Machine Learning for Solar Flare Prediction

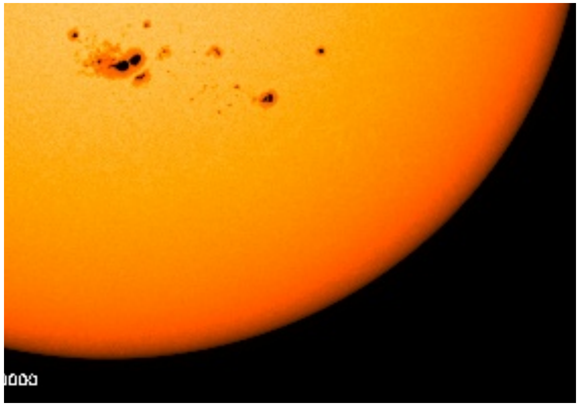
UCL
Université catholique de Louvain

UCL Université catholique de Louvain

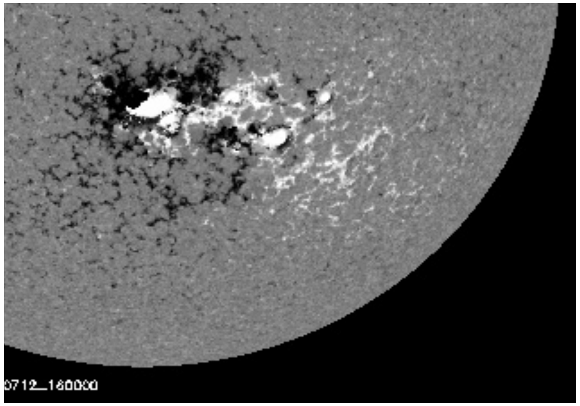
The Task

Solar Flare Prediction

Use Machine Learning to predict
solar flare phenomena



Continuum or white light image



Magnetogram at surface of Sun

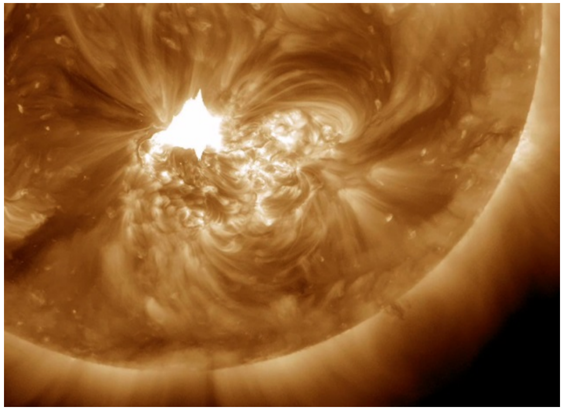
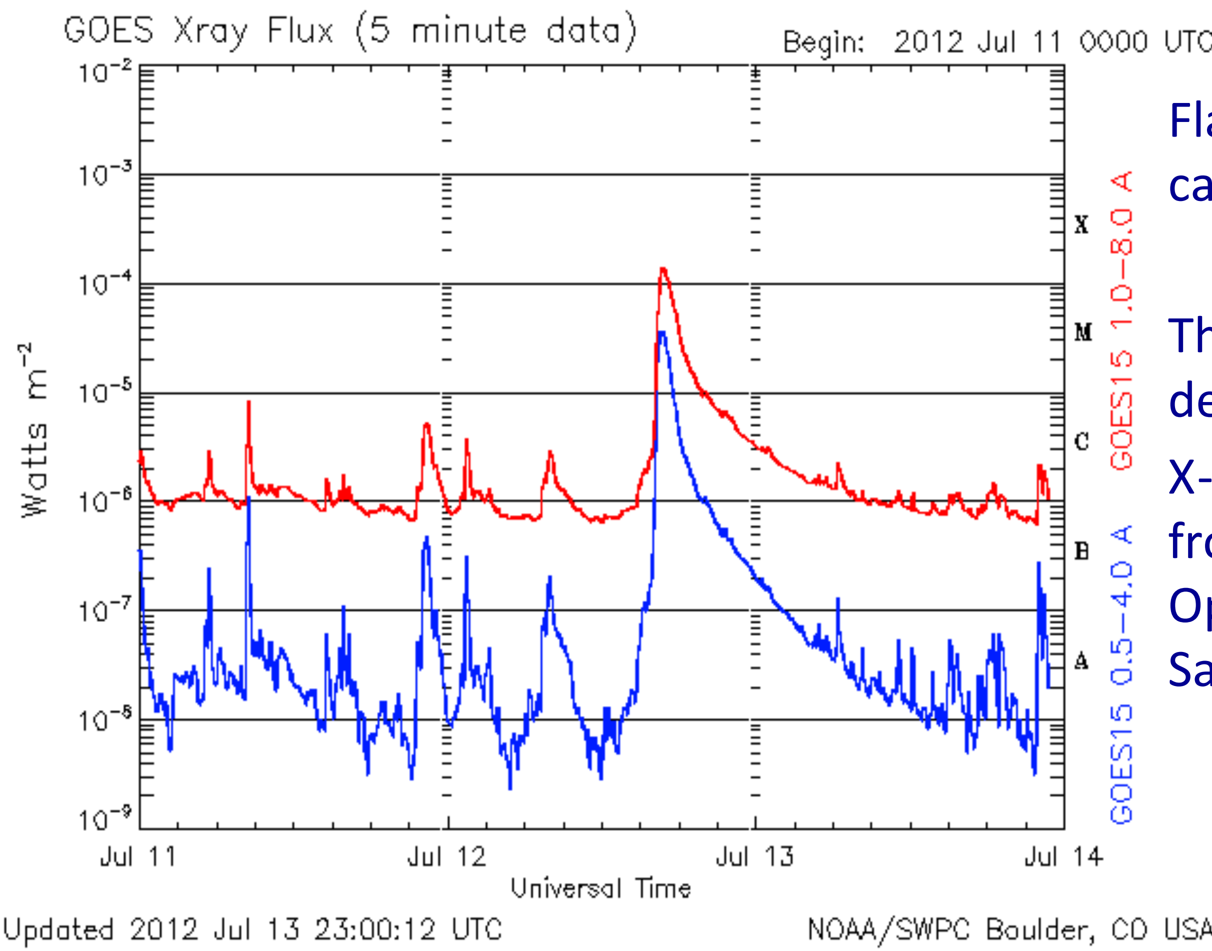


Image of solar atmosphere

The Task

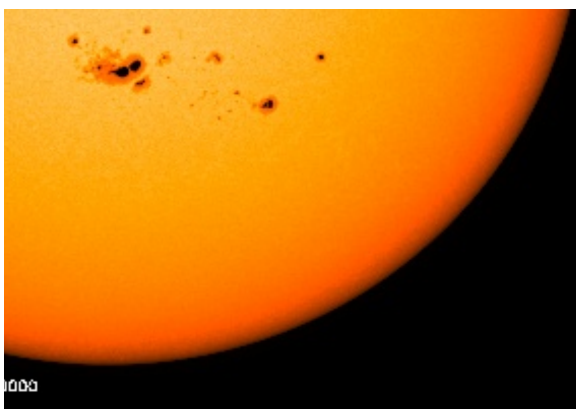


Flare can be categorized in 5 categories: A,B,C,M,X.

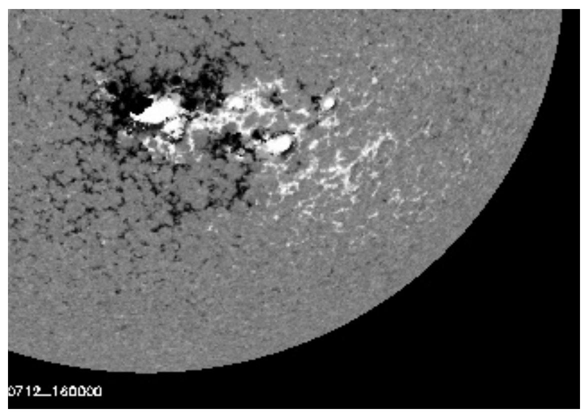
The category of the flare is decided according to the X-ray flux measurements from GOES (Geostationary Operational Environmental Satellites).

Flares can affect our society :

- interfere with radio communication and GPS signals
- damage space assets and perturb satellite launch
- harm astronauts in space



