The Tyranny of the Average

John Langford

Sources

Tom Dietterich's machine learning summary (first 10 pages):

http://citeseer.nj.nec.com/dietterich97machine.html

Robert Schapire's boosting summary:

http://www.cs.princeton.edu/~schapire/uncompress-papers.cgi/msri.ps

Averaging

$$h(x) = \mathsf{sign}\left(\sum_{k=1}^{K} \alpha_k h_k(x)\right)$$

Examples of Averaging Classifiers

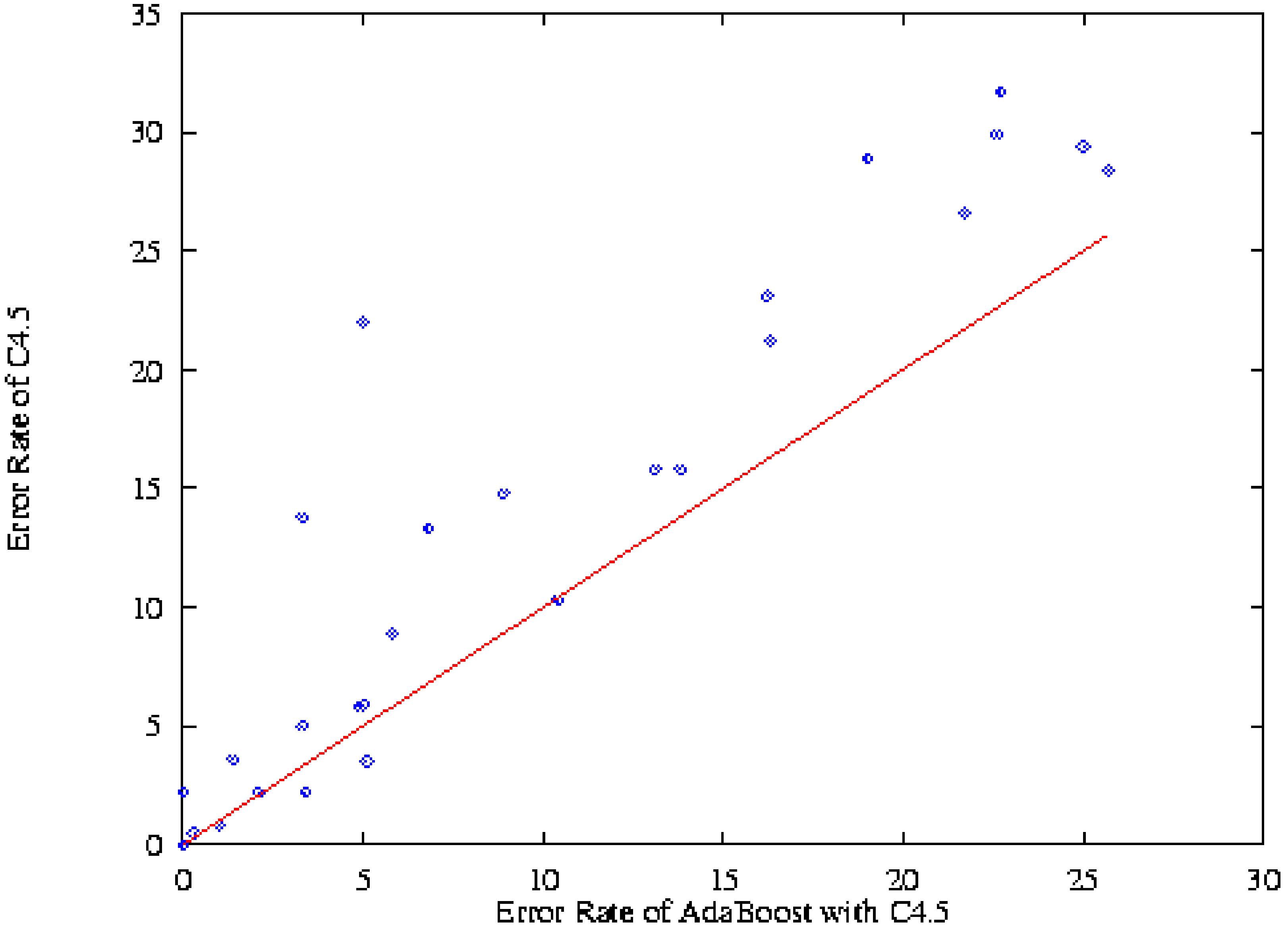
1. Adaboost (Freund and Schapire 1996)

