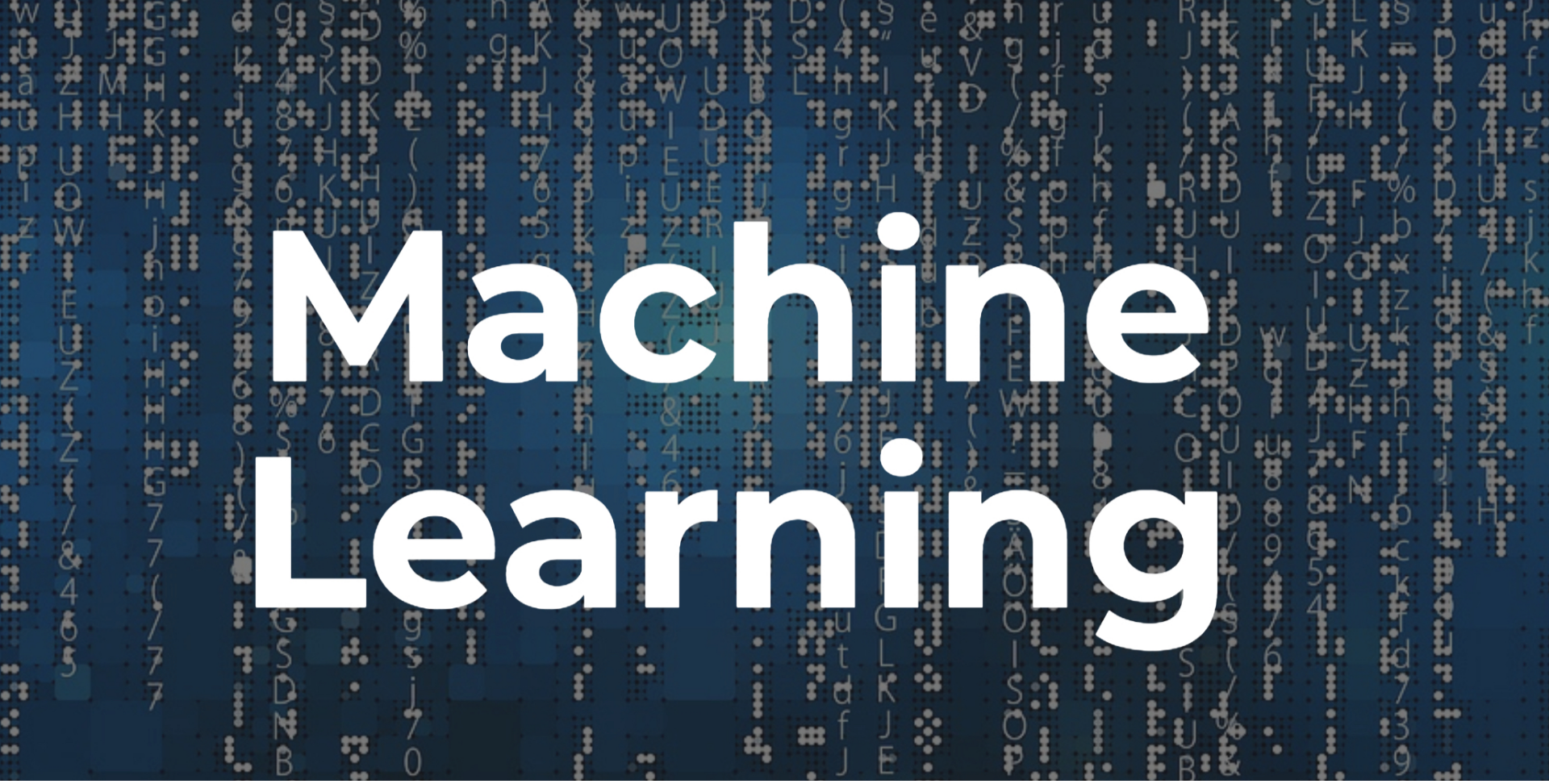


CSCI 4622 Fall 2019

Prof. Claire Monteleoni

Today: Lecture 11

- Ensemble Methods
  - Boosting

- Loss functions

With slide credits: S. Dasgupta, T. Jaakkola, J. Shotton, and A. Torralba

Ensemble methods

An ensemble classifier combines a set of weak "base" classifiers into a "strong" ensemble classifier.

- "boosted" performance
- more robust against overfitting

- Decision Forests, Random Forests [Breiman '01], Bagging
- Voted-Perceptron
- Boosting
- Online learning with expert advice
- ….

Boosting

- A simple algorithm for learning robust ensemble classifiers
  - Freund & Shapire, 1995
  - Friedman, Hastie, Tibshhirani, 1998

- Easy to implement, no external optimization tools needed.

Boosting

- Defines a classifier using an additive model:

$$F(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \alpha_3 f_3(x) + \ldots$$

Strong classifier

Data point

Weight

Weak classifier

Boosting

- Defines a classifier using an additive model:

$$F(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \alpha_3 f_3(x) + \ldots$$

Strong classifier

Data point

Weight

Weak classifier

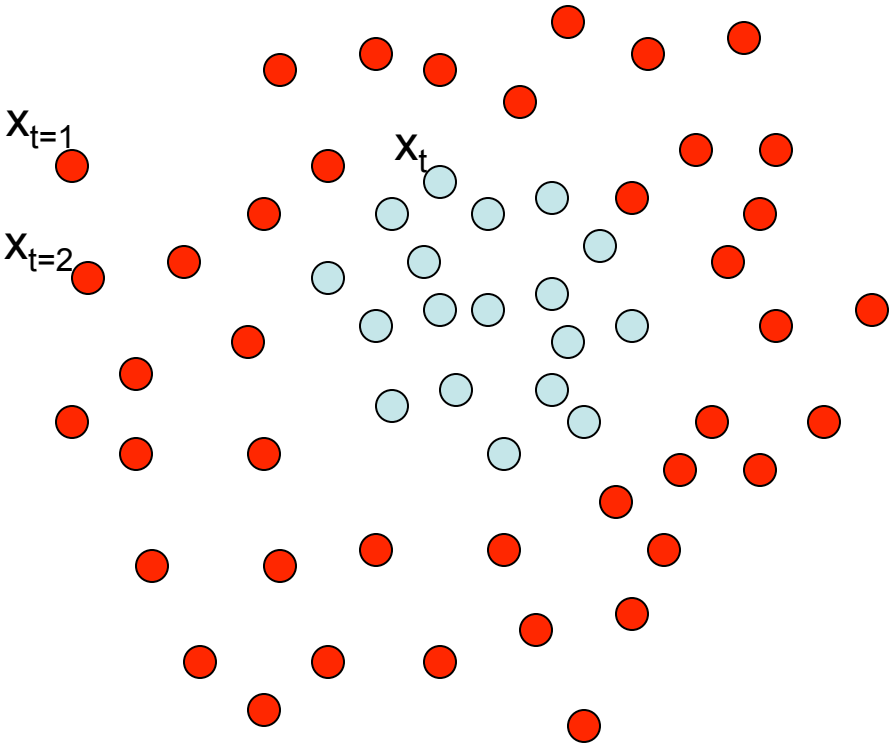
- We need to define a family of weak classifiers

$$f_k(x)$$

  - E.g. linear classifiers, decision trees, or even decision stumps (threshold on one axis-parallel dimension)