Continuum or white light image



Magnetogram at surface of Sun

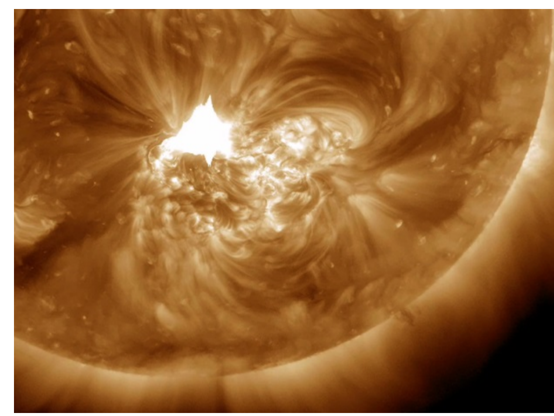


Image of solar atmosphere

PROBLEM DESCRIPTION

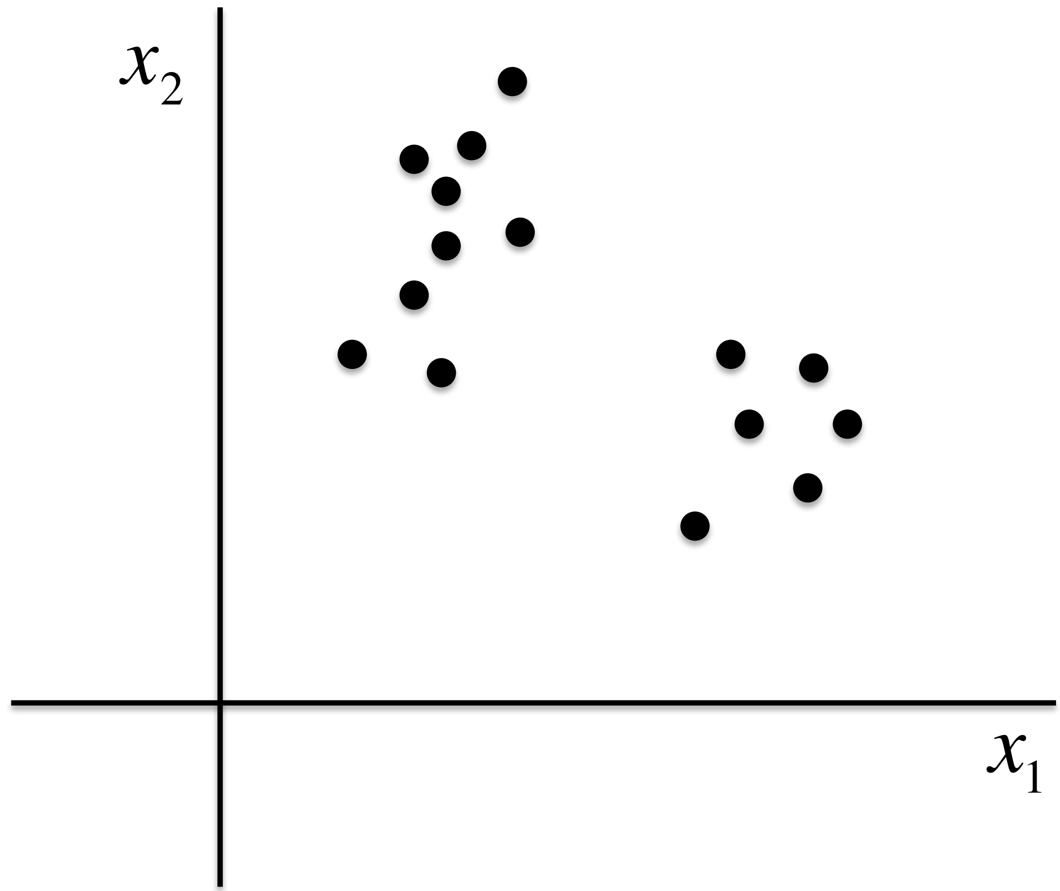
$$\bar{x} = (x_1, \ldots, x_m)^T$$

Descriptors

$$\{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_N\}$$

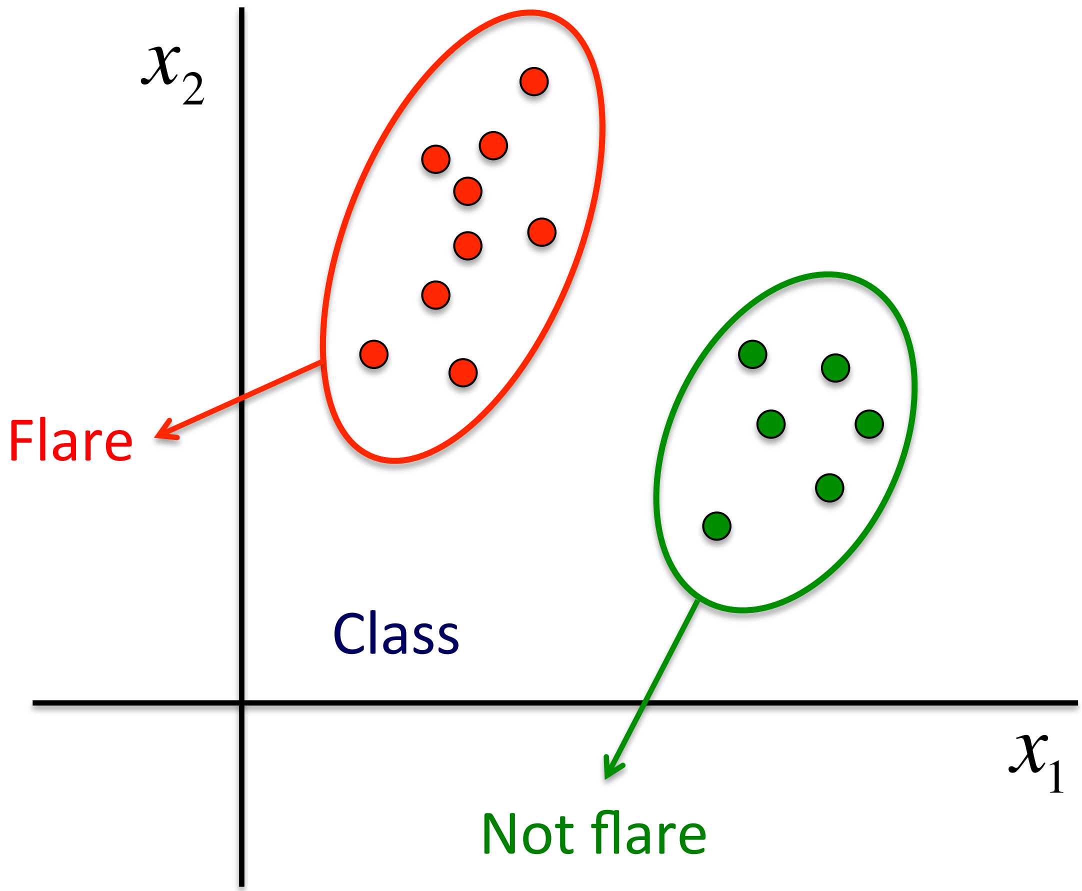
Dataset

$$\bar{x} = (x_1, x_2)^T$$



CLASSIFICATION PROBLEM

Class={Flare, Not flare}



CLASSIFICATION MODEL

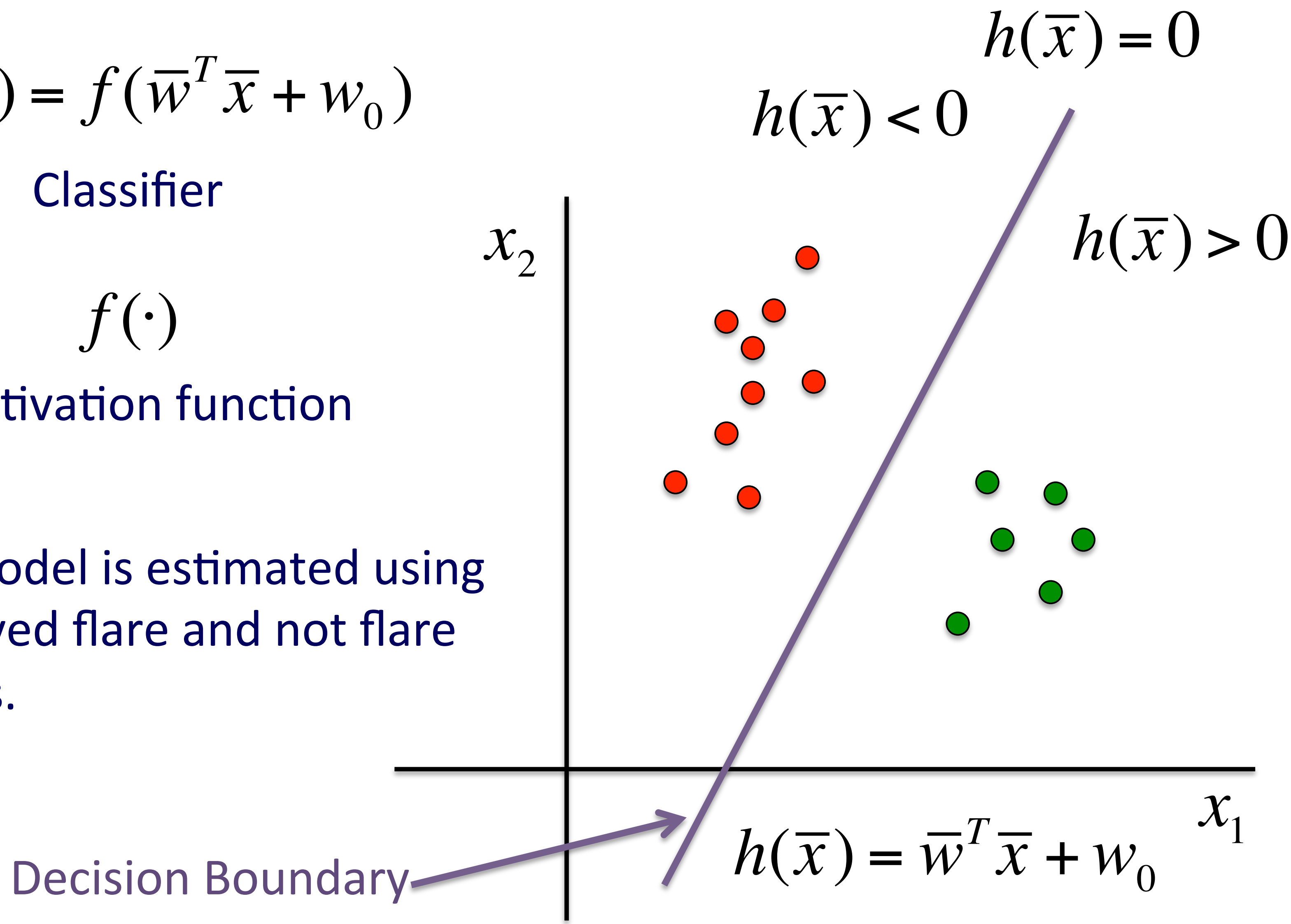
$$h(\bar{x}) = f(\bar{w}^T \bar{x} + w_0)$$

Classifier

$$f(\cdot)$$

Activation function

The model is estimated using observed flare and not flare events.



CLASSIFIER TESTING

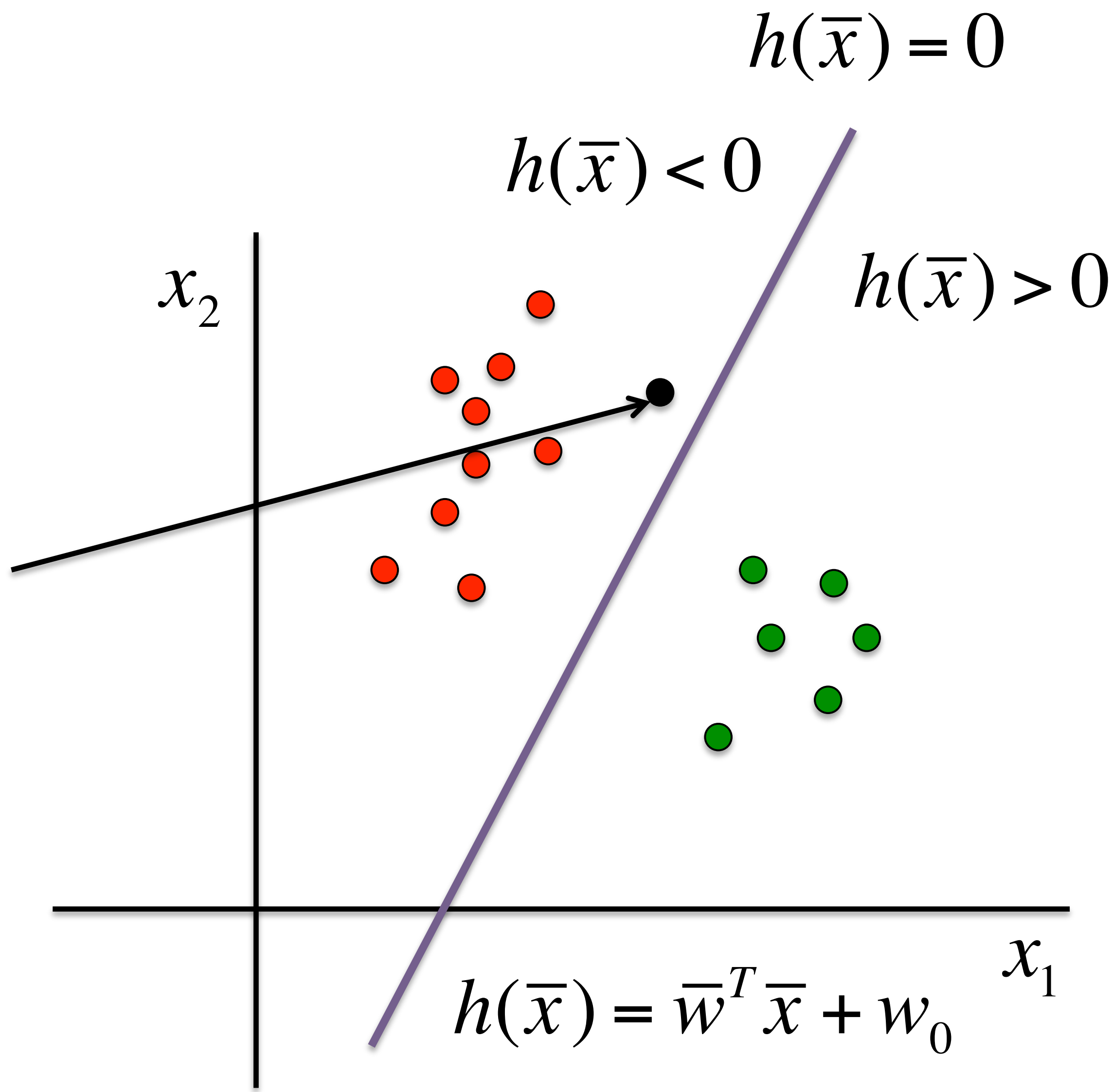
New Sample

$$h(\overline{x}) < 0$$

↓

Flare

Predicted Class



Data-Classifier-Output

Data

Classifier

Output

DATA

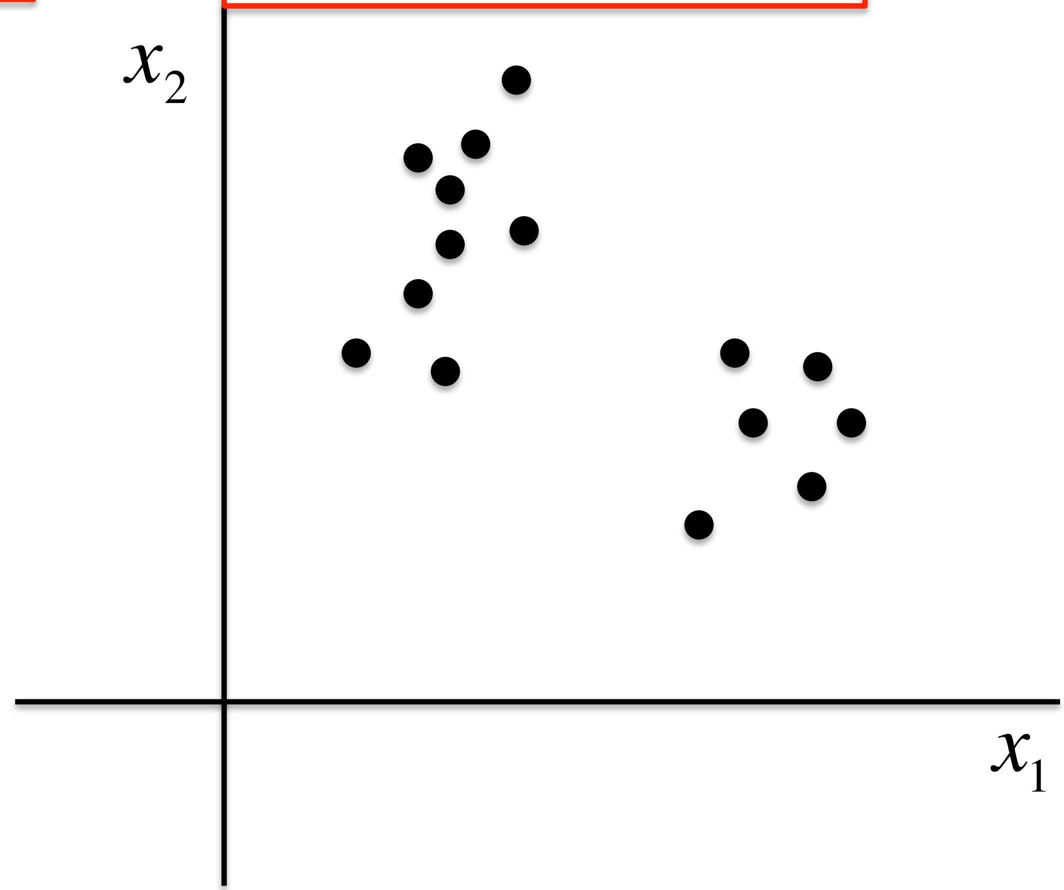
$$\overline{x} = (x_1, \ldots, x_m)^T$$

Descriptors

$$\{\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_N\}$$

Dataset

$$\overline{x} = (x_1, x_2)^T$$



DATA-CLASSIFIER-OUTPUT

Data

Classifier

Output

Flares are related to sunspots/active regions

We need:

- labeled data in order to train the classifier
- the definition and representation of solar features with high predictive power.

DATA REPOSITORY

NOAA's National Geophysical Data Center (NGDC) keeps record of data from several observatories around the world and holds one of the most comprehensive publicly available databases for solar features and activities.

It provides catalogues of

- Sunspot events
- Solar flare events

The availability of this data is important in order to train the classifiers and to provide the ground truth for the classification.

UCL Université catholique de Louvain

The Data

DATA REPOSITORY

Active Region

Joint USAF/NOAA Solar Region Summary
SRS Number 182 Issued at 0030Z on 01 Jul 2014
Report compiled from data received at SWO on 30 Jun
I. Regions with Sunspots. Locations Valid at 30/2400Z

NOAA#	Location	Lo	Area	Z	LL	NN	Mag Type
2096	N09W36	356	0020	Cso	03	02	Beta
2097	N12W30	350	0020	Hsx	01	01	Alpha
2100	N09E11	310	0060	Dai	09	12	Beta
2102	N12E41	280	0020	Cro	06	03	Beta
2104	S11E51	270	0350	Dkc	07	07	Beta-Gamma-Delta
2105	S06E11	310	0010	Cai	03	06	Beta
2106	N15E60	261	0020	Dai	05	04	Beta
2107	S20E59	262	0250	Dhi	10	09	Beta-Gamma

Solar flare

#Event	Begin	Max	End	Obs	Q	Type	Loc/Frq	Particulars		Reg#(NOAA#)
7570	0351	0355	0400	G15	5	XRA	1-8A	C1.5	5.6E-04	2104
7660	0510	0608	0633	G15	5	XRA	1-8A	C4.8	1.6E-02	2106
7680	0716	0737	0805	G15	5	XRA	1-8A	C6.6	1.3E-02	2107
7720	0857	0911	0931	G15	5	XRA	1-8A	C2.1	3.7E-03	2108

The Data

DATA REPOSITORY

Joint USAF/NOAA Solar Region Summary
SRS Number 182 Issued at 0030Z on 01 Jul 2014
Report compiled from data received at SWO on 30 Jun
I. Regions with Sunspots. Locations Valid at 30/2400Z

NOAA#	Location	Lo	Area	Z	LL	NN	Mag Type
2096	N09W36	356	0020	Cso	03	02	Beta
2097	N12W30	350	0020	Hsx	01	01	Alpha
2100	N09E11	310	0060	Dai	09	12	Beta
2102	N12E41	280	0020	Cro	06	03	Beta
2104	S11E51	270	0350	Dkc	07	07	Beta-Gamma-Delta
2105	S06E11	310	0010	Cai	03	06	Beta
2106	N15E60	261	0020	Dai	05	04	Beta
2107	S20E59	262	0250	Dhi	10	09	Beta-Gamma