2. Bagging (Breiman, 1996)

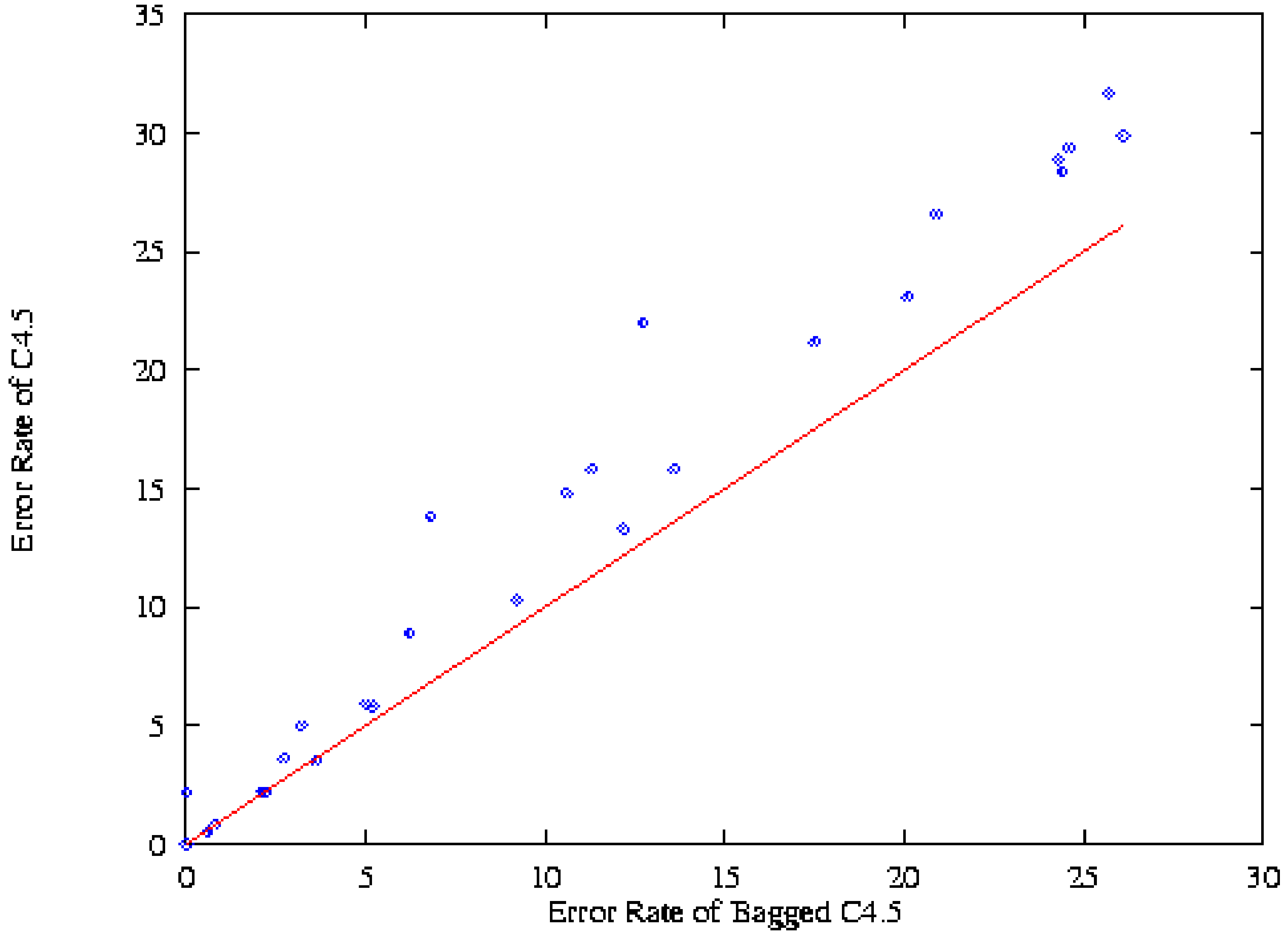
3. Cross-validated Committees (Permanto, Munro, and Doyle, 1996)