Boosting

- Run sequentially on a batch of n data points



Each data point has a class label:

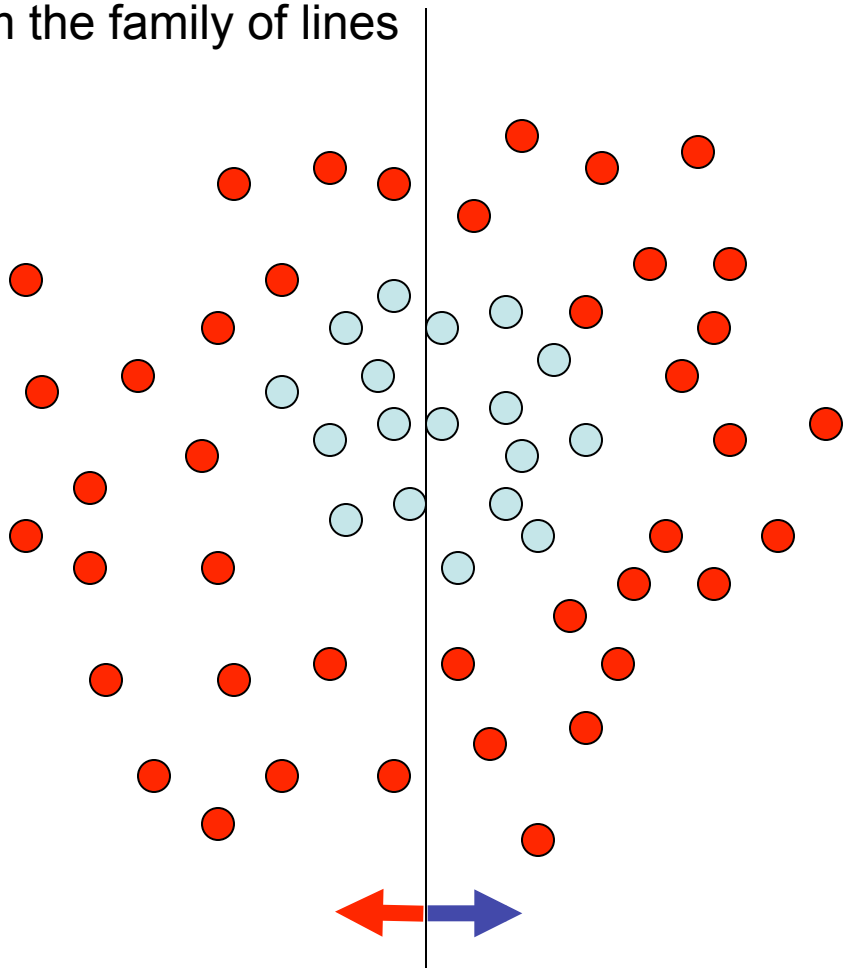
$$y_t = \begin{cases} +1 \ (\bullet) \\ -1 \ (\circ) \end{cases}$$

and a weight, w_t.

- we initialize all $w_t = 1$

Example using linear separators

Weak learners from the family of lines



Each data point has a class label:

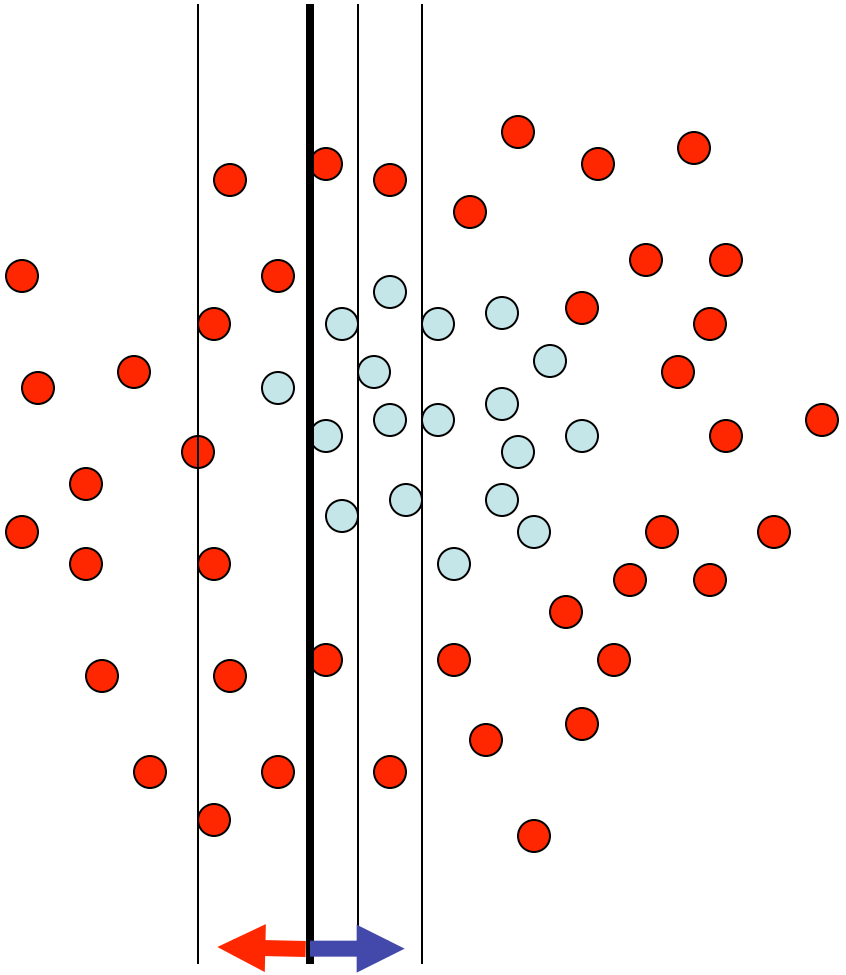
$$y_t = \begin{cases} +1 & (\bullet) \\ -1 & (\circ) \end{cases}$$

and a weight:

$$w_t = 1$$

This linear separator has error rate 50%

Example using linear separators



Each data point has a class label:

