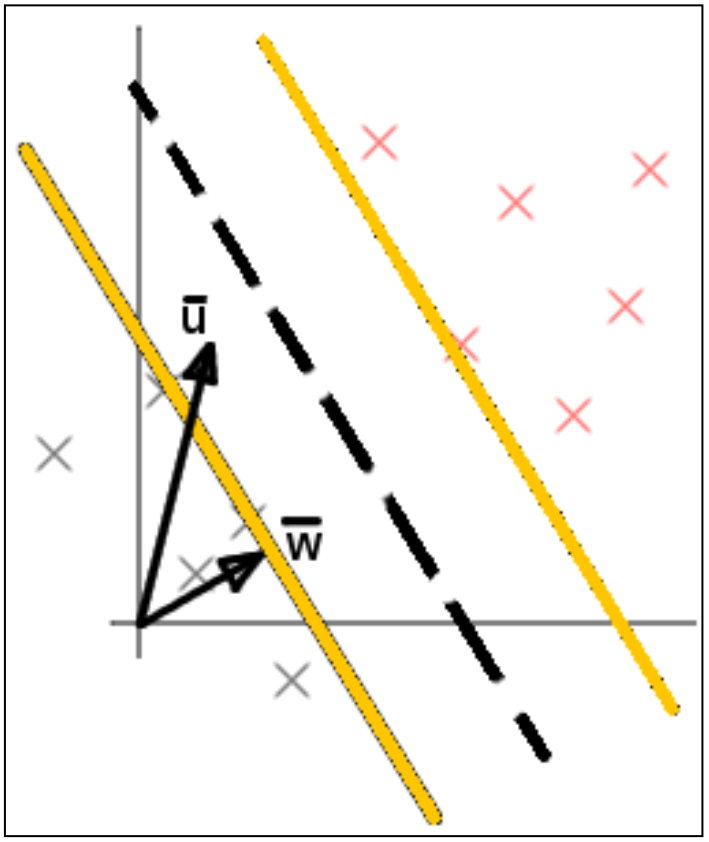
Support Vector Machine SVM

u is in the red part of in the black part?

$$\bar{w} \cdot \bar{u} \geq c$$

Decision rule:

$$\bar{w} \cdot \bar{u} + b \geq 0 \;\; Then\ is\ a\ red\ X$$



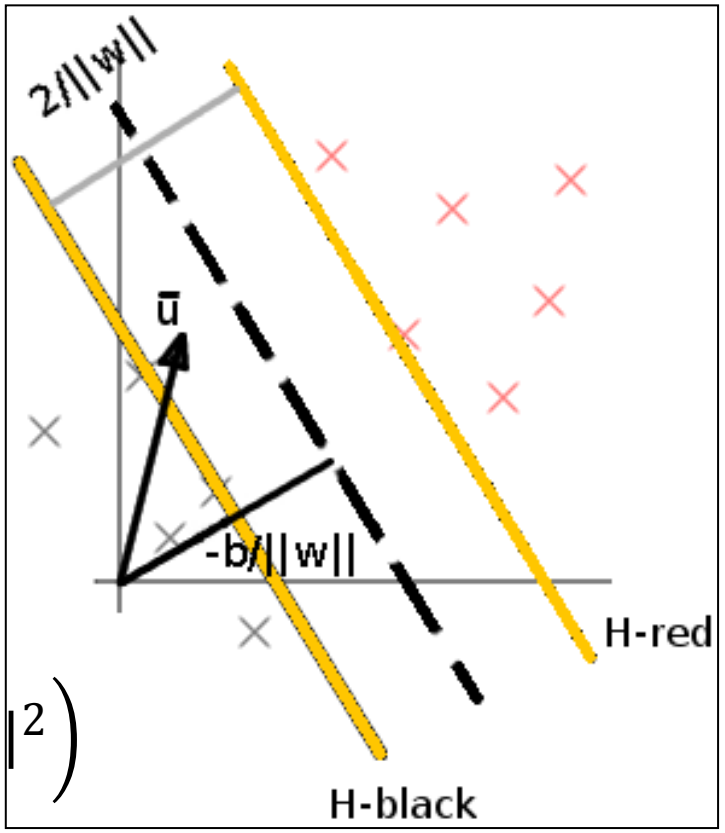
Support Vector Machine SVM

Define

$$\begin{cases} y_i = +1 \; for \; red \; X \\ y_i = -1 \; for \; black \; X \end{cases}$$

$$H_{red} \rightarrow \frac{|1-b|}{\|w\|} \qquad H_{black} \rightarrow \frac{|-1-b|}{\|w\|}$$

$$\max_{f(\cdot)} \left(\frac{2}{\|w\|} \right) \Longrightarrow min(\|w\|) \Longrightarrow min\left(\frac{1}{2} \|w\|^2 \right)$$