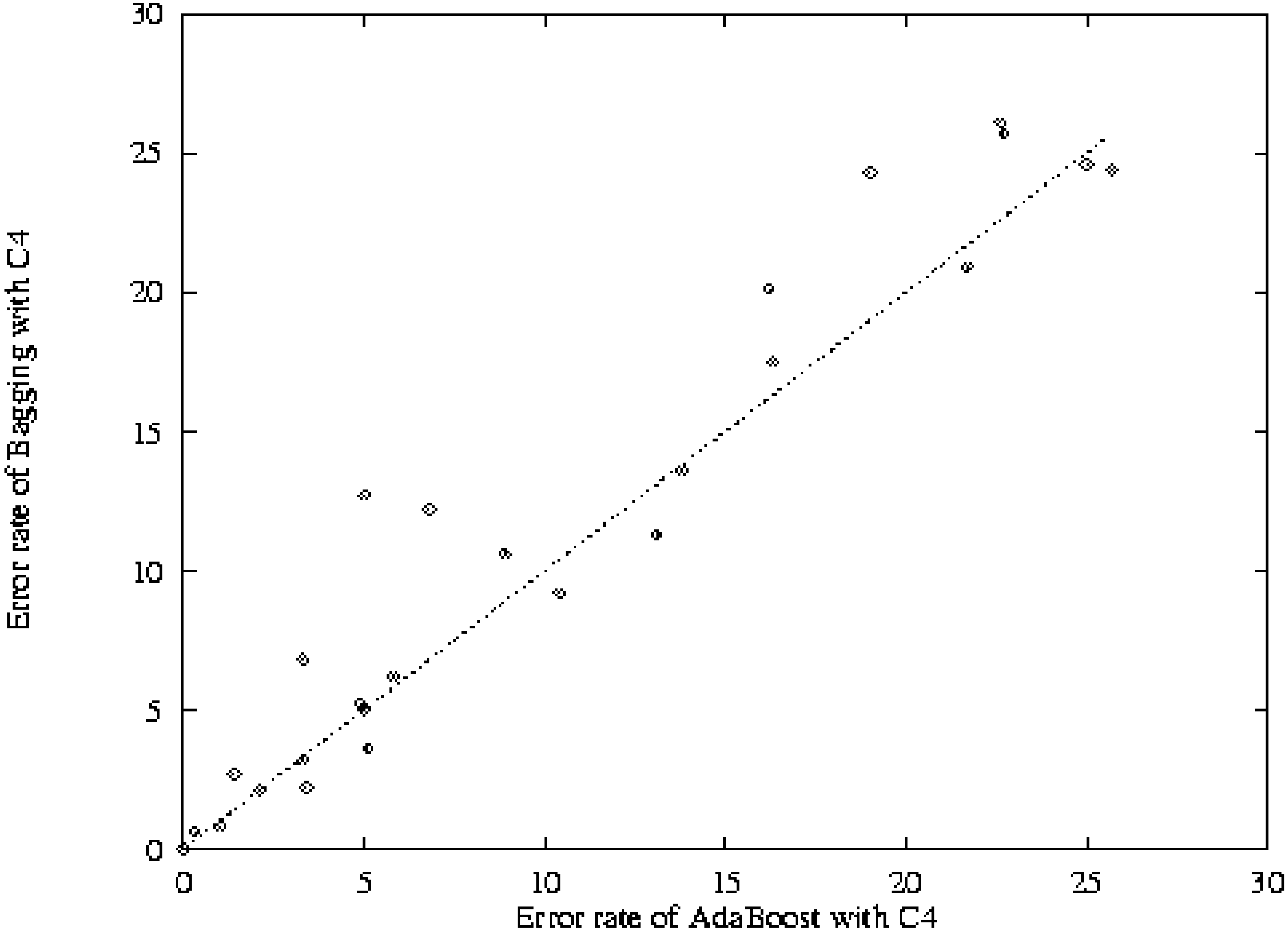
4. Bayes Optimal

5. Maximum Entropy (Jaakola, Meila, Jebara 1999)



Error Rate of C4.5
Error Rate of AdaBoost with C4.5
0
5
10
15
20
25
30
35


Error Rate of C4.5
Error Rate of Bagged C4.5
0
5
10
15
20
25
30
35



Error rate of Bagging with C4
Error rate of AdaBoost with C4
0
5
10
15
20
25
30

Outline

1. General Theoretical Motivations

   (a) Independent Errors

   (b) Sample Complexity Theory

   (c) What would Bayes do?

2. A Zoology of Averages

Independent Errors

Suppose each h_i errs independently:

$$\mathsf{Pr}(h_1(x) \neq y \wedge h_2(x) \neq y ... | y)$$

$$= \prod_k \mathsf{Pr}\left(h_k(x) \neq y | y\right)$$

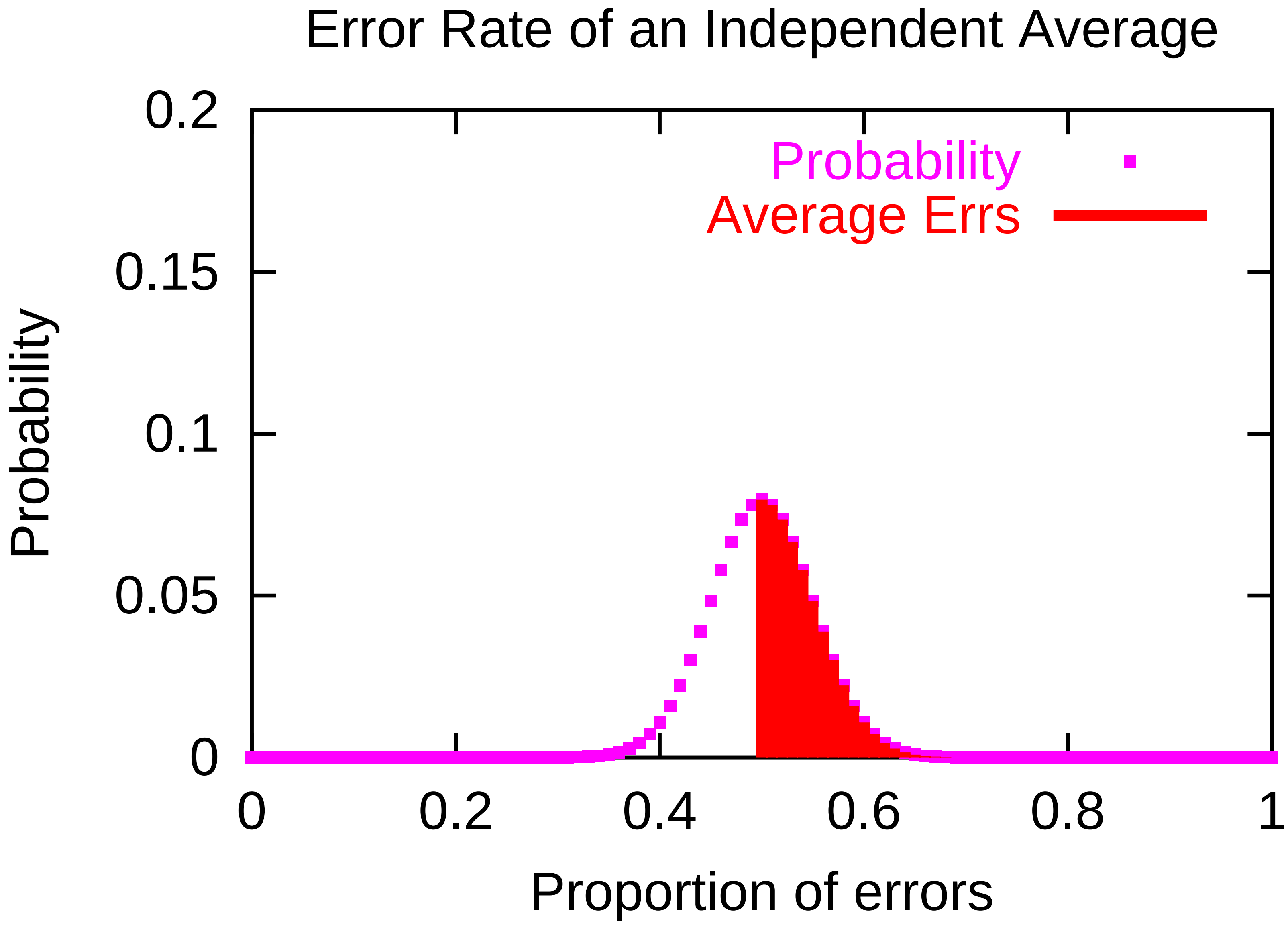
What is the probability that the average misclassifies?