Active Region

Solar flare

#Event	Begin	Max	End	Obs	Q	Type	Loc/Frq	Particulars		Reg#(NOAA#)
#-------										
7570	0351	0355	0400	G15	5	XRA	1-8A	C1.5	5.6E-04	2104
7660	0510	0608	0633	G15	5	XRA	1-8A	C4.8	1.6E-02	2106
7680	0716	0737	0805	G15	5	XRA	1-8A	C6.6	1.3E-02	2107
7720	0857	0911	0931	G15	5	XRA	1-8A	C2.1	3.7E-03	2108

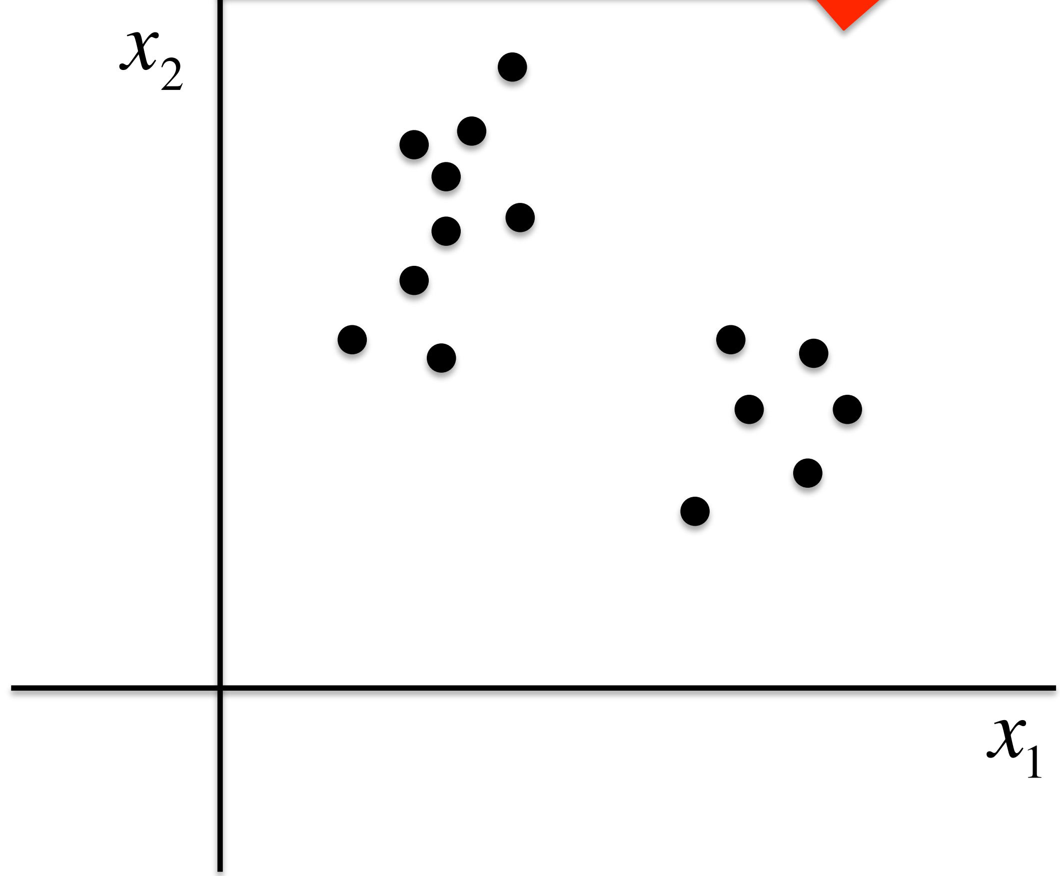
DATA

$$\bar{x} = (x_1, \ldots, x_m)^T$$

Descriptors

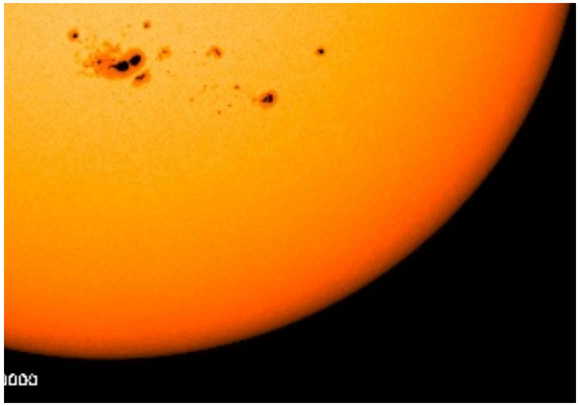
$$\{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_N\}$$

Dataset

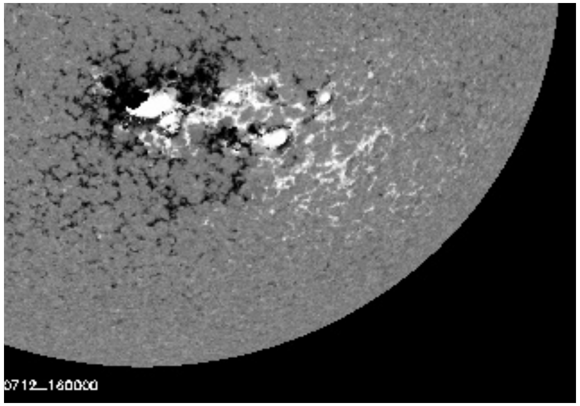


$$\bar{x} = (x_1, x_2)^T$$

The Data



Continuum or white light image



Magnetogram at surface of Sun

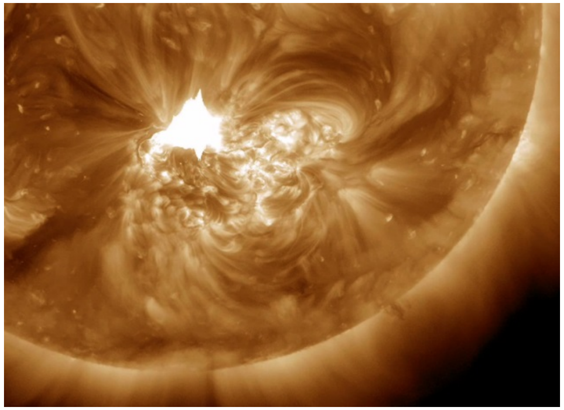
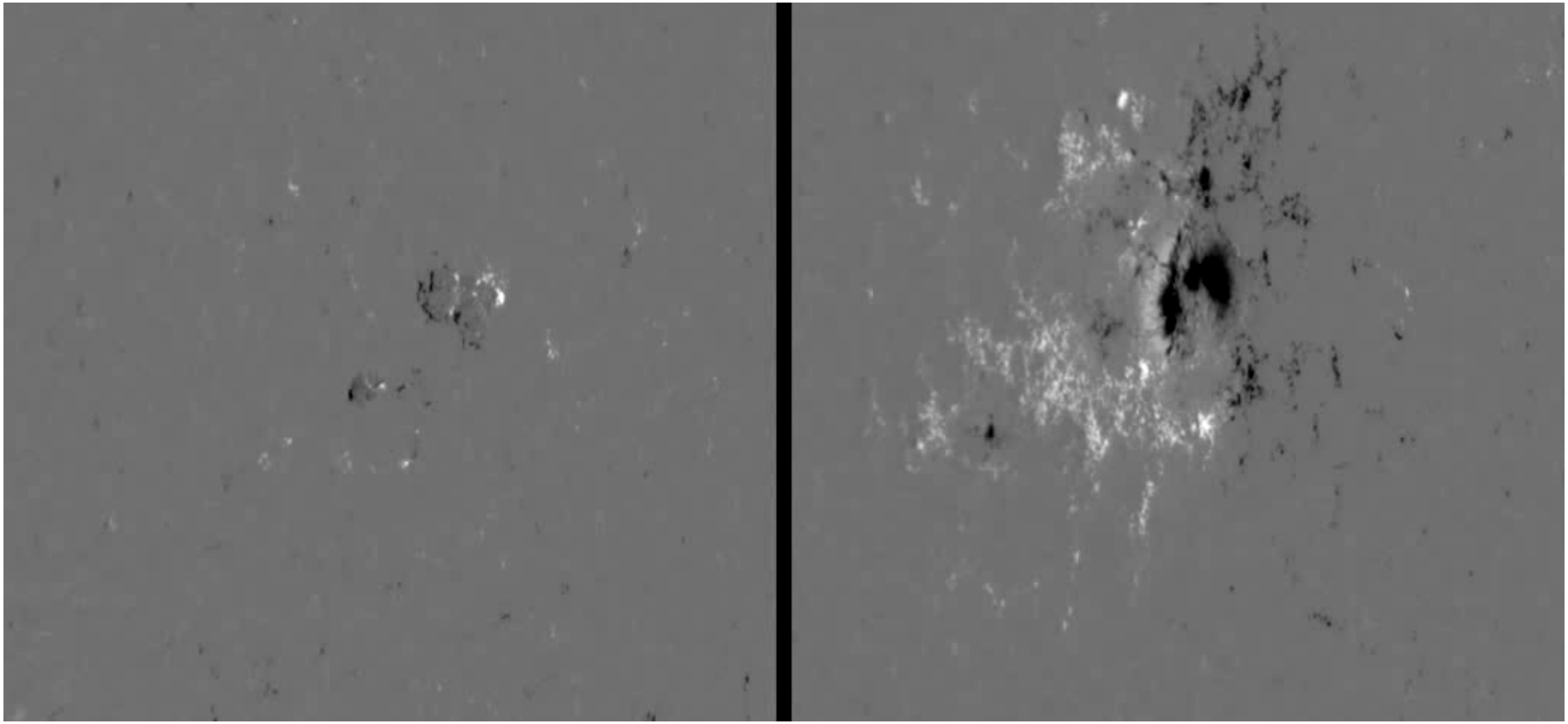


Image of solar atmosphere

The Data

How to generate the descriptors



McIntosh Classification

The 3 component McIntosh is based on the general form 'Zpc', where 'Z' is the modified Zurich Class, 'p' describes the penumbra of the principal spot, and 'c' describes the distribution of spots in the interior of the group.

There are 60 valid McIntosh classification

Examples: Dao, Eao, Ekc, Fai, Fkc, Fko.

The Data

McIntosh Classification

Z-values: (Modified Zurich Sunspot Classification).

- A - A small single unipolar sunspot. Representing either the formative or final stage of evolution.
- B - Bipolar sunspot group with no penumbra on any of the spots.
- C - A bipolar sunspot group. One sunspot must have penumbra.
- …

p-values:

- x - no penumbra (group class is A or B)
- r - rudimentary penumbra partially surrounds the largest spot.
- s - small, symmetric (like Zurich class J)
- …

c-values

- x - undefined for unipolar groups (class A and H)
- o - open. Few, if any, spots between leader and follower
- …

Example of proposed features

Qu et Al. '03 (H-alpha images):

- Feature 1: mean brightness of the frame
- Feature 2: standard deviation of brightness
- Feature 3: variation of mean brightness between consecutive images
- Feature 4: absolute brightness of a key pixel
- Feature 5: radial position of the key pixel
- Feature 6: contrast between the key pixel and the minimum value of its neighbors in a 7 by 7 window
- Feature 8: standard deviation of the pixels in a 50 by 50 window
- Feature 9: difference of the mean brightness of the 50 by 50 window between the current and the previous images.

Li et Al. '06 (Continuum + Manual Classification):

- the area of the sunspot group
- magnetic classification
- McIntosh classification
- 10 cm radio flux

(Cui'06, Yu '09) (Magnetogram)

- maximum horizontal gradient
- the length of the neutral line
- the number of singular points

The Data

Example of proposed features

(Jing et Al '06) (Magnetogram):

- the mean spatial magnetic field gradient at the strong-gradient magnetic neutral line
- the length of a strong-gradient magnetic neutral line
- the total magnetic energy dissipation

(Song '09) (Magnetogram):

- the total unsigned magnetic
- the length of the strong-gradient neutral line
-
- the total magnetic dissipation

Which Descriptors ?

No Well Defined Set of Descriptors

Find a set of features highly discriminative in the task is an open challenge.

Existing descriptors collect information from both the whole image and from local portion of the image.

Limited attention has been given to the temporal evolution of the phenomena.

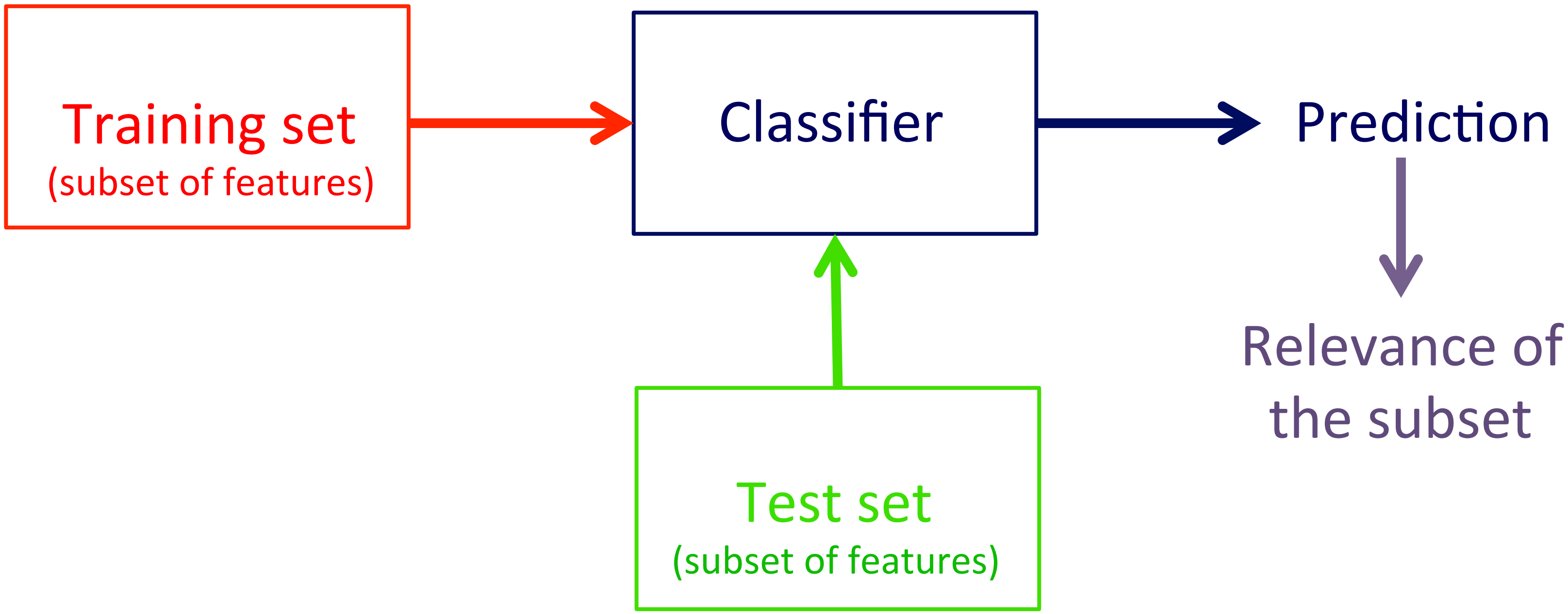
Features Selection

Feature selection aims at reducing the dimensionality of the data while preserving the interpretation of the original features