$$y_t = \begin{cases} +1 & (\bullet) \\ -1 & (\circ) \end{cases}$$

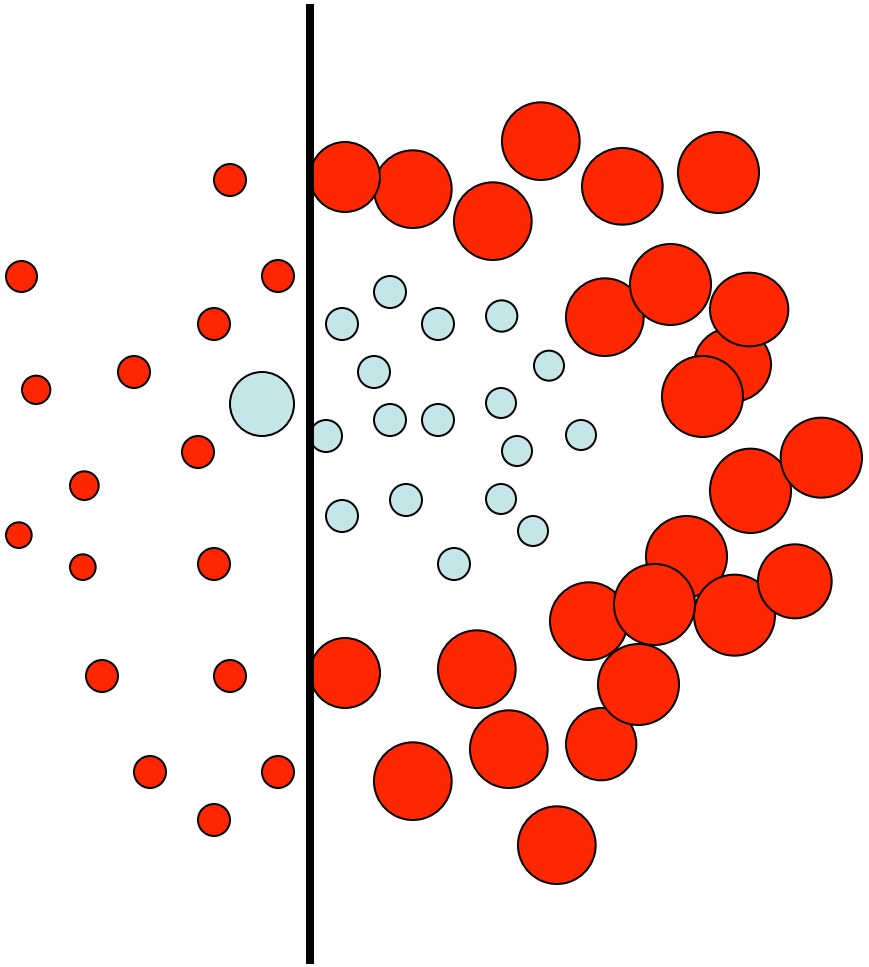
and a weight:

$$w_t = 1$$

This one seems to be the best, call it f_1

This is a '**weak classifier**' : Its error rate is slightly less than 50%.

Example using linear separators



Each data point x_i has a class label:

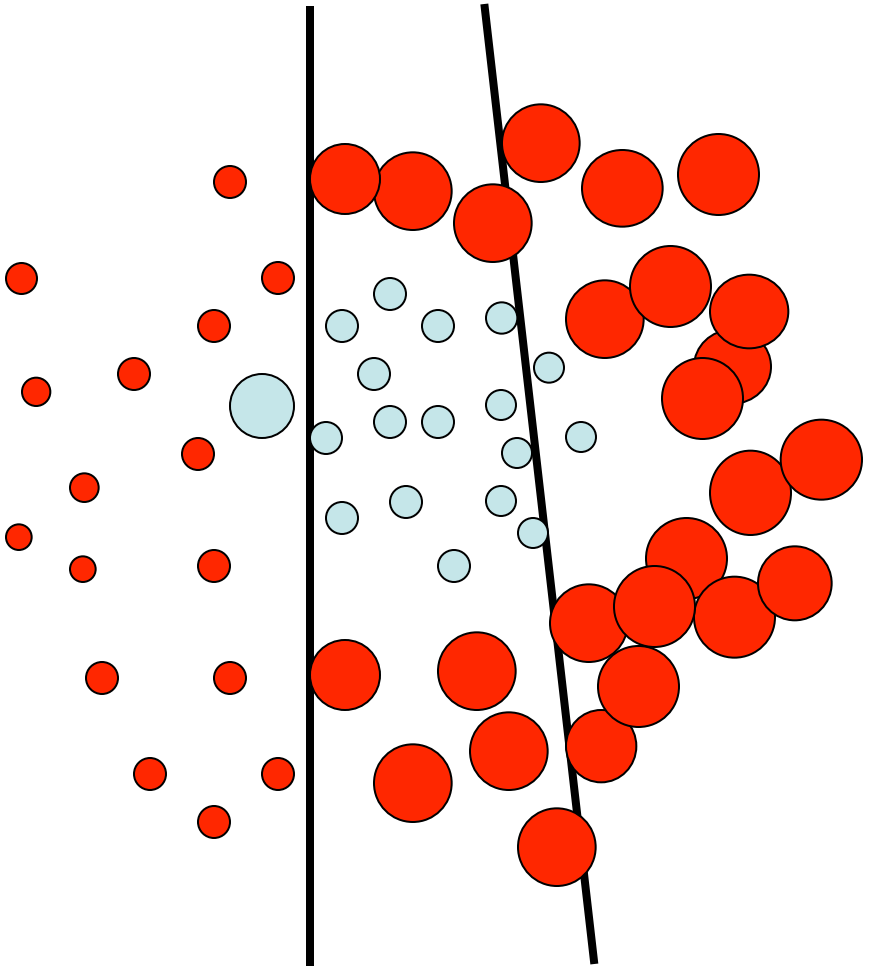
$$y_i = \begin{cases} +1 \\ -1 \end{cases}$$

We update the weights:

$$w_t \leftarrow w_t \exp\{-y_t f_m(x_t)\}$$

- Re-weight the points such that the previous weak classifier now has 50% error
- Iterate: find a weak classifier for this new problem

Example using linear separators



Each data point x_i has a class label:

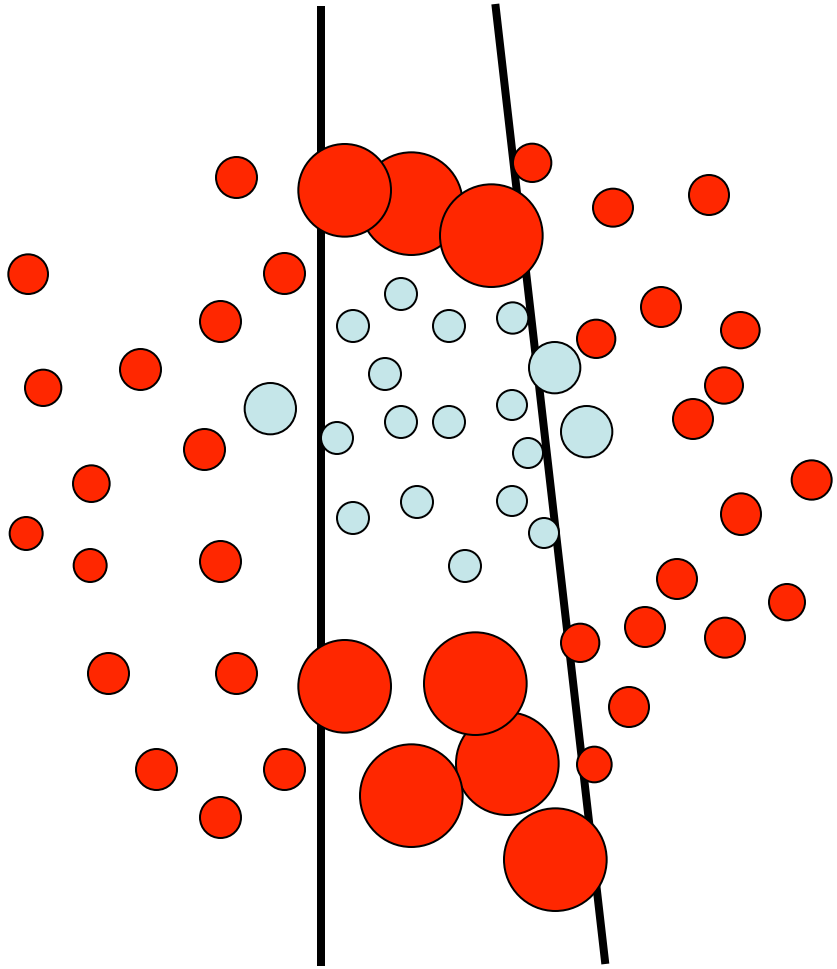
$$y_i = \begin{cases} +1 \\ -1 \end{cases}$$