Independent Errors II

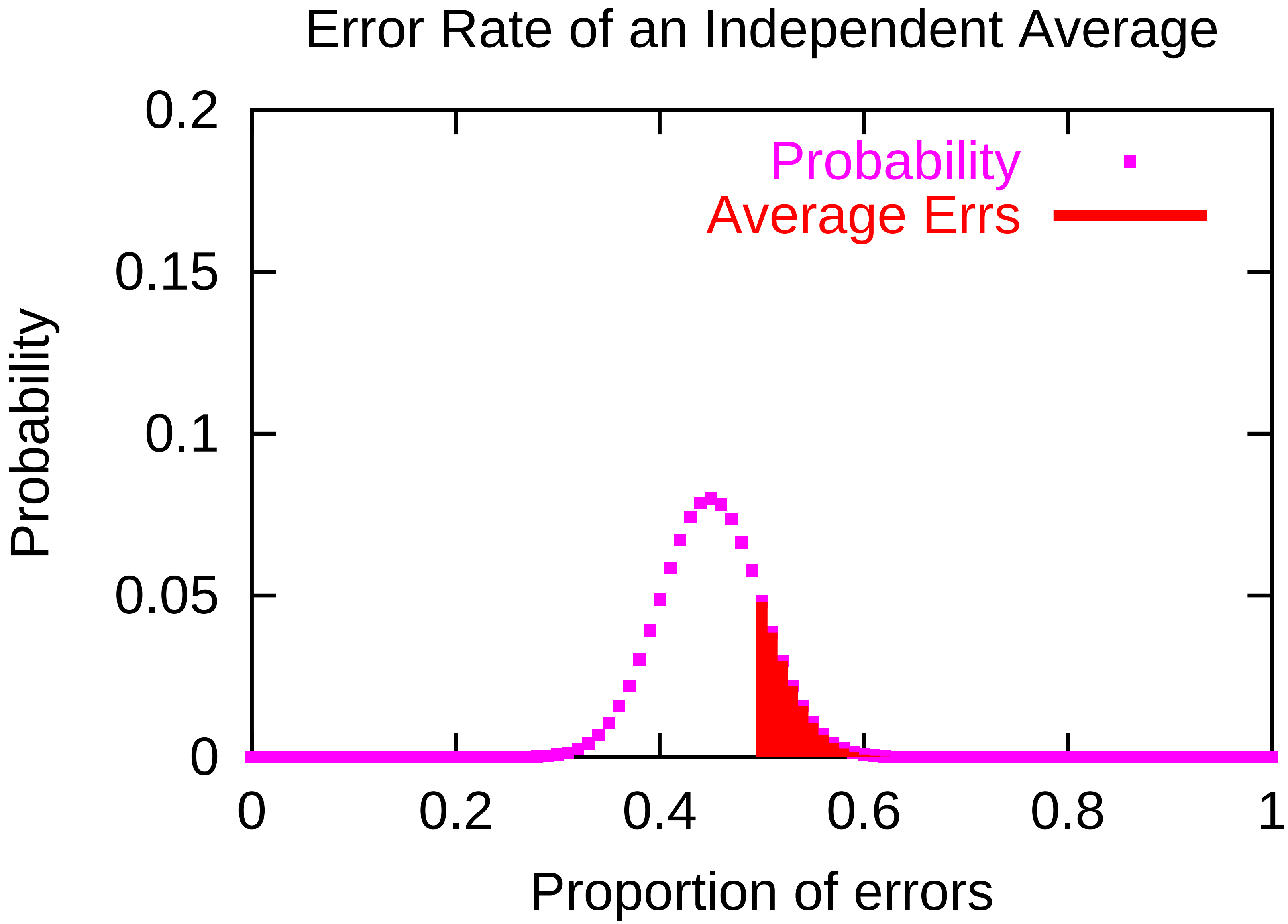
Suppose $\Pr(h_k(x) \neq y | y) = \mu$.