- **Filters methods** use only the data + class labels:
  - simple, fast, generally univariate
- **Wrappers** take the performance of the classifier into account
  - Multivariate as soon as the classifier is multivariate
  - Often computing intensive
- **Embedded methods** take the structure of the classifier into account
  - More elegant and often faster than wrappers, not always better in terms of performance
  - A way to get an insight into a black-box classifier
  - Convex optimization plays a key role

- A **feature relevance** can be defined according to the distance between the average feature value in each class
- The larger the distance the better, relatively to standard deviations
- The distance is computed according to a t-Test statistics
- *p-values* assess the significance of the difference between the two class means
- A feature is selected if its associated *p-value* is below a prescribed threshold

Wrappers



- Estimate a classifier from a given subset of all possible features
- Select the feature subset that optimizes the performance of the classifier (usually on an independent validation set)
  - Feature selection depends on the evaluation protocol of the classifier
  - There are $O(2^p)$ possible subsets

Embedded Methods

- Define the feature selection and the classifier estimation as a combined optimization process

- The features are selected as a by-product of the estimated classifier and its parameters

$$h(\overline{x}) = \overline{w}^T \overline{x} + w_0$$

- $|w_j|$ is a measure of the importance of the jth feature

DATA

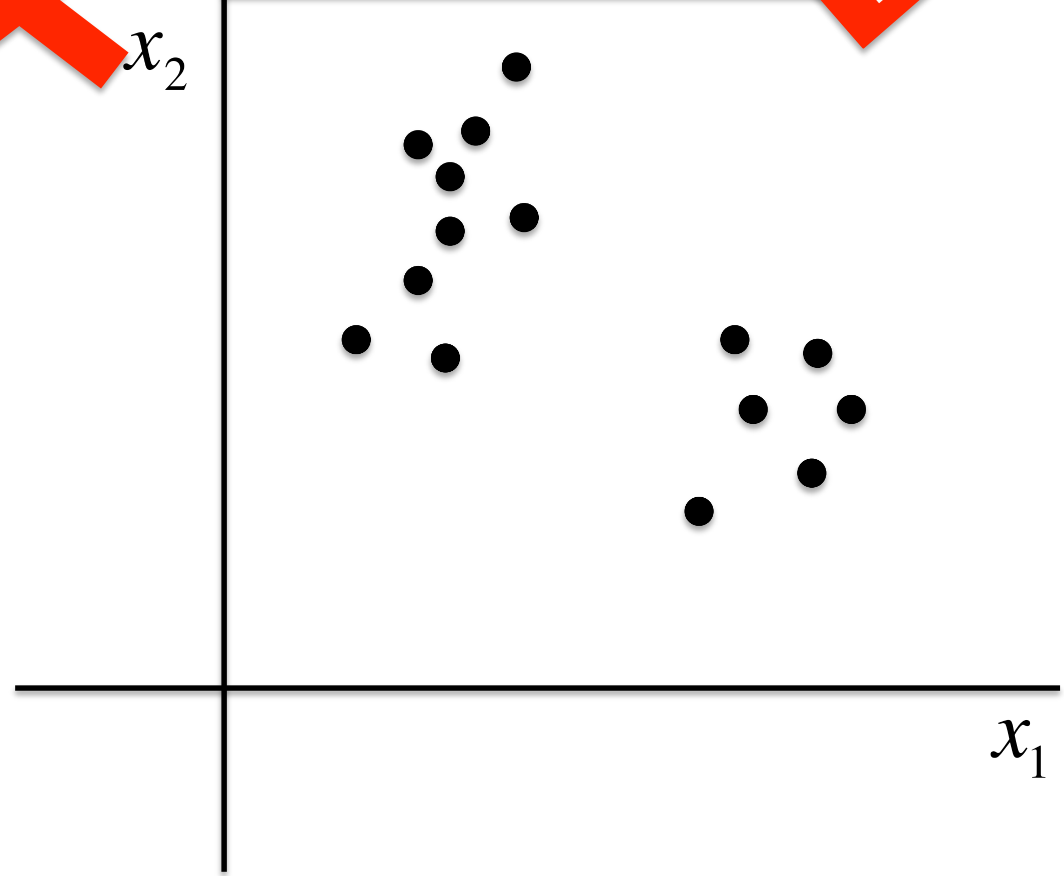
$$\overline{x} = (x_1, \ldots, x_m)^T$$

Descriptors

$$\{\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_N\}$$

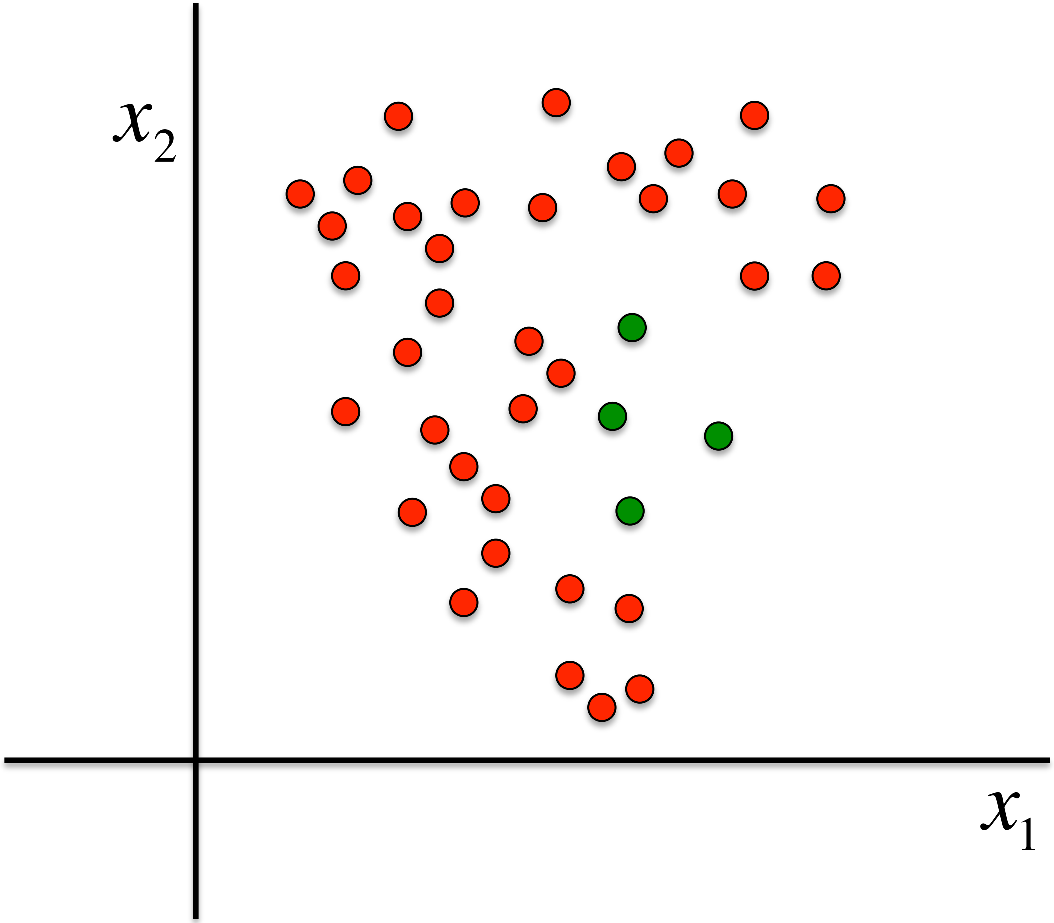
Dataset

$$\overline{x} = (x_1, x_2)^T$$



Another issue is the presence of a skeweness among the distribution of the samples.

Only the 10% of the active regions produce an M- or X-flare!