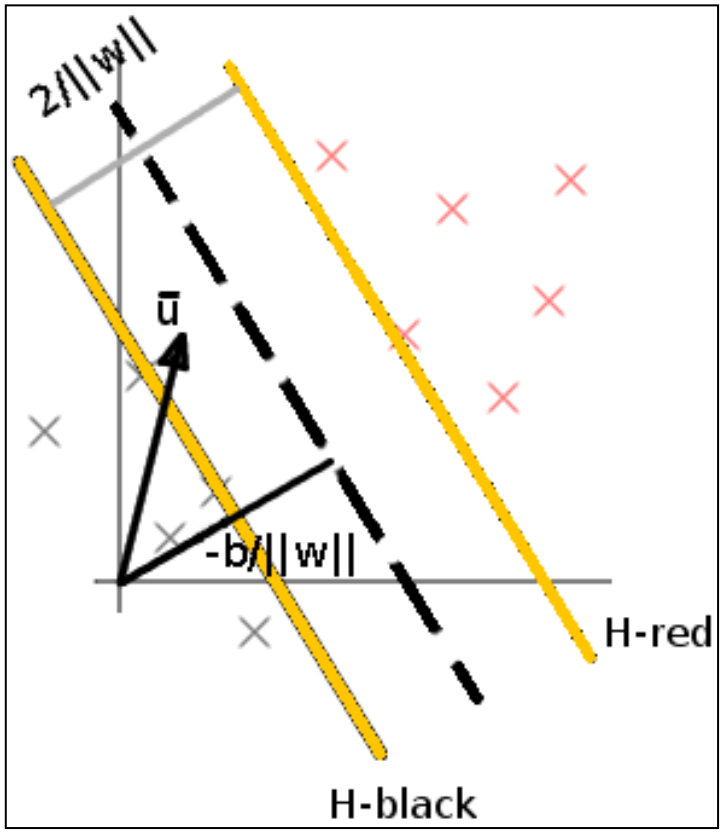


Support Vector Machine SVM

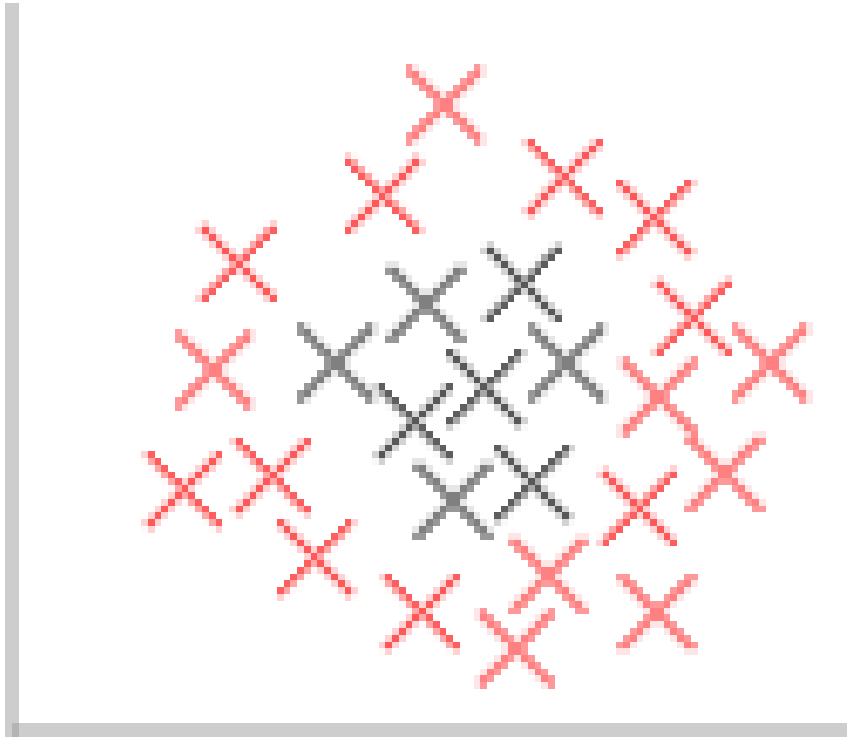
$$L = \sum \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j \, y_i y_j \, \bar{\boldsymbol{x}}_{\boldsymbol{i}} \cdot \bar{\boldsymbol{x}}_{\boldsymbol{j}}$$