We update the weights:

$$w_t \leftarrow w_t \exp\{-y_t f_m(x_t)\}$$

- Re-weight the points such that the previous weak classifier now has 50% error
- Iterate: find a weak classifier for this new problem

Example using linear separators



Each data point x_i has a class label:

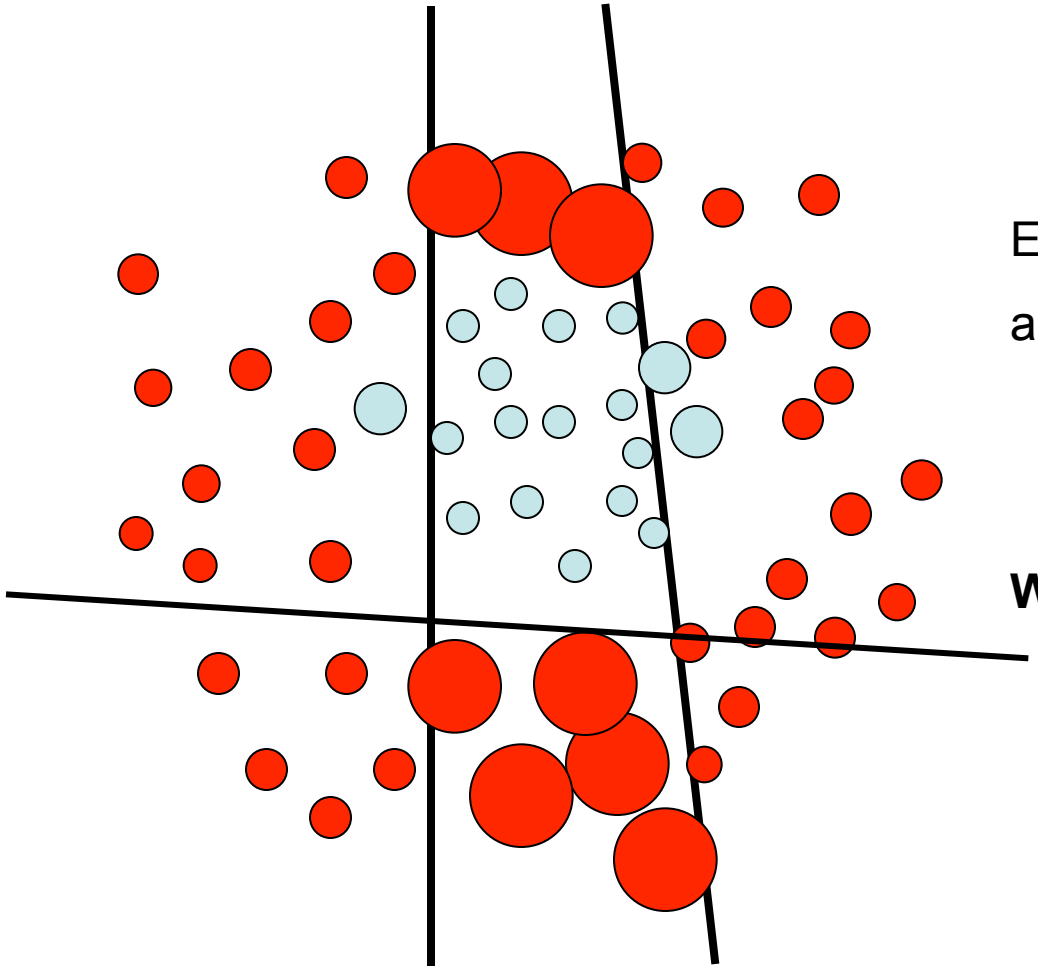
$$y_i = \begin{cases} +1 \\ -1 \end{cases}$$

We update the weights:

$$w_t \leftarrow w_t \exp\{-y_t f_m(x_t)\}$$

- Re-weight the points such that the previous weak classifier now has 50% error
- Iterate: find a weak classifier for this new problem

Example using linear separators



Each data point x_i has a class label:

$$y_i = \begin{cases} +1 \\ -1 \end{cases}$$

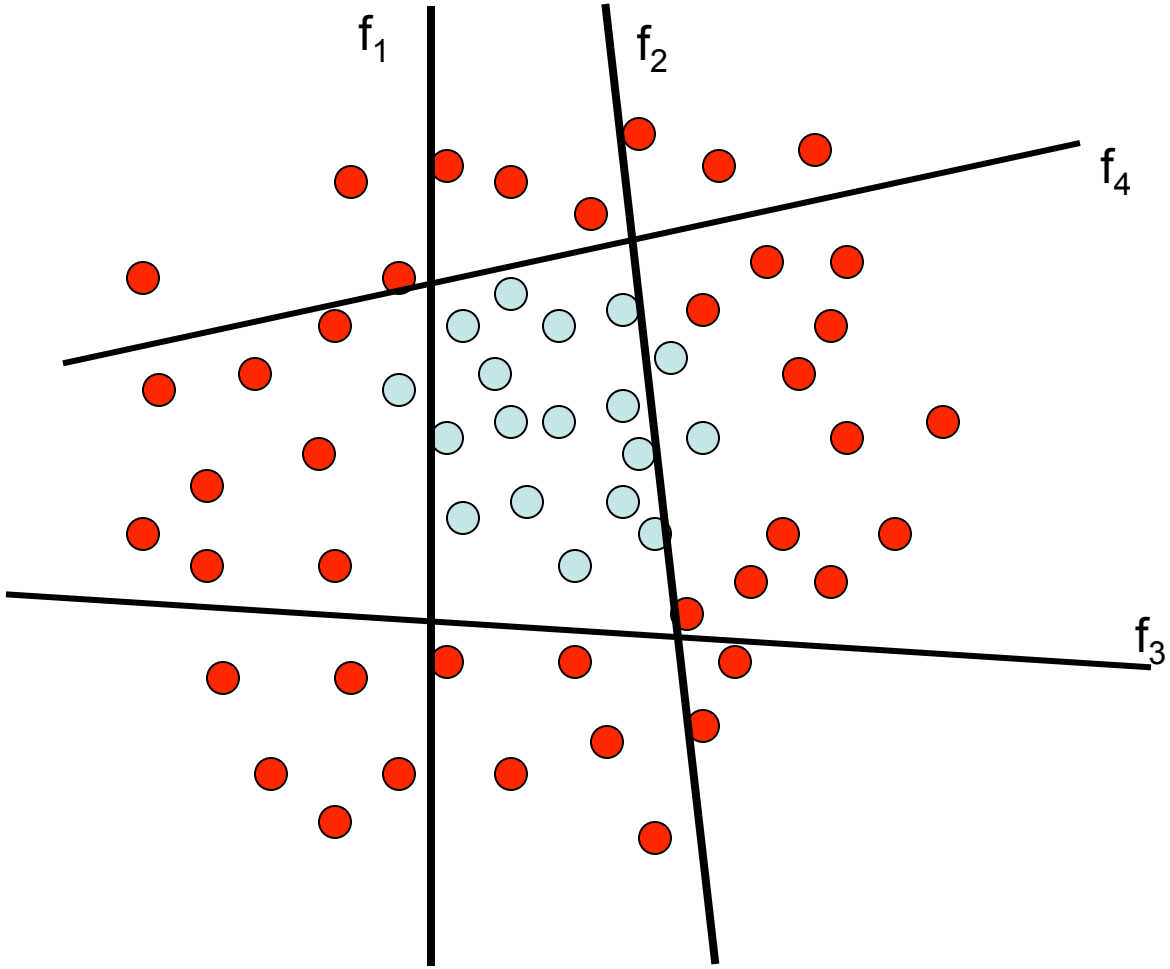
We update the weights:

$$w_t \leftarrow w_t \exp\{-y_t f_m(x_t)\}$$

- Re-weight the points such that the previous weak classifier now has 50% error
- Iterate: find a weak classifier for this new problem

Torralba, ICCV 05 Short Course

Example using linear separators



The strong (non-linear) ensemble classifier is built as a weighted combination of all the weak (linear) classifiers.

Torralba, ICCV 05 Short Course

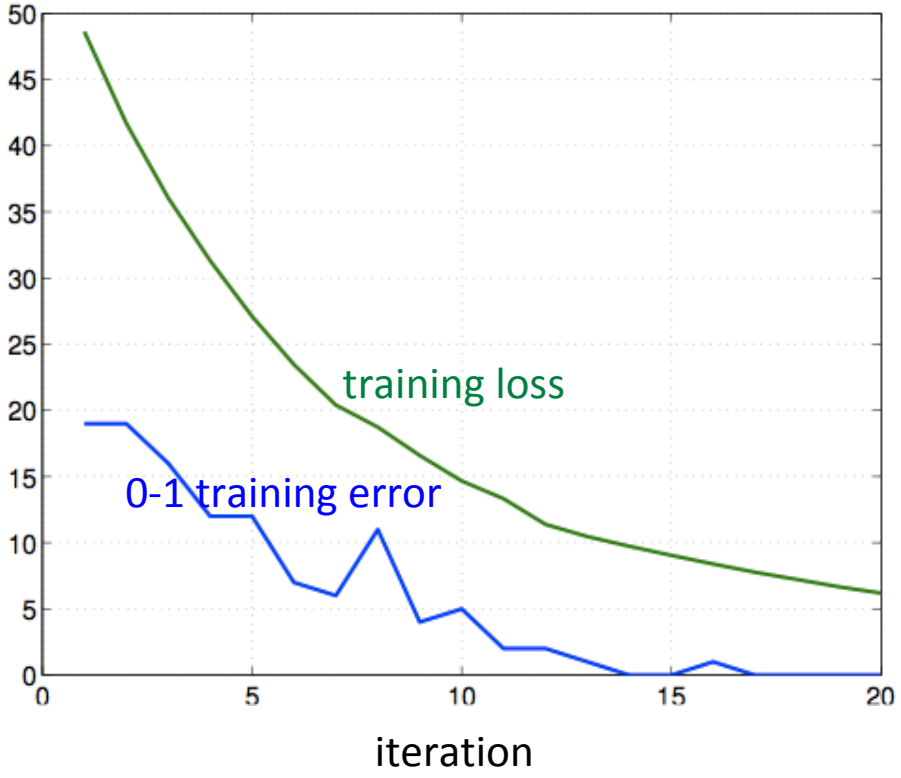
Boosting

- AdaBoost (Freund and Shapire, 1995)
- Real AdaBoost (Friedman et al, 1998)
- LogitBoost (Friedman et al, 1998)
- Gentle AdaBoost (Friedman et al, 1998)
- BrownBoosting (Freund, 2000)
- FloatBoost (Li et al, 2002)
- …

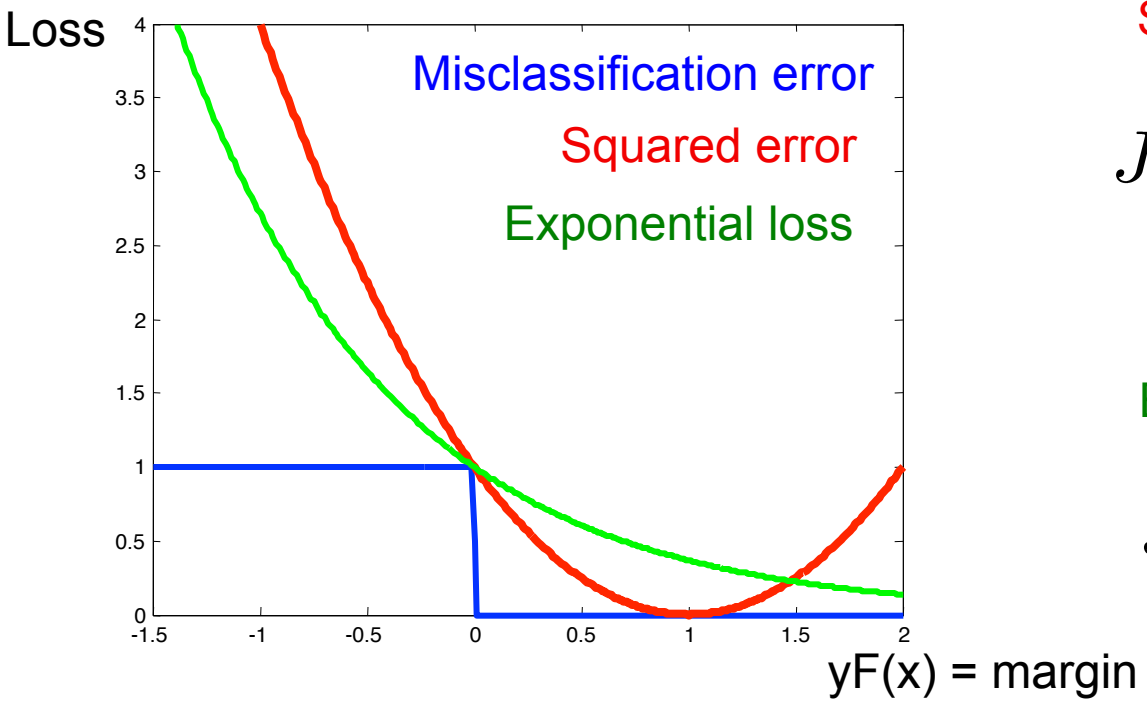
Mostly differ in choice of loss function and how it is minimized.