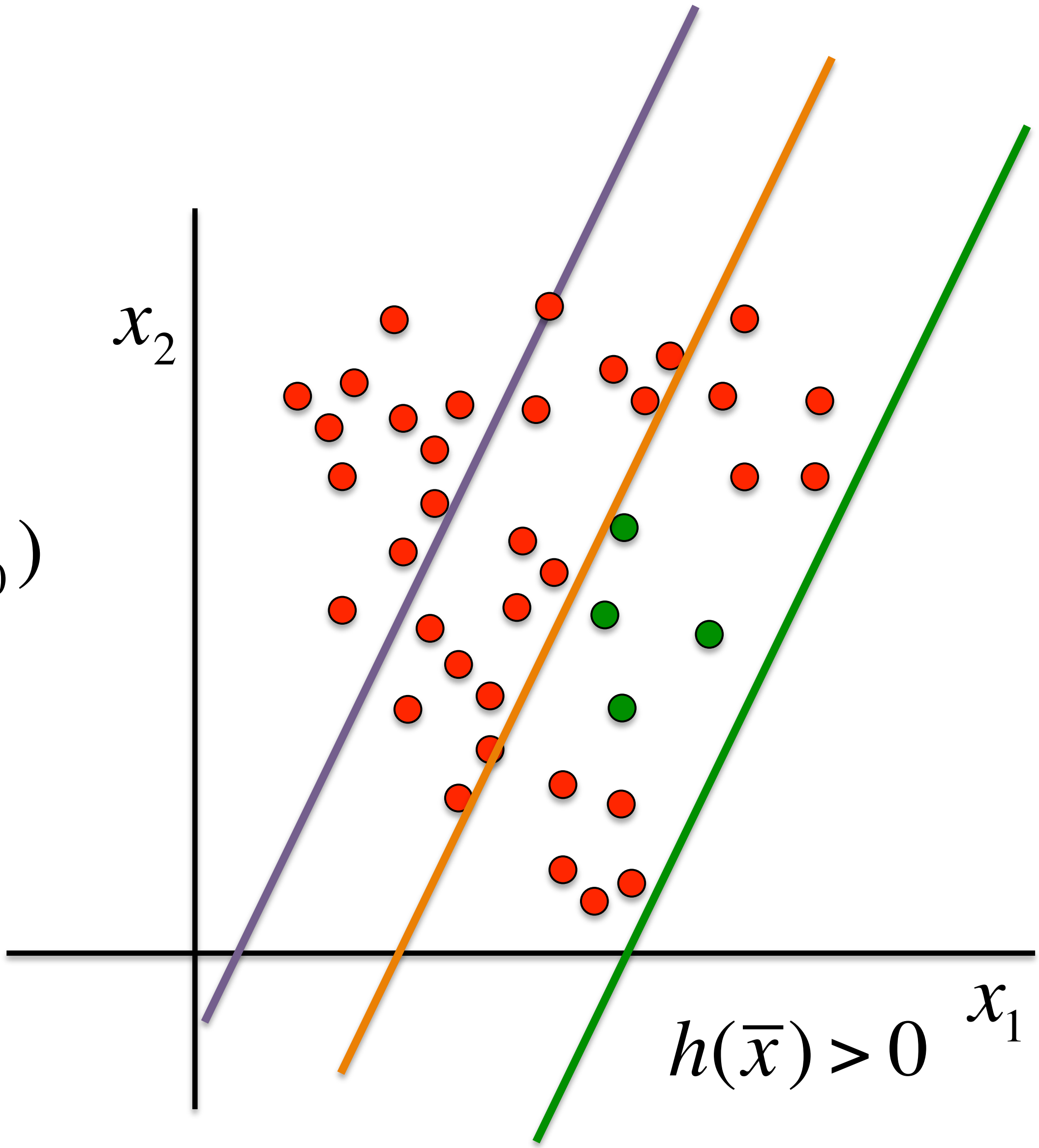

x_2
x_1

IMBALANCED PROBLEM

$$h(\overline{x}) = f(\overline{w}^T \overline{x} + w_0)$$

$$L(\omega(\overline{x}), h(\overline{x}))$$

Loss Function

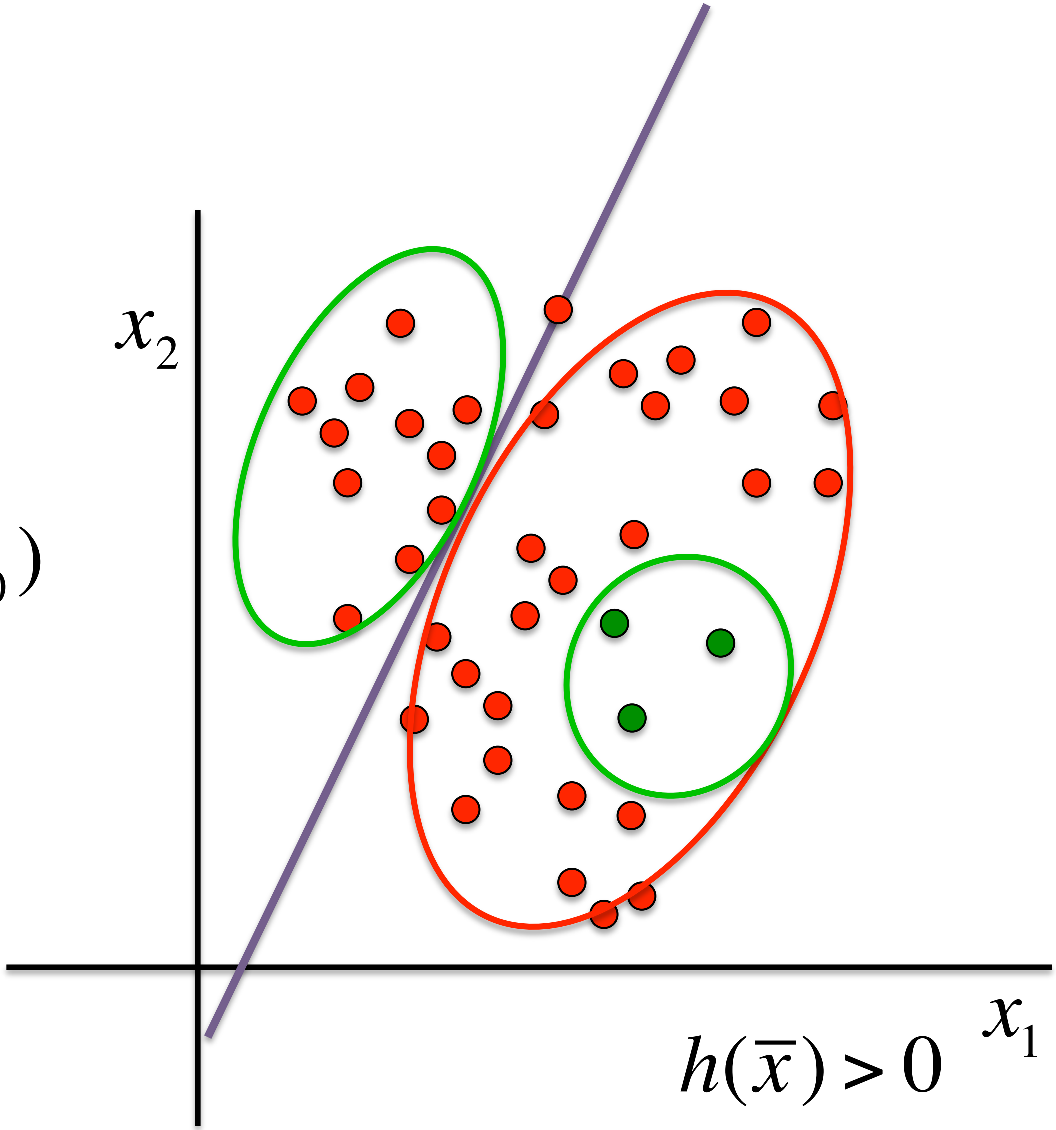


IMBALANCED PROBLEM

$$h(\overline{x}) = f(\overline{w}^T \overline{x} + w_0)$$

$$L(\omega(\overline{x}), h(\overline{x}))$$

Loss Function



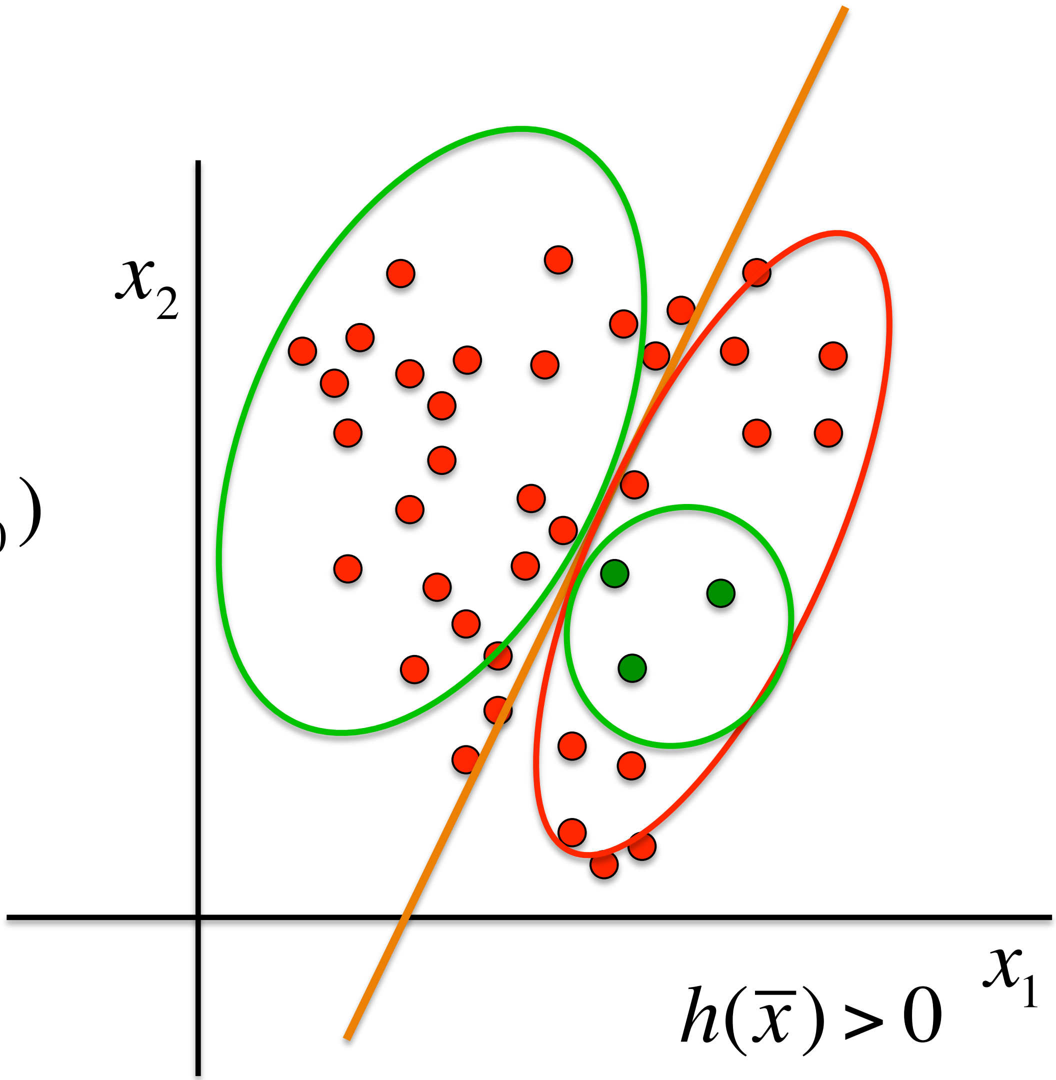
Imbalanced Problem

$$h(\overline{x}) = f(\overline{w}^T \overline{x} + w_0)$$

↑

$$L(\omega(\overline{x}), h(\overline{x}))$$

Loss Function

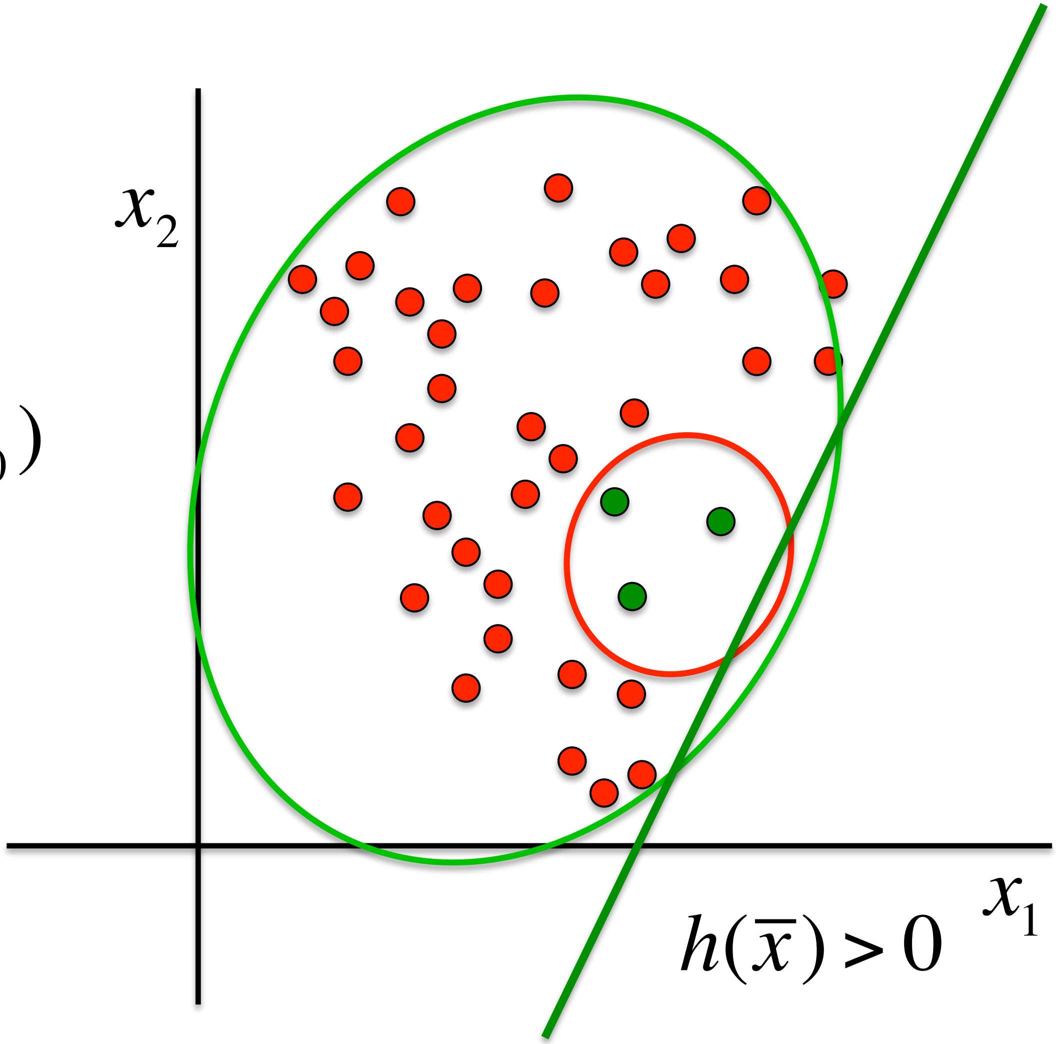


IMBALANCED PROBLEM

$$h(\overline{x}) = f(\overline{w}^T \overline{x} + w_0)$$

$$L(\omega(\overline{x}), h(\overline{x}))$$

Loss Function

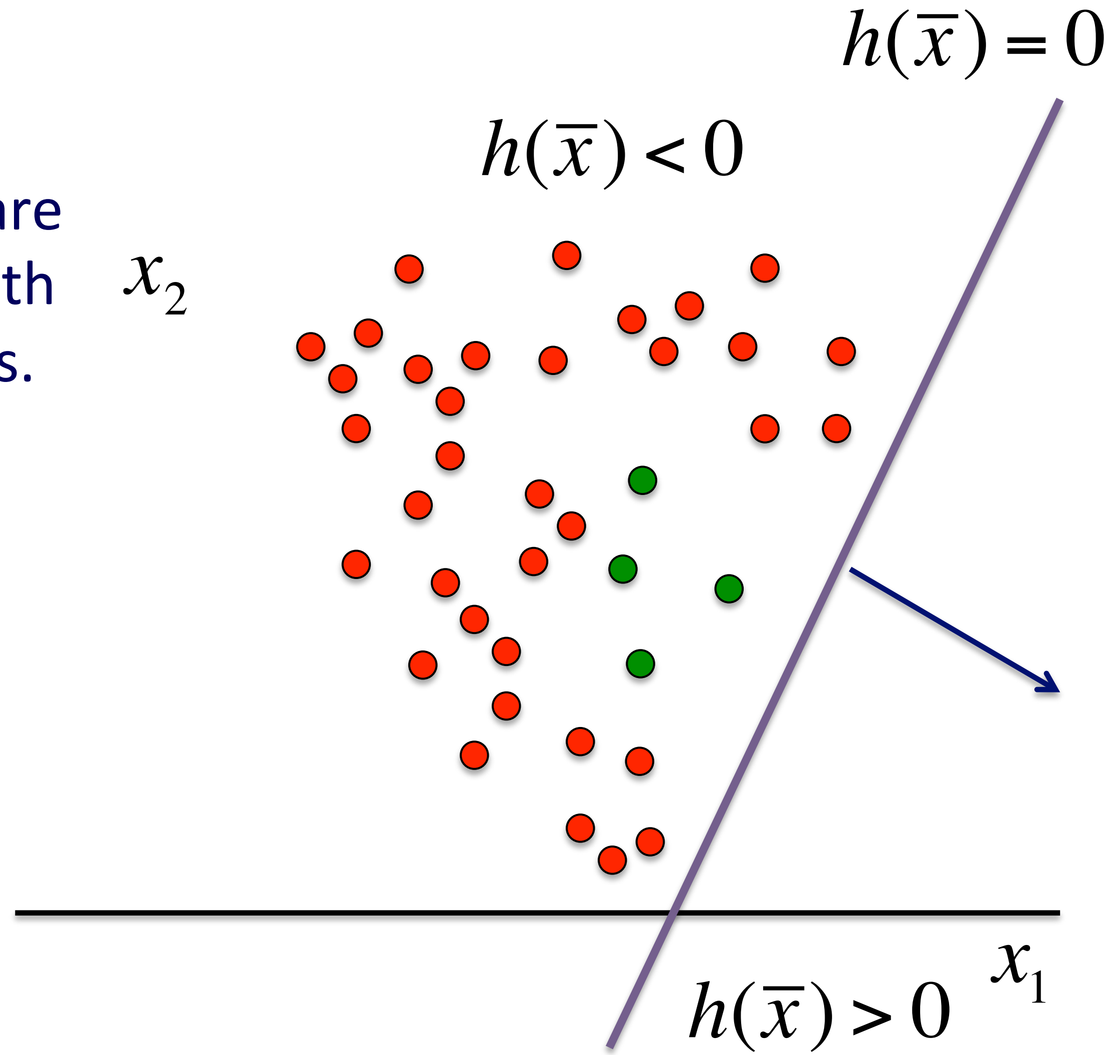


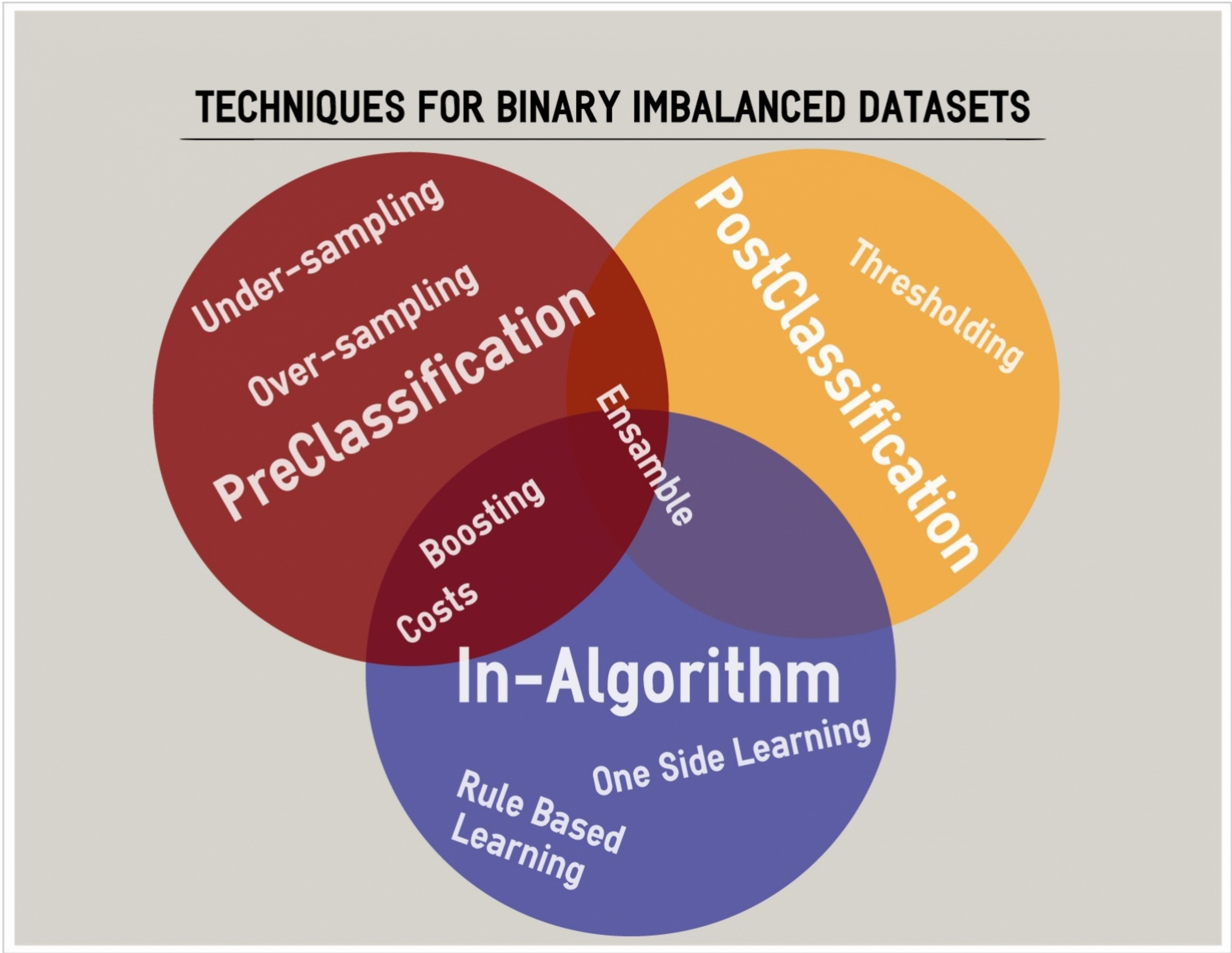
Imbalanced Problem

Imbalanced Data

One or more classes are underrepresented with respect to the others.