Depends on the dot product of samples, so decision rule is:

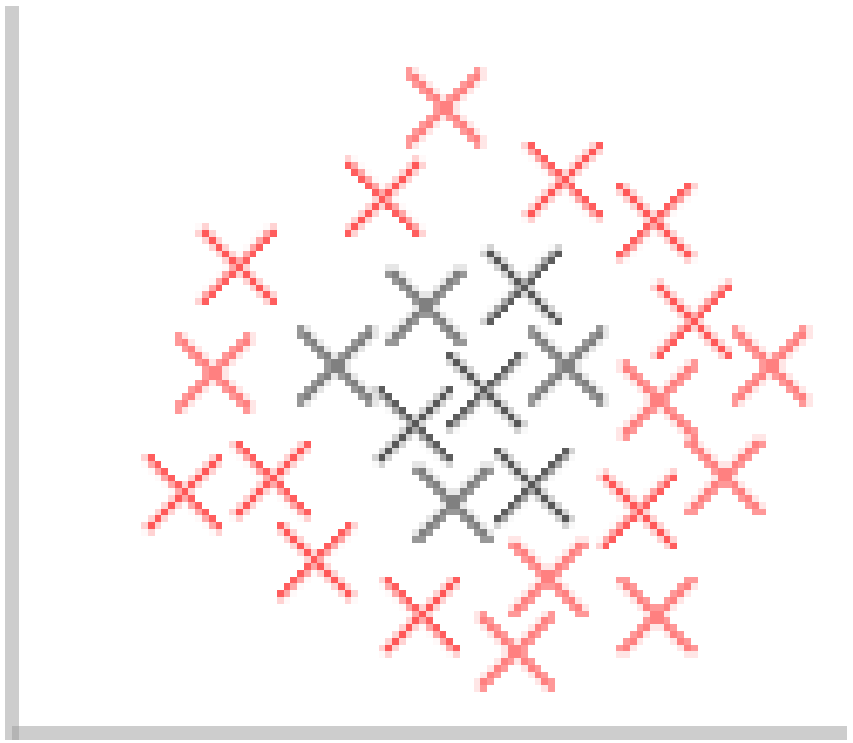
$$\sum \alpha_i y_i \bar{\boldsymbol{x}}_{\boldsymbol{i}} \cdot \bar{\boldsymbol{u}} + b \geq 0 \implies RED\ X$$