Then,

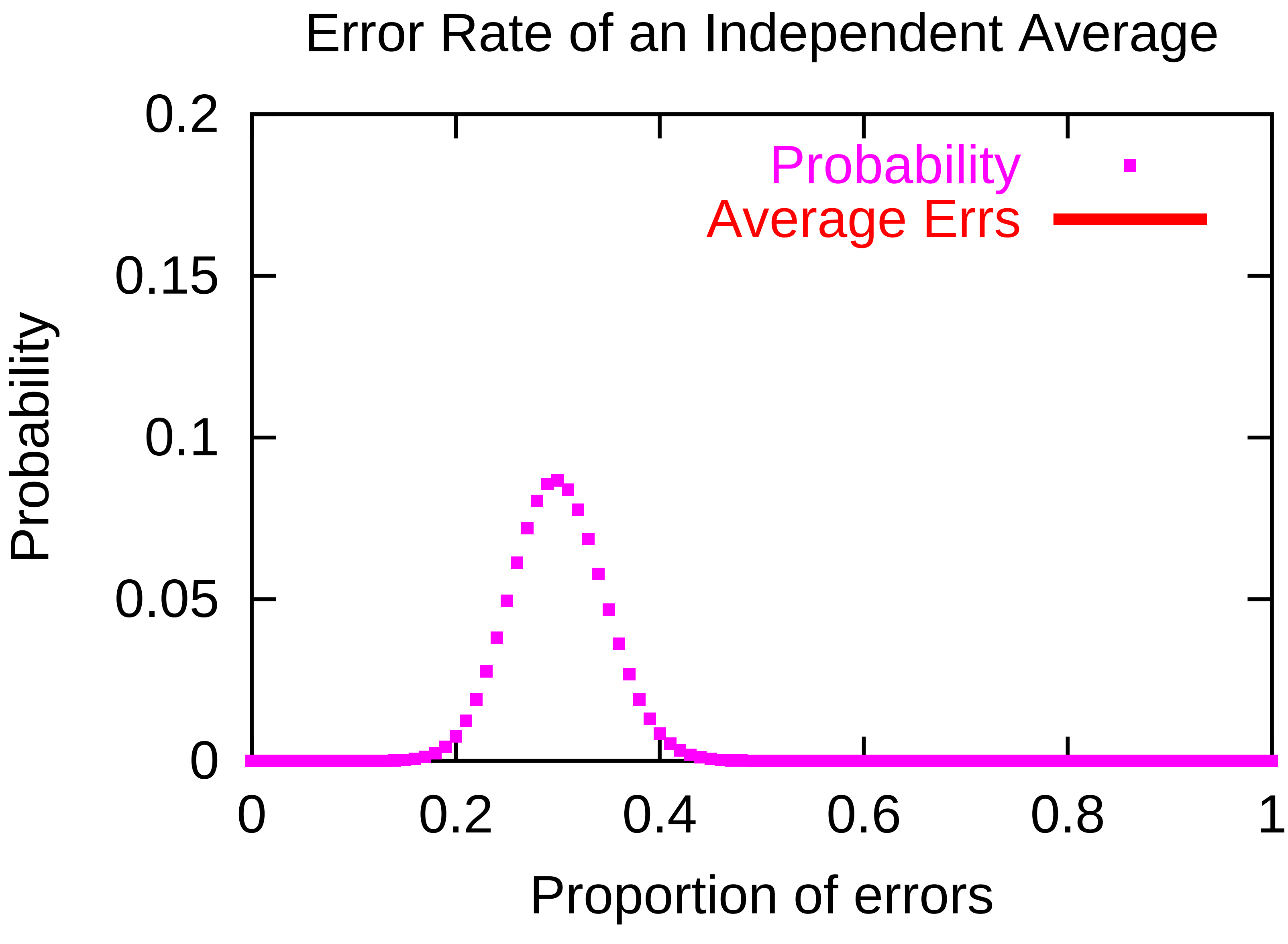
$$\Pr\left(\sum_k I(h_k(x) \neq y) \geq \frac{K}{2} | y\right) = 1 - \mathrm{Bin}\left(\frac{\frac{K}{2}}{K}, \mu\right)$$



Error Rate of an Independent Average
Probability
Average Errs
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1
Probability
Proportion of errors



Error Rate of an Independent Average
Probability
Average Errs
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1
Probability
Proportion of errors


Error Rate of an Independent Average
Probability
Average Errs
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1
Probability
Proportion of errors

Sample Complexity

Occam's Razor bound does *not* motivate averages.

... but remember side note: there are many other train set bounds, many of which motivate averages.