Loss functions: motivation

- We want a smooth upper bound on 0-1 training error.



Loss functions



Squared error

$$J = \sum_{t=1}^{N} [y_t - F(x_t)]^2$$

Exponential loss

$$J = \sum_{t=1}^{N} e^{-y_t F(x_t)}$$

Boosting

Sequential procedure. At each step we add

$$F(x) \leftarrow F(x) + f_m(x)$$

to minimize the residual loss

$$(\phi_m) = \arg\min_{\phi} \sum_{t=1}^{N} J\left(y_i, F(x_t) + f(x_t; \phi)\right)$$

Parameters of weak classifier

Desired output

input

where J is the loss function

For more details: Friedman, Hastie, Tibshirani. "Additive Logistic Regression: a Statistical View of Boosting" (1998)

How to set the ensemble weights?

- Prediction on a new data point *x* is typically of the form:

$$F(x) = \sum_{m=1}^{k} \alpha_m f_m(x)$$

- How to set the α_m values?
- Depends on the algorithm (due to different loss functions, etc.)
  E.g. in AdaBoost:

$$\alpha_m = \frac{1}{2} \ln \frac{1 - \epsilon_m}{\epsilon_m}$$

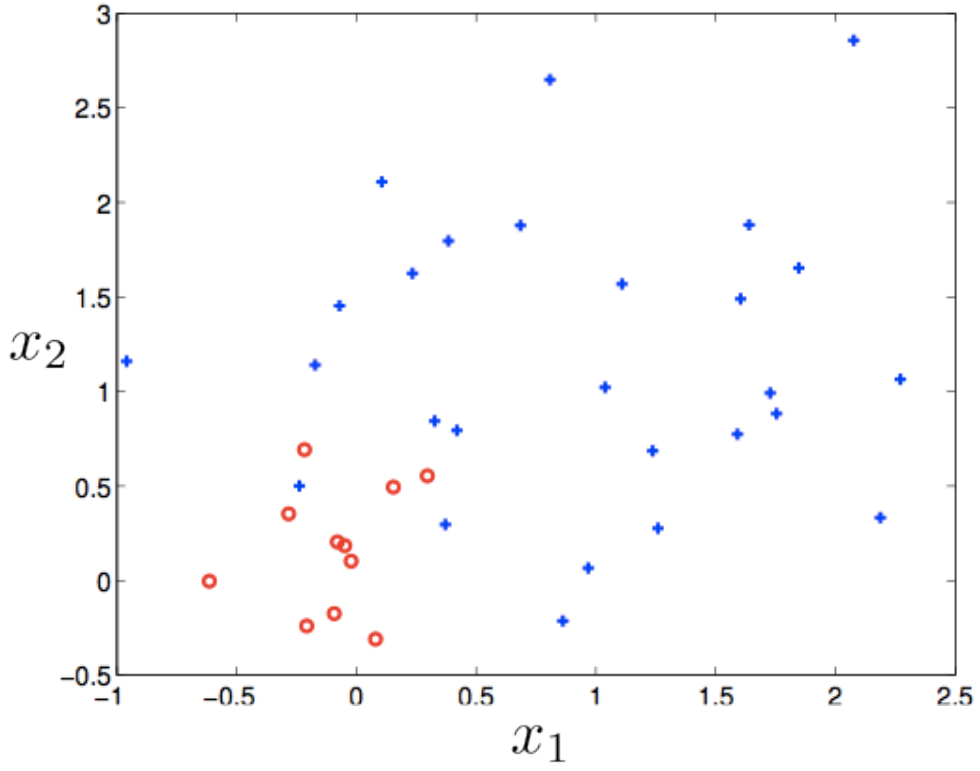
- Where ϵ_m is the training error of f_m on the (currently) weighted data set.

Understanding boosting

- There are four different kinds of “error” in boosting:
  - weighted error that the base learner achieves at each iteration
  - weighted error of the base learner relative to just updated weights (i.e., trying the same base learner again)
  - training error of the ensemble as a function of the number of boosting iterations
  - generalization error of the ensemble