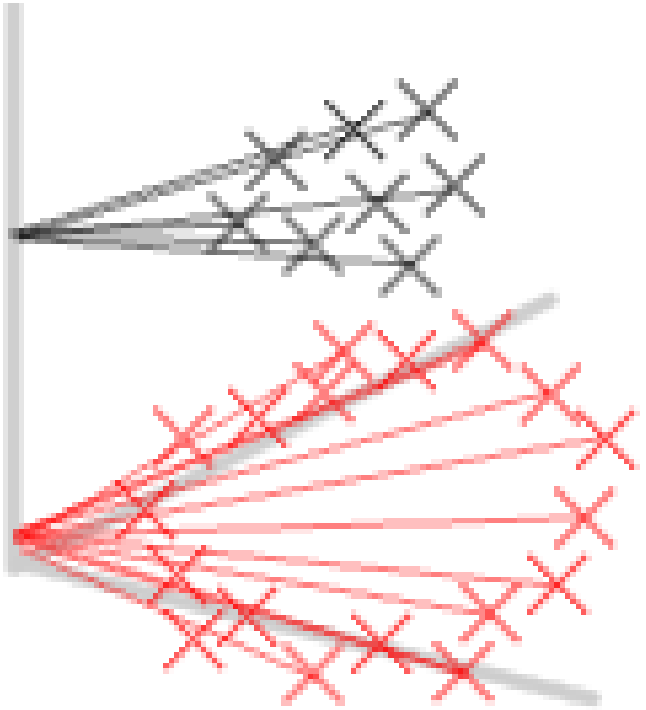


Kernel functions



Kernel functions

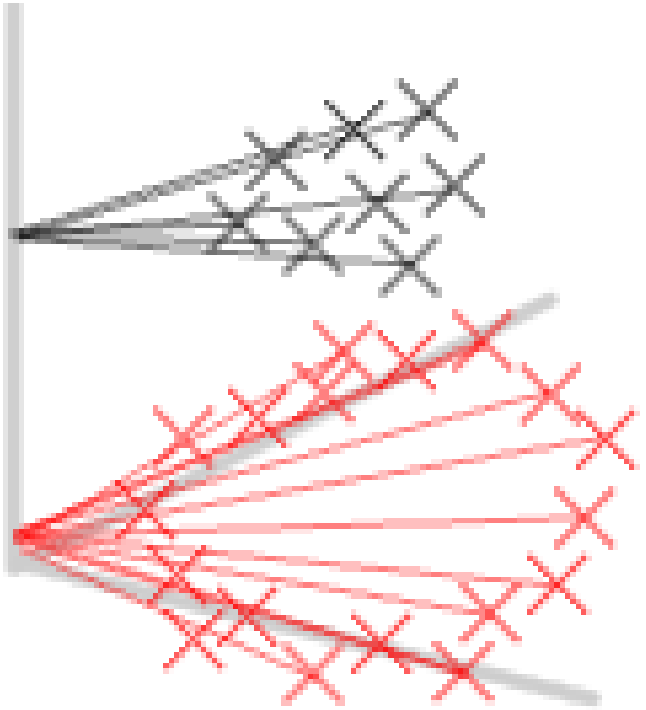




Kernel functions

Function that computes the dot product of x_i and x_j in other space:

$$K(x_i, x_j) = \Phi(\bar{x}_i) \cdot \Phi(\bar{x_j})$$

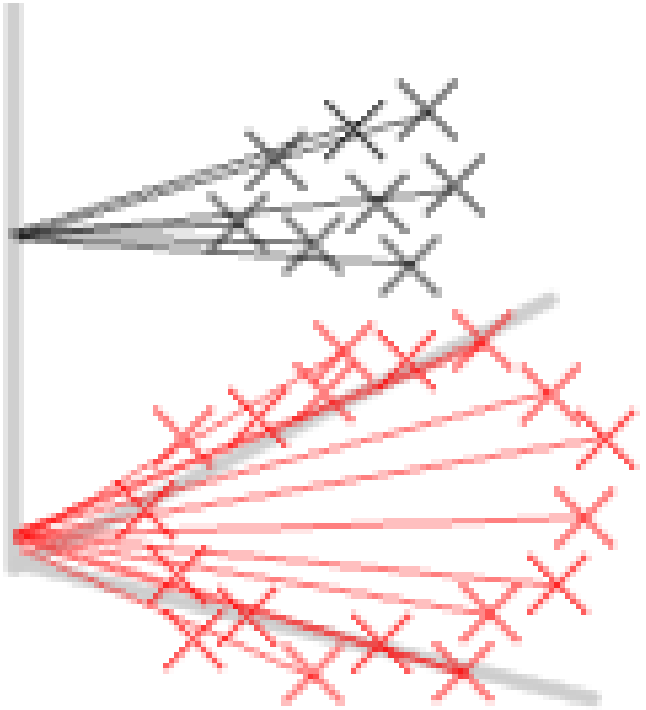


Kernel functions

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^n$$

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$$

$$K(x_i, x_j) = tanh(\kappa\, x_i \cdot x_j - \delta)$$



N-ary classification

one-vs-all

N classifiers

f_i is classifier i,

- Compute the positive count of all the samples in red X,
- Compute negative count of all the samples in black X,

classify by: $f(x) = argmax_i \, f_i(x)$

N-ary classification

all-vs-all

N(N-1) classifiers.

f_{ij} is the classifier where:

- class i are red Xs
- class j are black Xs.

So, we have a matrix and the general classifier is:

$$f(x) = argmax_i \; f_{ij}(x)$$

Next Steps

- Collect a big enough set of real data (log files).
  Evaluate use of UC4 and Automation Cockpit.
- Initial data analysis
- Study possible clustering inside data
- Apply techniques explained in point 3 to decide how to work with the data
- Apply SVM to data collected, including different kernel functions. Keep also an eye on DNN.
- Analysis of results and conclusions.

XBRL validation logs analysis and classification using supervised learning methods

Eduardo A. González Blanco

e.gonblan@acm.org