1. Margin Bound (Schapire, Freund, Bartlett, Lee, 1998)
2. PAC-Bayes Bound (McAllester, 1999)
3. Stochastic Margin Bound (Langford and Shawe-Taylor, 2002)
4. (many others...)

What would Bayes do?

$P(h_k)$ = prior over h_k

$Q(h_k|S) = \frac{P(S|h_k)P(h_k)}{P(S)}$

Bayes optimal Prediction:

$$h(x) = \text{sign}\left(\sum_k Q(h_k|S)h_k(x)\right)$$

Outline

1. Theoretical Motivations
   (a) Independent Errors (all methods)
   (b) Sample Complexity Theory (all methods)
   (c) What would Bayes do? (some methods)
2. A Zoology of Averages

Outline

1. Theoretical Motivations

2. A Zoology of Averages

   (a) Bagging

   (b) Boosting

Given: m training examples

1. Repeat $k = 1...K$ times

   (a) $S' = \emptyset$

   (b) Repeat m times:

      i. $(x, y) =$ an example from the uniform distribution on m training examples

      ii. $S' \leftarrow S' \cup \{(x, y)\}$

   (c) $h_k =$learning algorithm on S'

2. Return $h(x) = \text{sign}\left(\sum_k \frac{1}{K} h_k(x)\right)$

Bagging: Analysis

Question: How many unique examples are in S'?

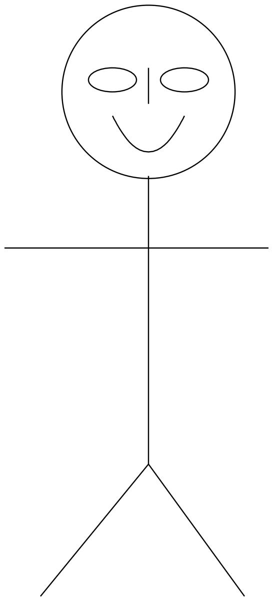
Answer: $1 - \frac{1}{e}$

Bagging: Analysis

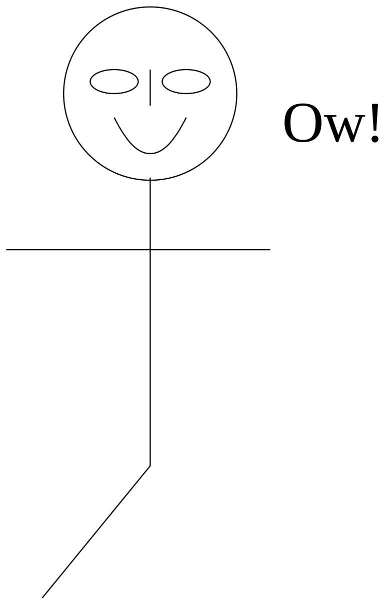
Question: What is the effect of duplicates?