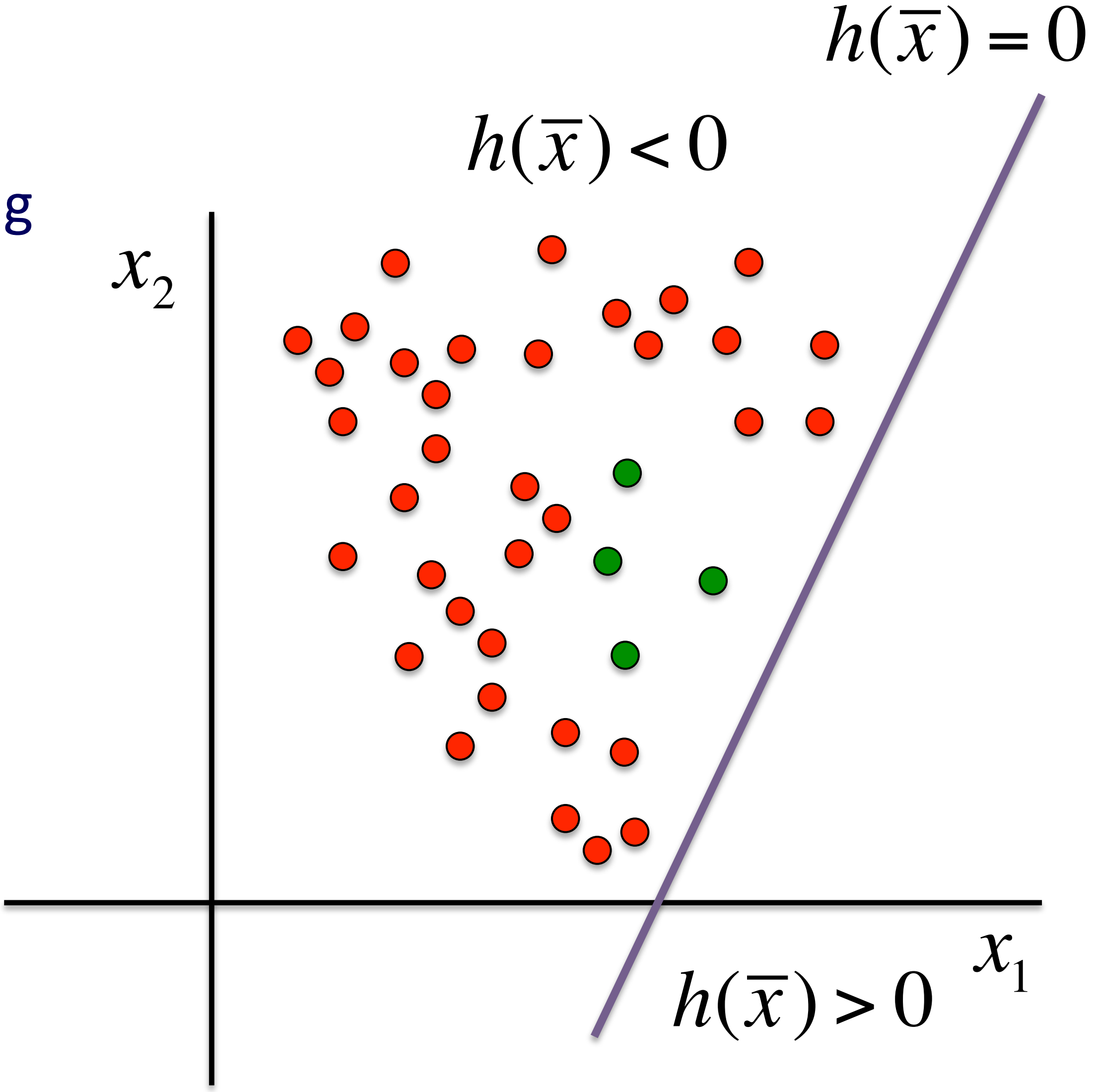
Almost all the real world domains are imbalanced




TECHNIQUES FOR BINARY IMBALANCED DATASETS
PreClassification
Under-sampling
Over-sampling
PostClassification
Thresholding
Ensamble
Boosting
Costs
In-Algorithm
One Side Learning
Rule Based Learning

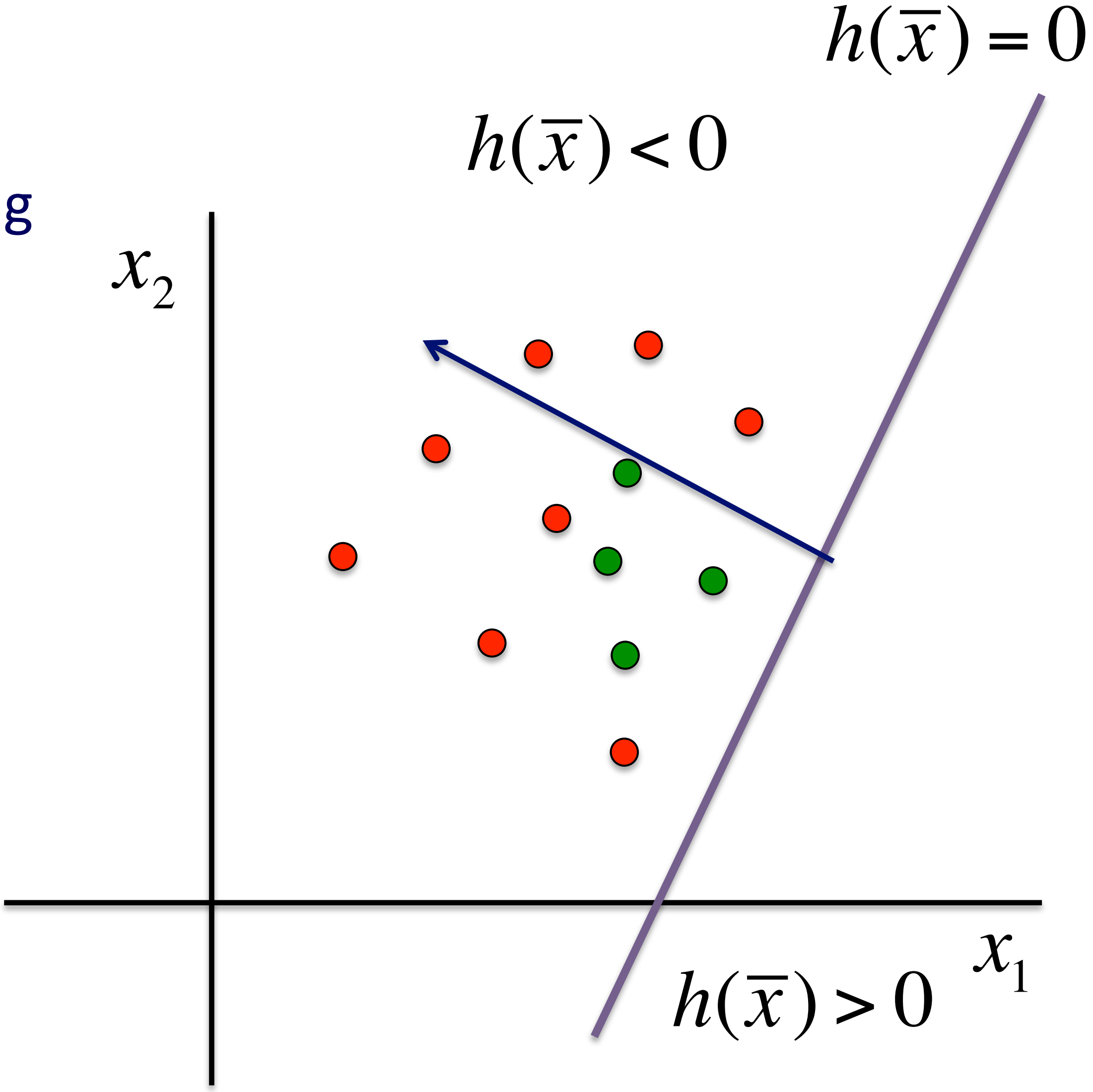
UNDER-SAMPLING

It consists in removing samples from the majority class.



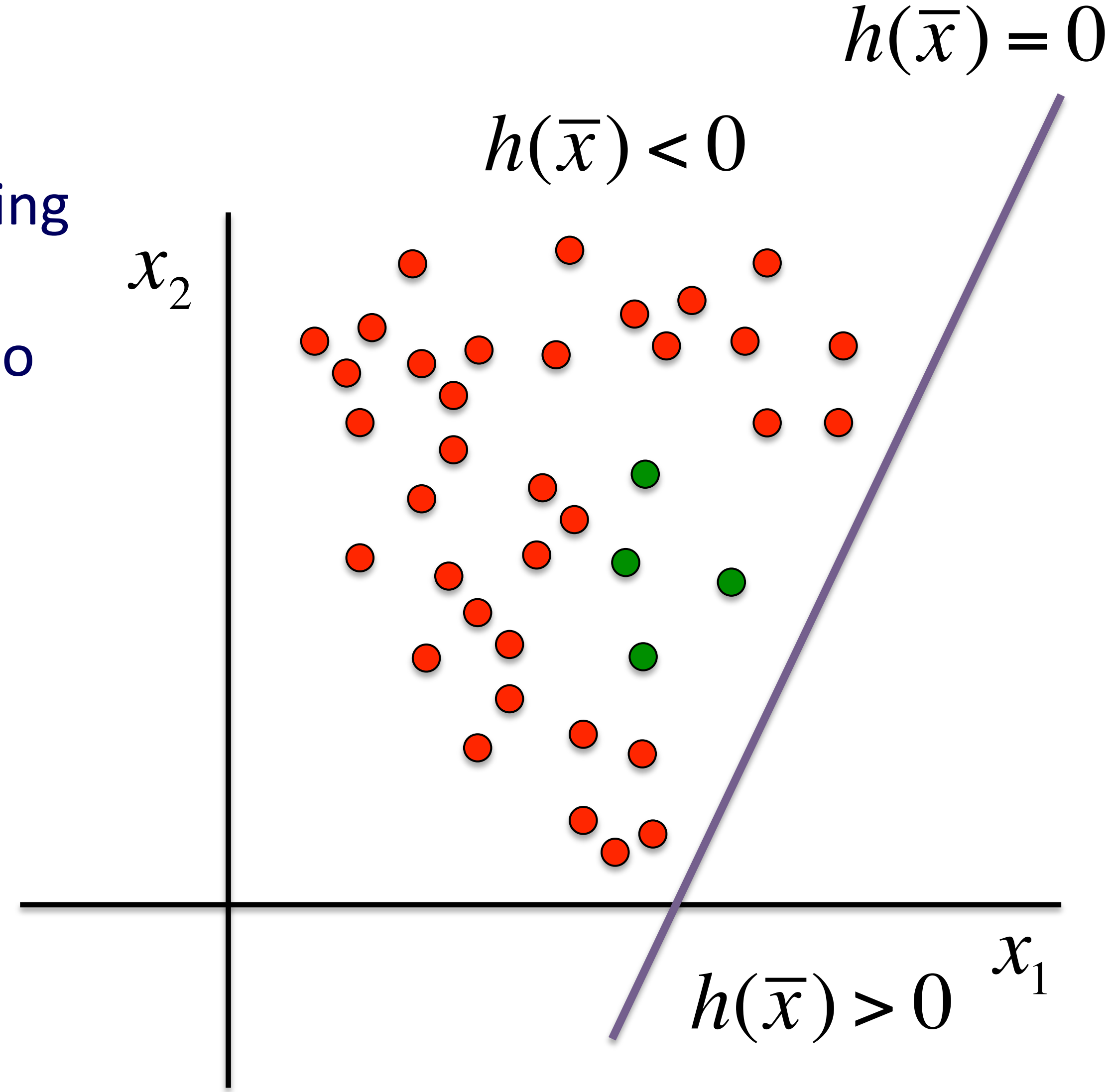
Under-Sampling

It consists in removing samples from the majority class.