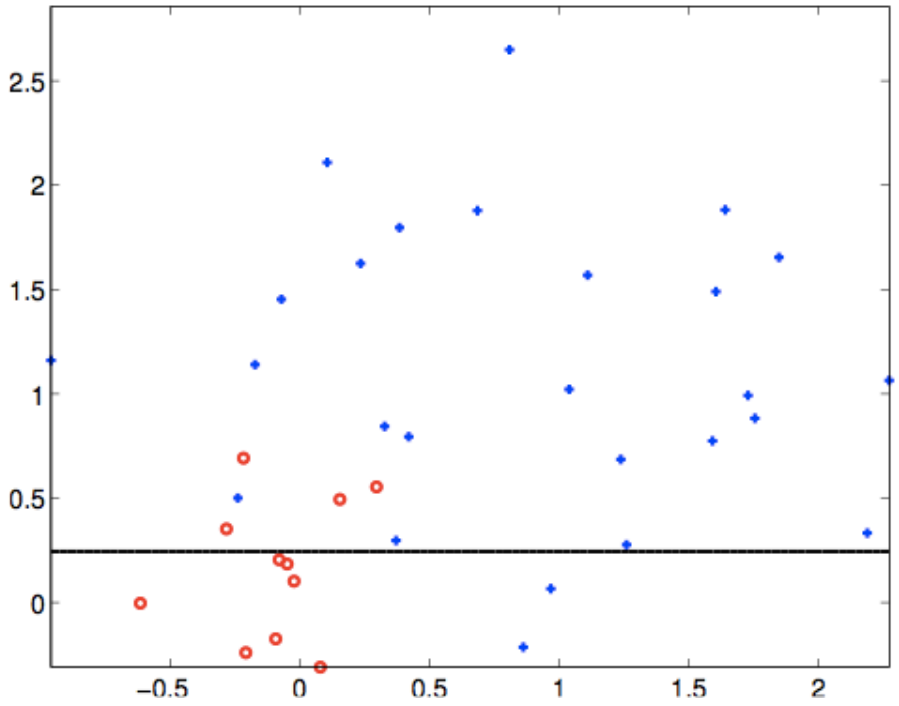
Boosting example

Logistic loss $\text{Loss}(z) = \log(1 + \exp(-z))$

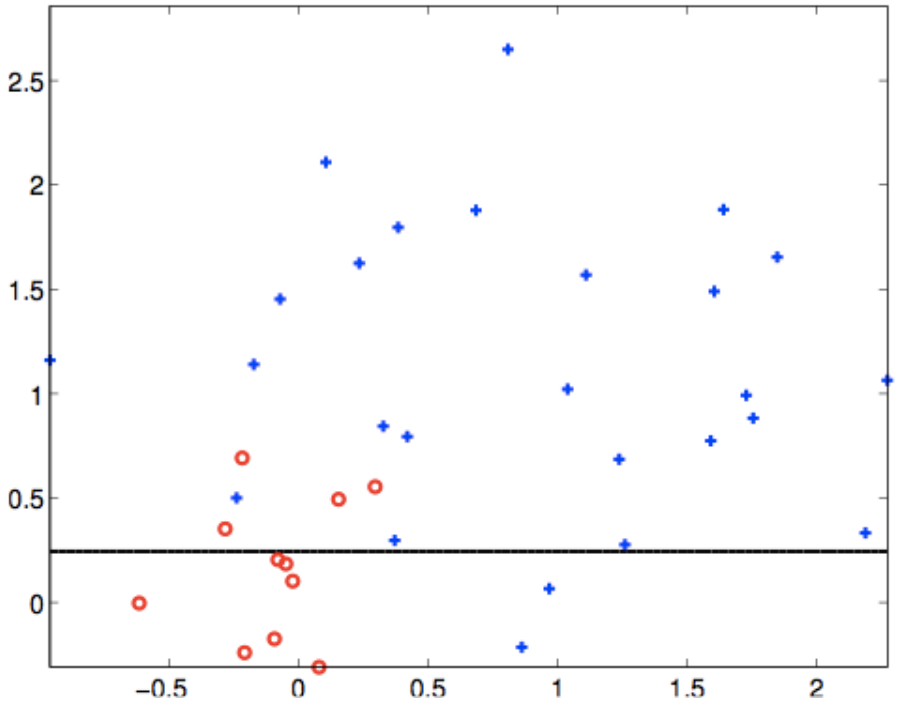


Boosting example

$h(\underline{x}; \hat{\underline{\theta}}_1)$

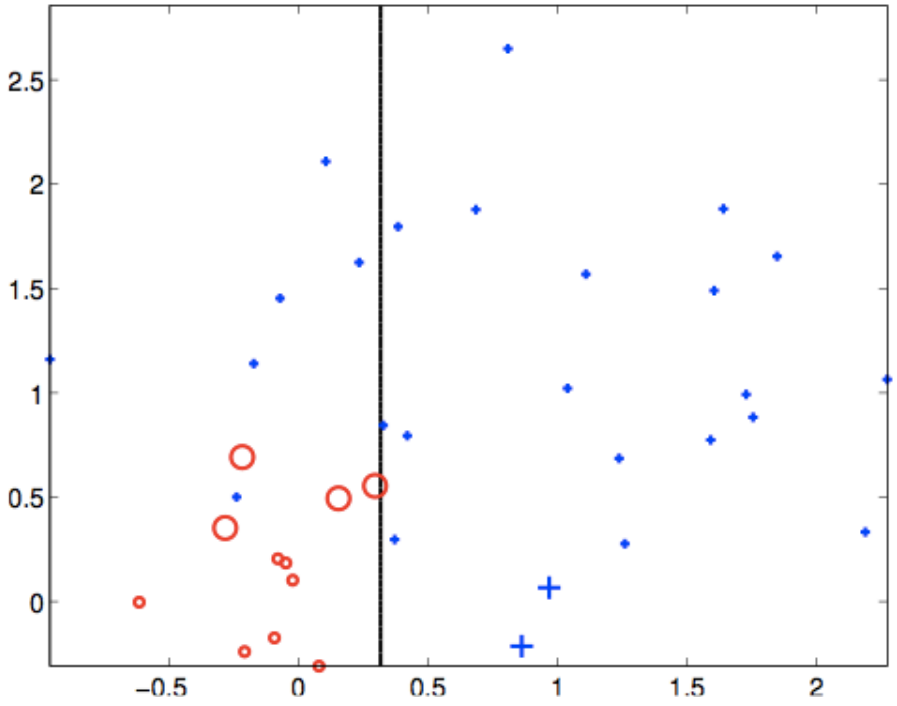


$h_1(\underline{x}) = \hat{\alpha}_1 h(\underline{x}; \hat{\underline{\theta}}_1)$