Answer: They can weaken complexity control

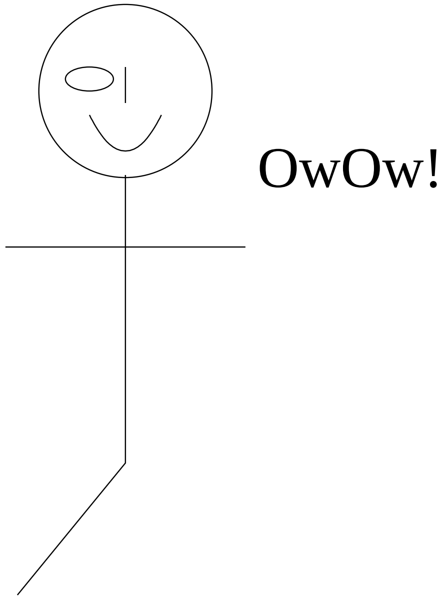
A Learning Algorithm



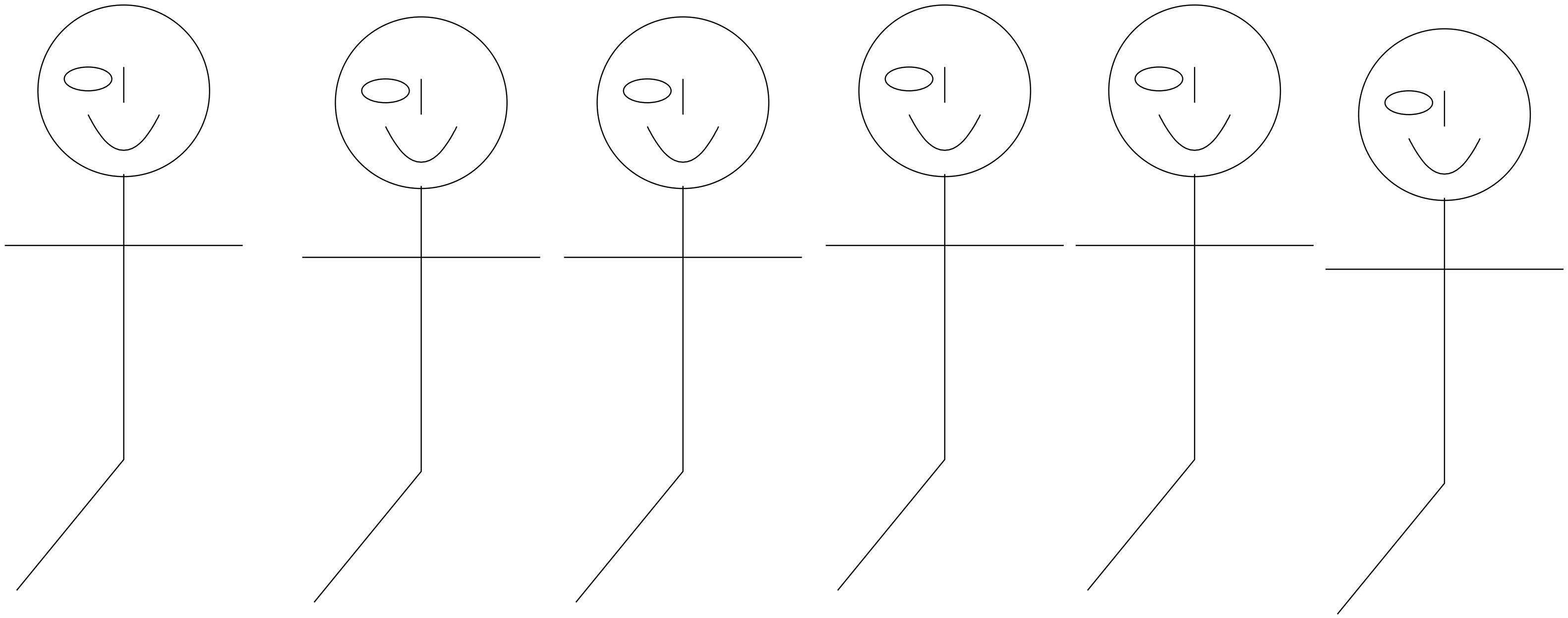
A Learning Algorithm missing $\frac{1}{e}$ of all examples



A Learning Algorithm missing $\frac{1}{e}$ examples and with duplicates



Bagging: Learning algorithm loses $\frac{1}{e}$ examples, gains duplicates, and is averaged



Boosting

Given:

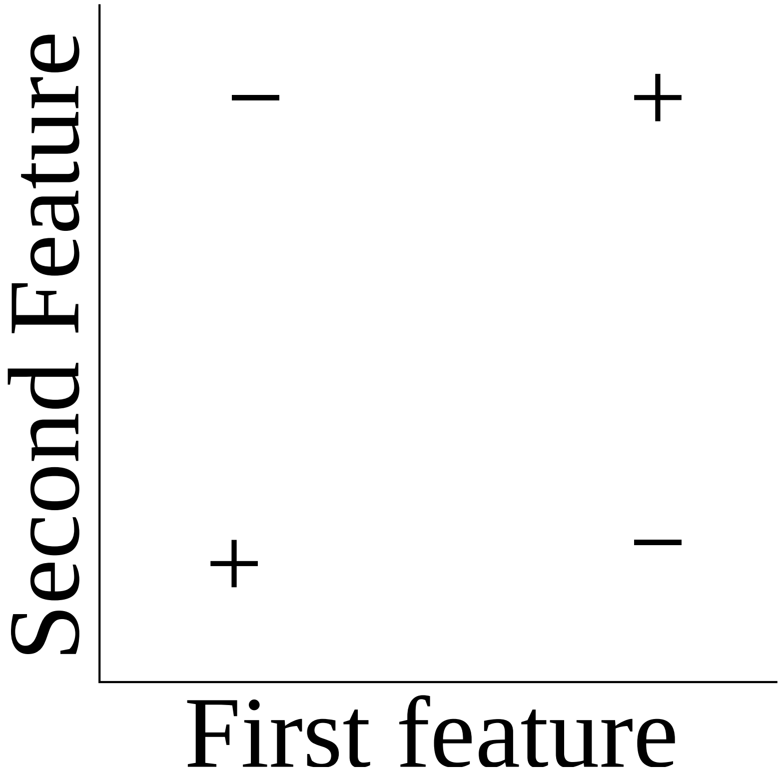
1. m labeled examples, $(x_1, y_1), ..., (x_m, y_m)$
2. A "weak" Classifier learning algorithm which takes a distribution $D(i)$ over the inputs

What is a “weak” Classifier learning algorithm?