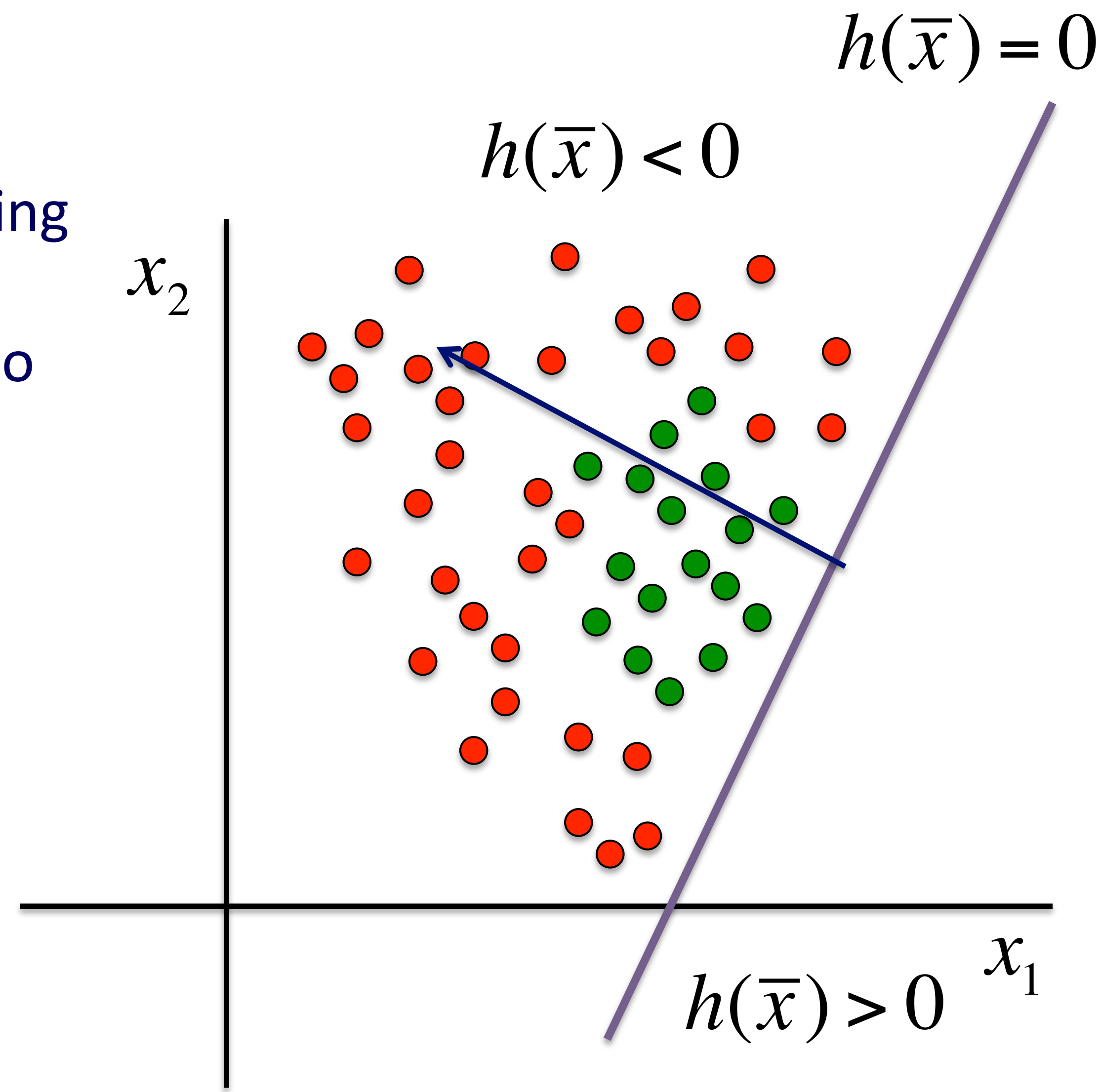
Over-Sampling

It consists in duplicating or generating new samples belonging to the minority class.



OVER-SAMPLING

It consists in duplicating or generating new samples belonging to the majority class.

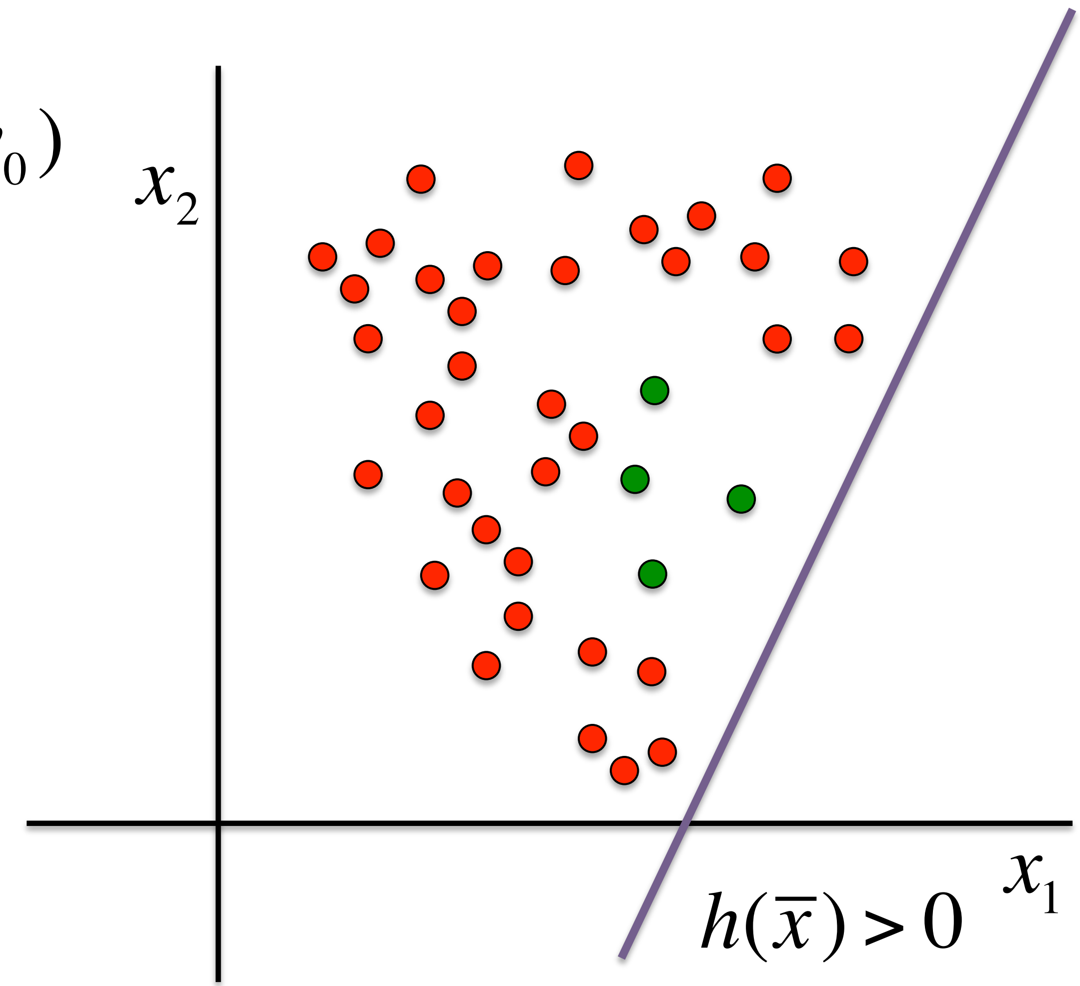


IN ALGORITHM

$$h(\overline{x}) = f(\overline{w}^T\overline{x} + w_0)$$

$$L(\omega(\overline{x}), h(\overline{x}))$$

Loss Function

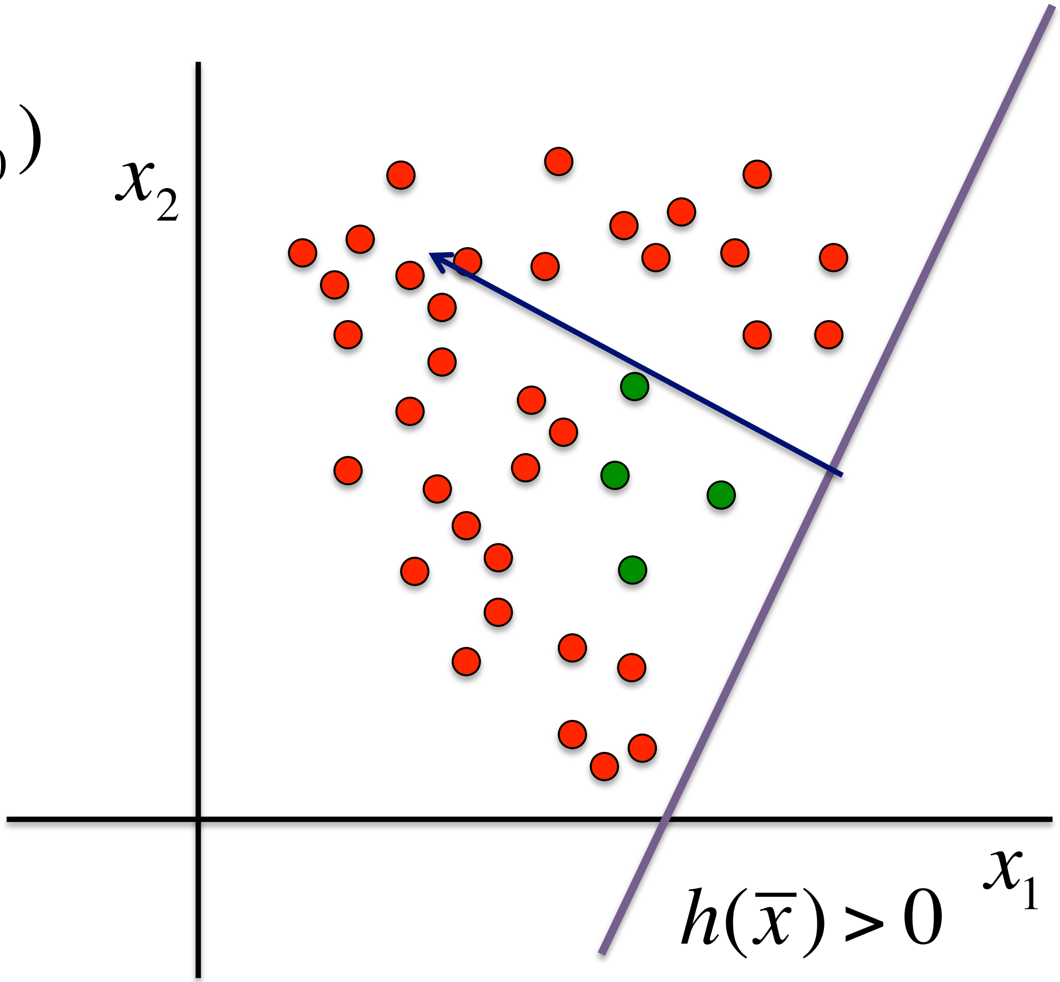


In Algorithm

$$h(\overline{x}) = f(\overline{w}^T \overline{x} + w_0)$$

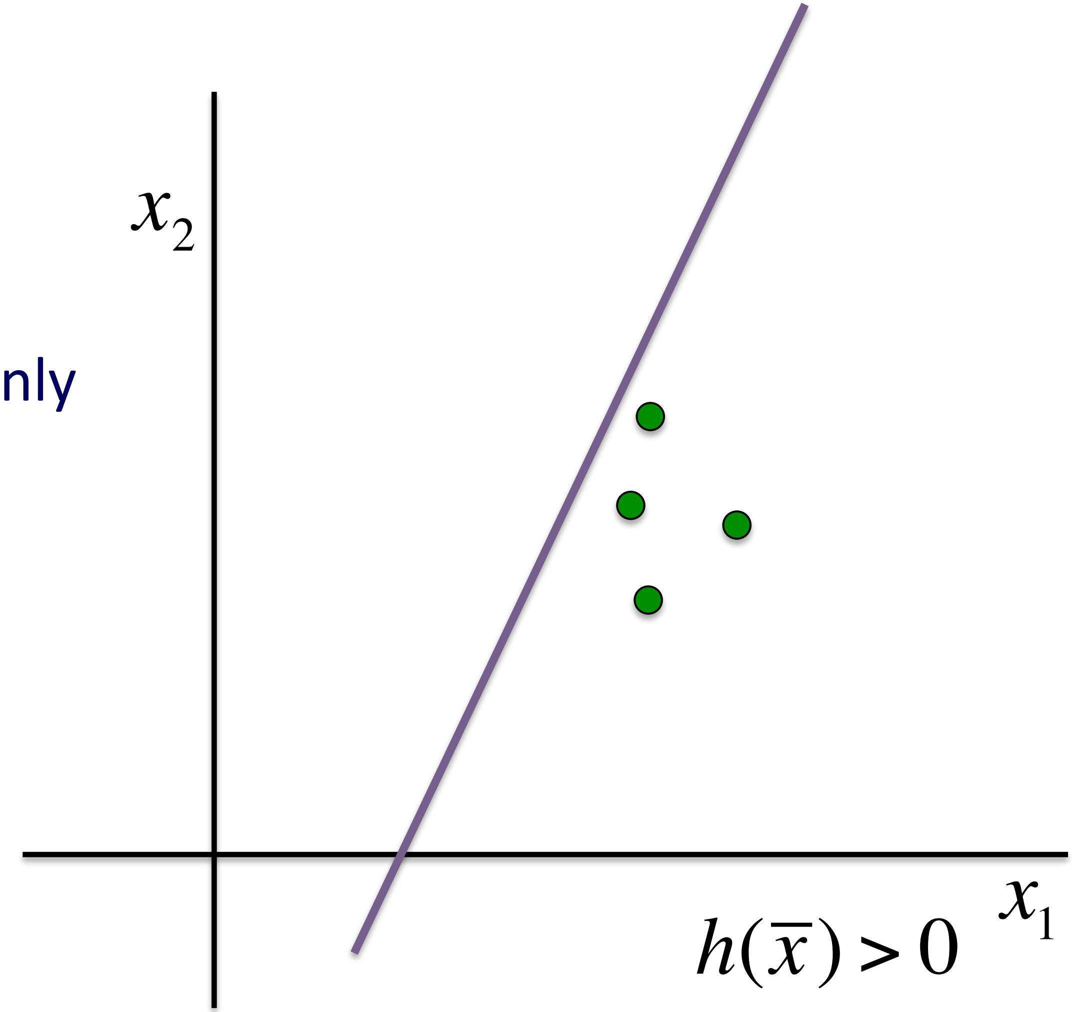
$$L(\omega(\overline{x}), h(\overline{x}))$$