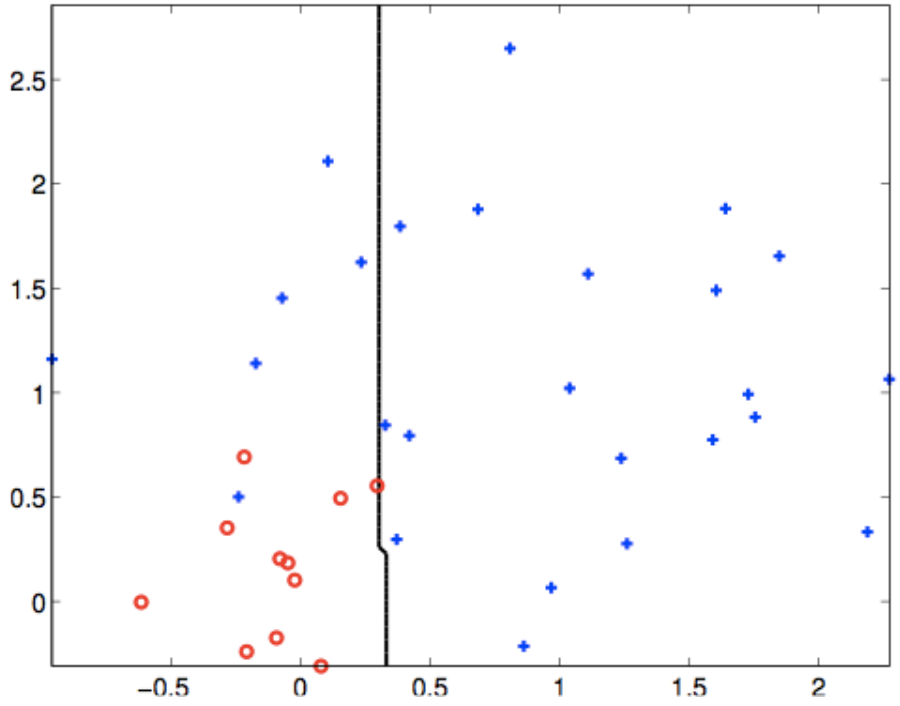


Boosting example

$h(\underline{x}; \hat{\underline{\theta}}_2)$

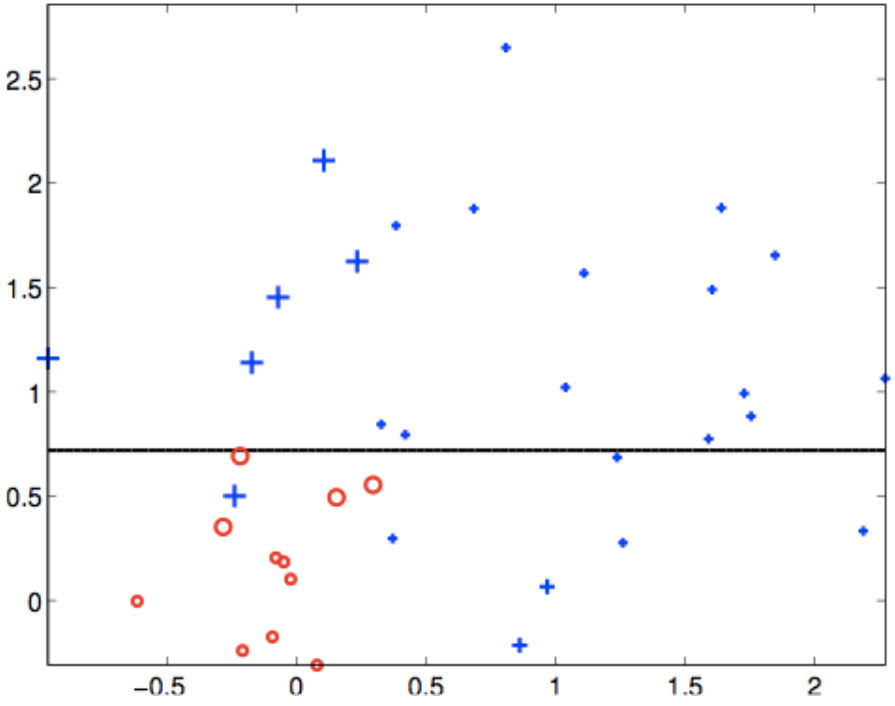


$h_2(\underline{x}) = \hat{\alpha}_1 h(\underline{x}; \hat{\underline{\theta}}_1) + \hat{\alpha}_2 h(\underline{x}; \hat{\underline{\theta}}_2)$

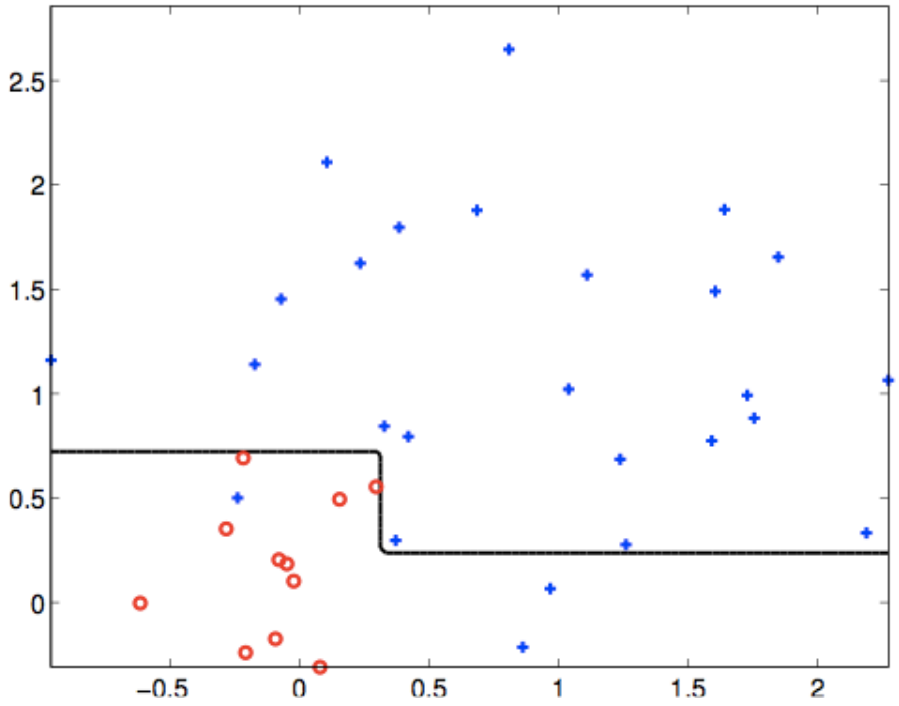


Boosting example

$h(\underline{x}; \hat{\underline{\theta}}_3)$

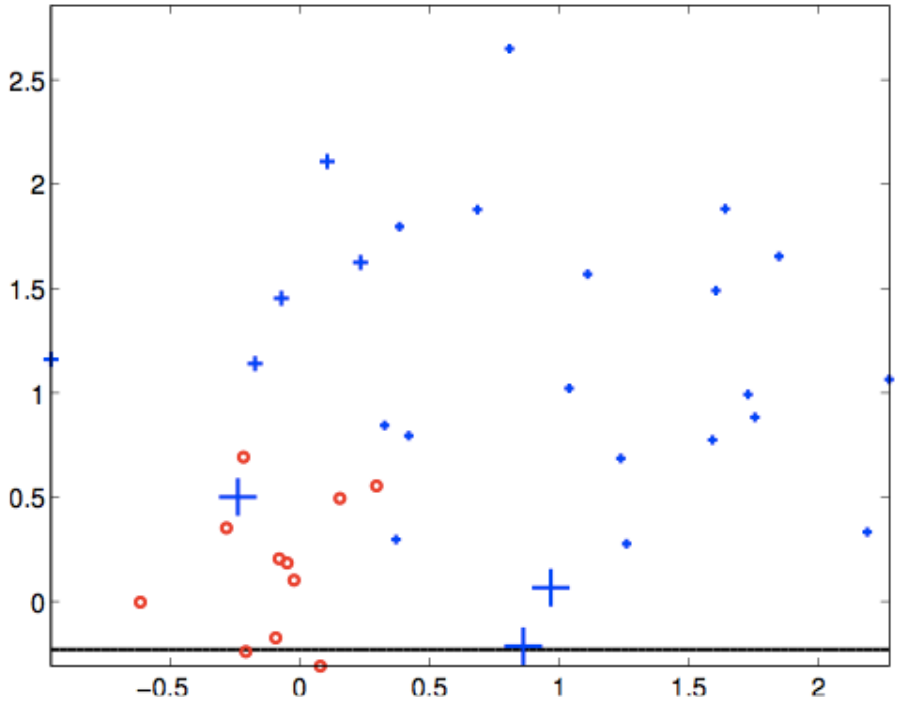


$h_3(\underline{x}) = \hat{\alpha}_1 h(\underline{x}; \hat{\underline{\theta}}_1) + \cdots + \hat{\alpha}_3 h(\underline{x}; \hat{\underline{\theta}}_3)$



Boosting example

$h(\underline{x}; \hat{\underline{\theta}}_4)$



$h_4(\underline{x}) = \hat{\alpha}_1 h(\underline{x}; \hat{\underline{\theta}}_1) + \cdots + \hat{\alpha}_4 h(\underline{x}; \hat{\underline{\theta}}_4)$