Loss Function

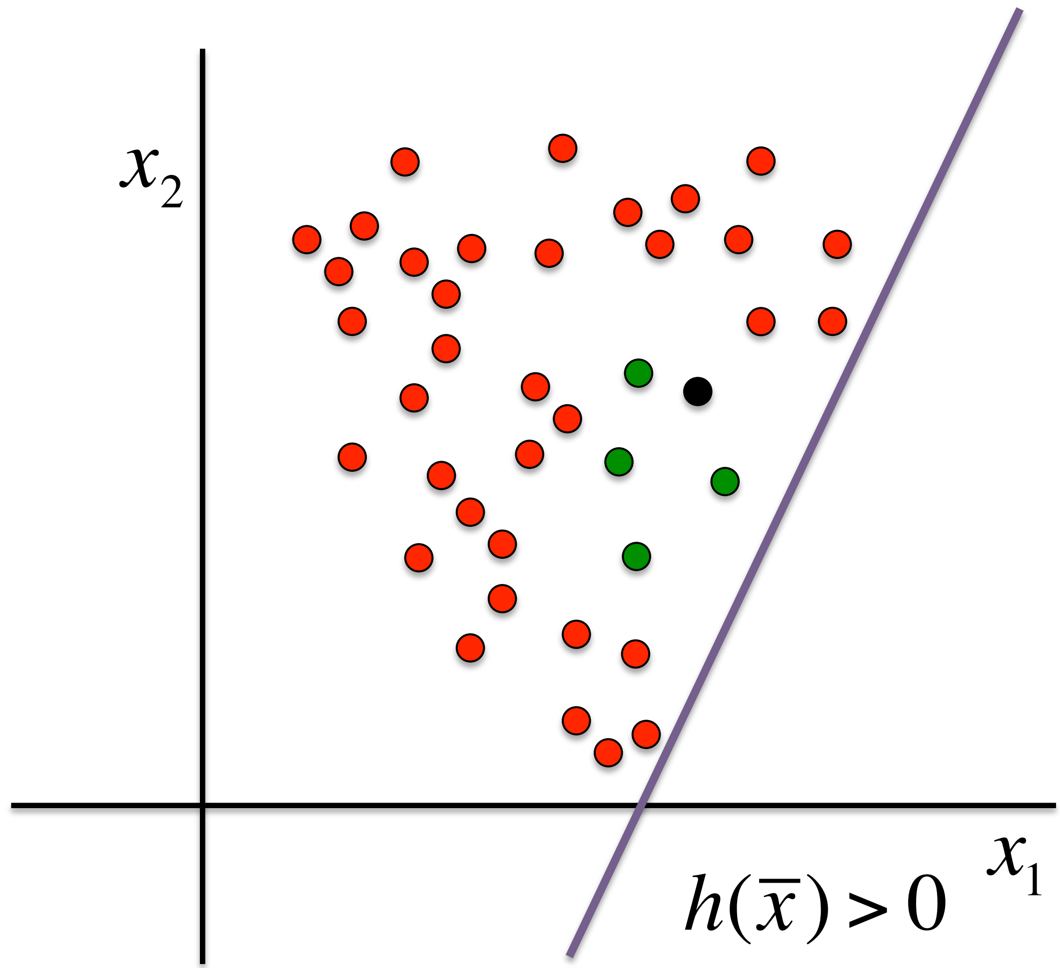


One Side Learning

The system focuses only on samples belonging to the minority class

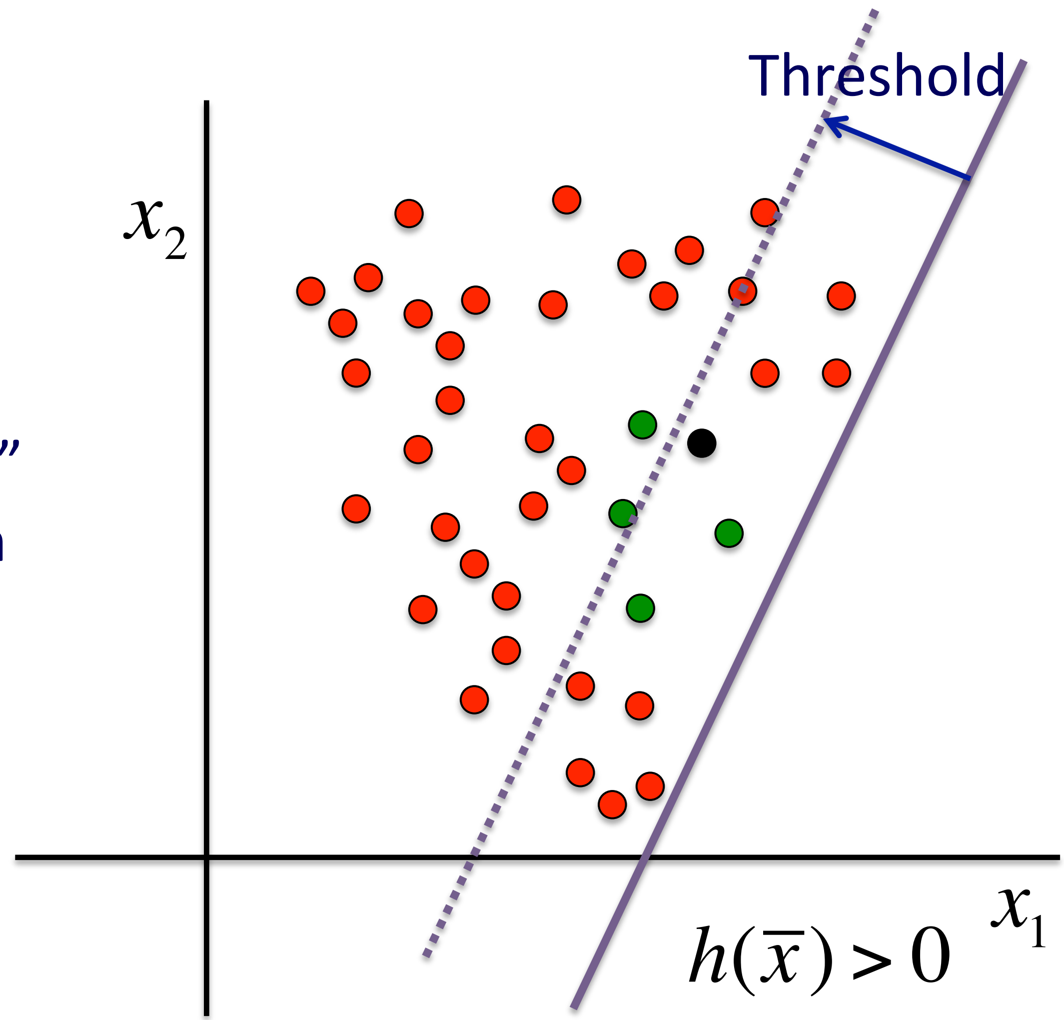


Thresholding



Thresholding

The decision of the system is "modified" after the prediction



DATA-CLASSIFIER-OUTPUT

Data

Classifier

Output

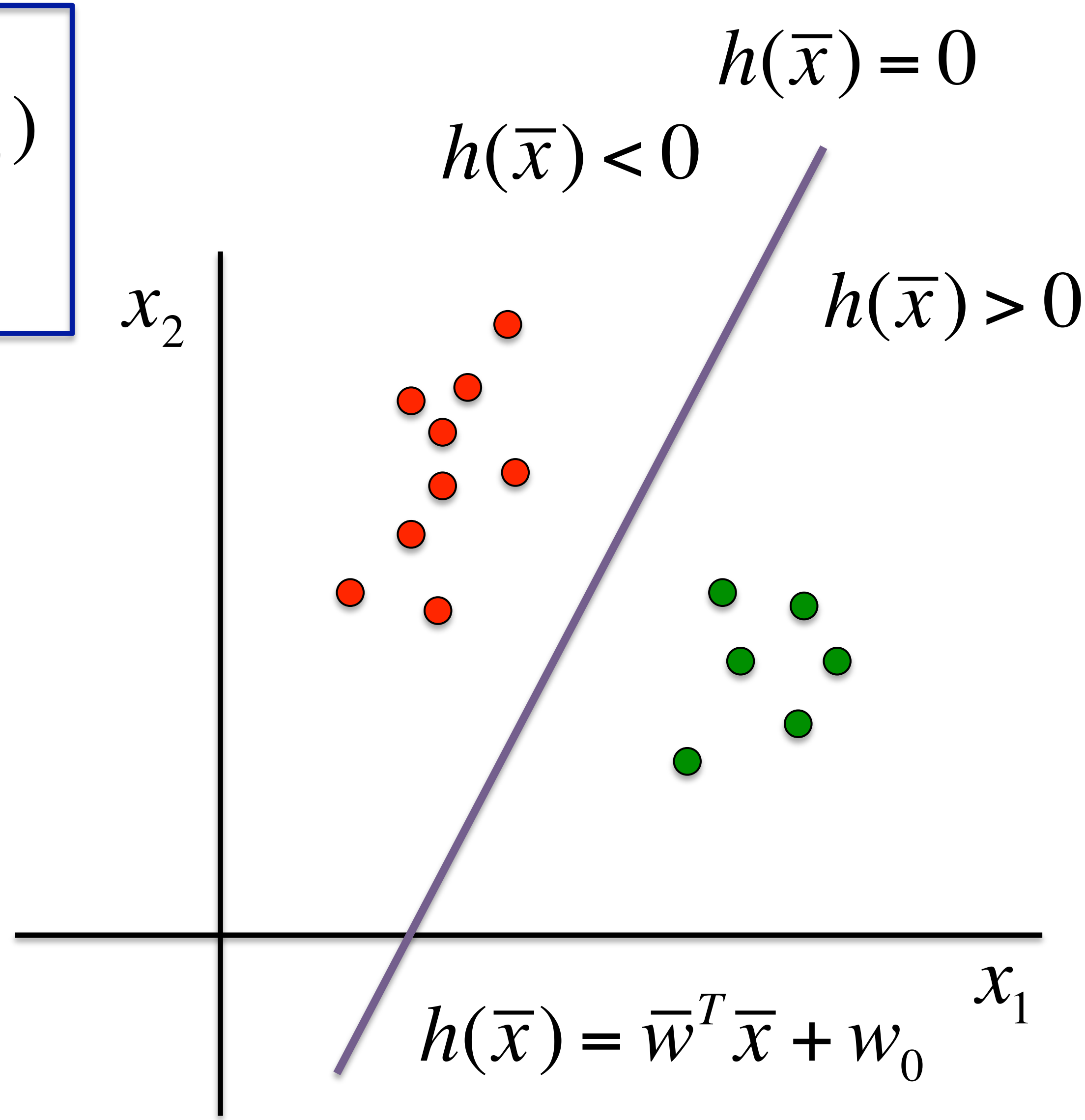
- Support Vector Machine (Qu ’03, Qahwaji ‘06)
- Radial Basis Function (Qu ’03, Qahwaji ‘06)
- Multi-Layer perceptron (Qu ‘03)
- Cascade-Correlation Neural Networks (Qahwaji ’06)
- SVM+kNN (Li ’07)
- Neural Network (Colak ‘08)
- Logistic regression (Song ’09)
- C4.5 decision tree (Yu ‘09)

Classifier

$$h(\bar{x}) = f(\bar{w}^T \bar{x} + w_0)$$

Classifier

The last step!



Data

Classifier

Output

Which metrics is reliable to evaluate the prediction of an automatic system?

Confusion Matrix

Ground Truth

Predicted Class

	p	n
P̃	**T**rue **P**ositive	**F**alse **P**ositive
Ñ	**F**alse **N**egative	**T**rue **N**egative
	P	N

ACCURACY

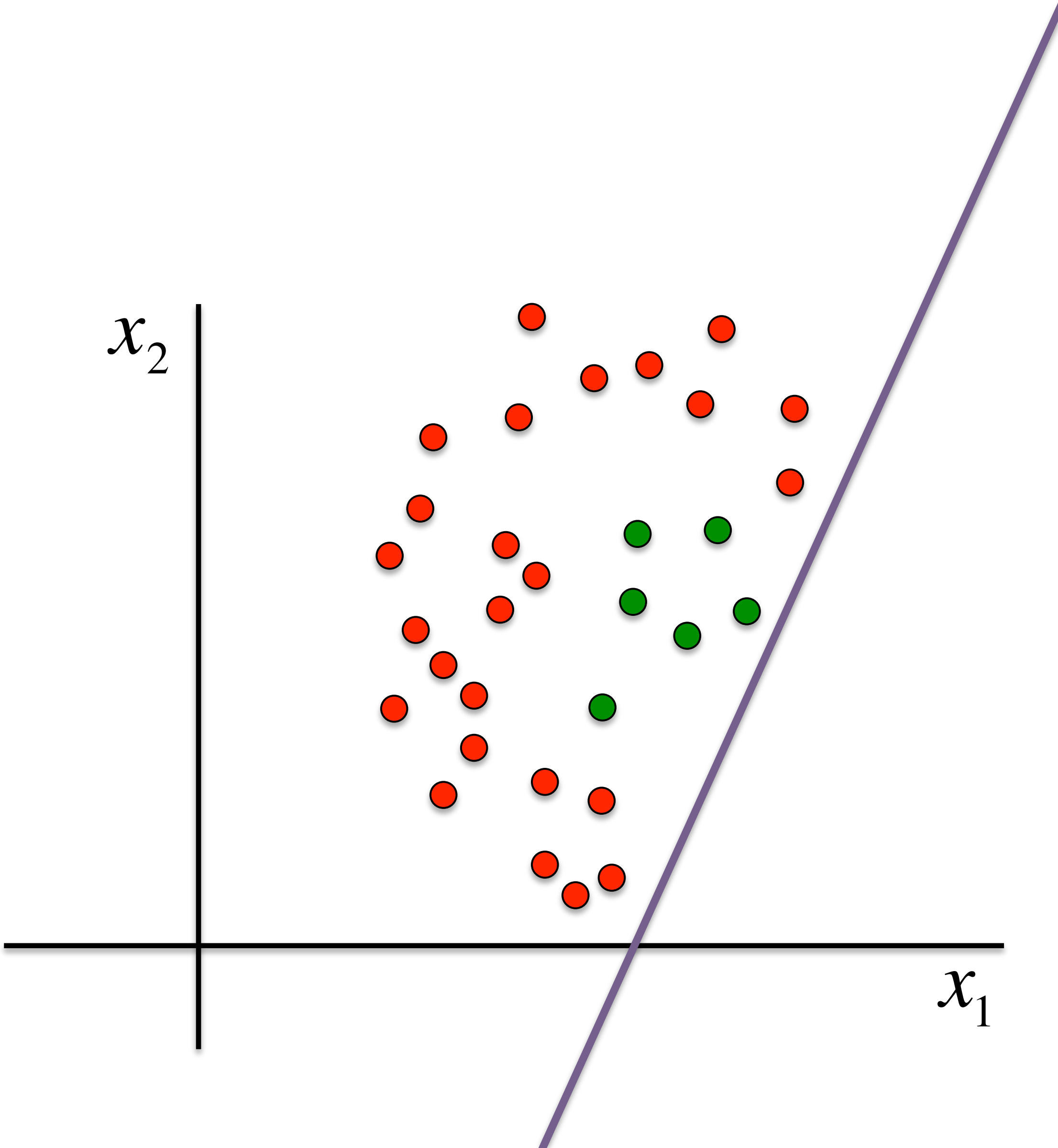
$$Acc = \frac{TP+TN}{P+N}$$

The number of samples correctly classified over the total number of samples

	p	n
$\tilde{P}$	**T**rue **P**ositive	**F**alse **P**ositive
$\tilde{N}$	**F**alse **N**egative	**T**rue **N**egative
	P	N

Accuracy Limitations

Accuracy = 80%



TRUE SKILL STATISTICS

True Skill Statistics was proposed as a standard metric to compare flare forecasts (Bloomfield '12) :

TSS=%True Positive - % False Positive

This formulation is equivalent to the **Balanced Classification Rate**

$$BCR=\frac{1}{2}\left(TP / (TP + FN) + TN / (TN + FP)\right)$$

Conclusions

- Public repository of data
- No well defined set of descriptors
  - Features selection
- Imbalanced problem
  - Tecniques for imbalanced datasets
- Classifier: the last step!
- Metrics: balanced

Conclusions

- Public repository of data
- No well defined set of descriptors
  - Features selection
- Imbalanced problem
  - Techniques for imbalanced datasets
- Classifier: the last step!
- Metrics: balanced

The end

Thanks for your attention