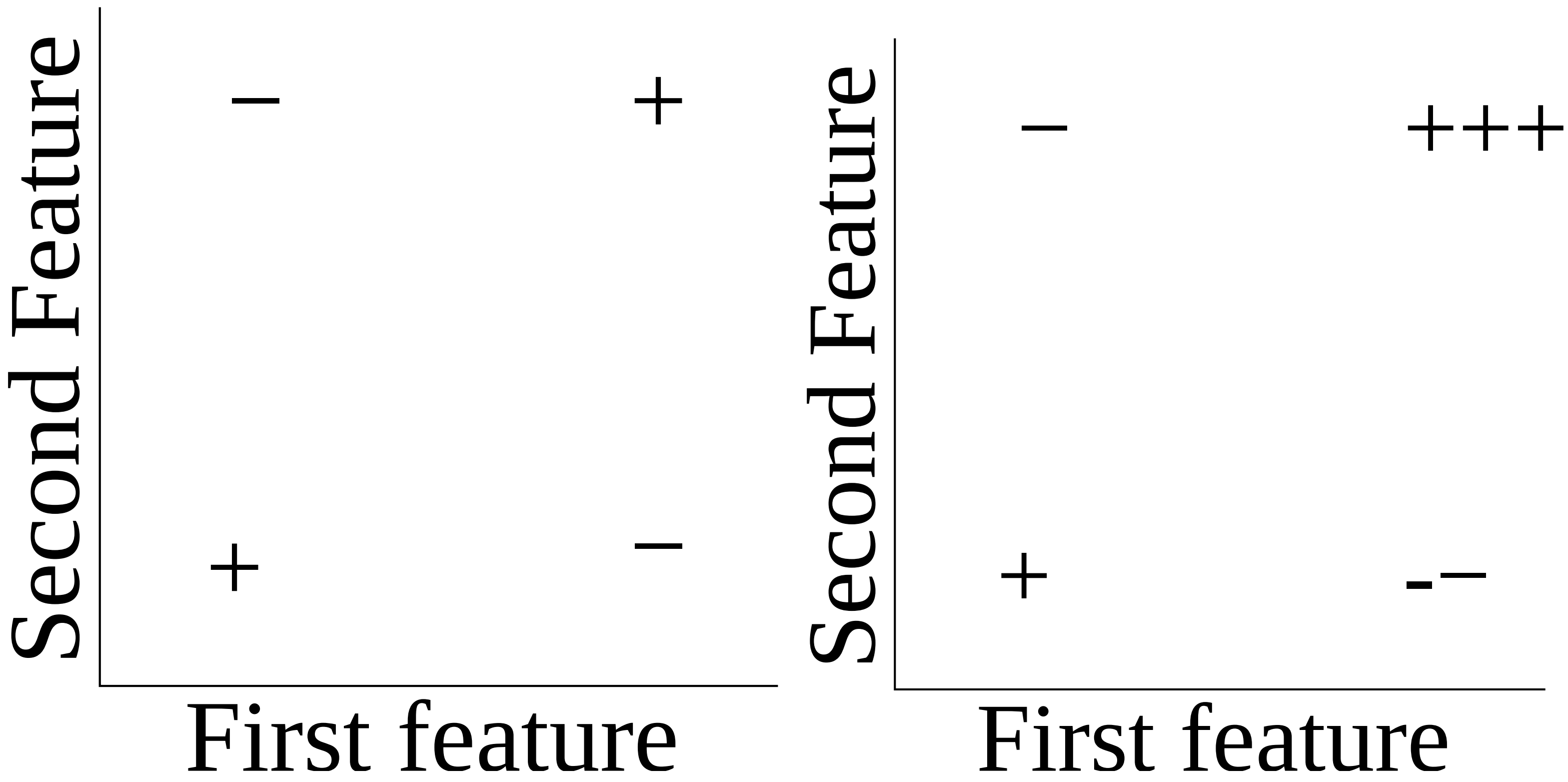
1. An algorithm which we hope predicts better than random.

2. An algorithm which can learn with respect to different emphasis on the data.

What is a good classifier?



What is a good classifier?



A distribution is a soft version of cloning examples.

The learning algorithm should find:

$$\min_h \sum_{i=1}^{m} D(i) I(h(x) \neq y)$$

Weak Learning algorithms

1. Many algorithms easily modified to take distributions
   (a) Decision trees (or Decision "stumps")
   (b) Neural network classifier
   (c) Naive Bayes classifier
2. All classification algorithms can be made to work by rejection sampling according to $D(i)$.

Next: the Adaboost algorithm

1. $D_1(i) = \frac{1}{m}$

2. For $k = 1, ..., K$

   (a) $h_t = \text{LEARN}(D_t, S)$

   (b) $\epsilon_k = E_{x,y \sim D_k} I(h_k(x) \neq y)$

   (c) $\alpha_k = \frac{1}{2} \ln \frac{1-\epsilon_k}{\epsilon_k}$

   (d) $D_{k+1}(i) = D_k(i) \frac{e^{-\alpha_k y_i h_k(x_i)}}{Z_k}$

3. Output $h(x) = \text{sign}\left(\sum_{k=1}^{K} \alpha_t h_k(x)\right)$

Adaboost Analysis

Theorem: (Train set boosting) If the weak learning algorithm errs at most a $\frac{1}{2} - \epsilon$ portion of the time, then the train error rate of the average is at most $e^{-2K\left(\frac{1}{2}-\epsilon\right)^2}$.

Theorem: (boosting) If the train error is near to the true error then Adaboost is a boosting algorithm.

Boosting side notes

Variants for real-valued outputs

Variants for multiclass classification

Variants with different update functions

Much analysis

Outline

1. Theoretical Motivations

2. A Zoology of Averages

   (a) Bagging (A testament to the effectiveness of averaging)

   (b) Boosting (+ the boosting guarantee)

Conclusion

Averaging techniques *dominate* in supervised classification learning.

Some (Boosting for example) have more motivation than others.

All trade computation for accuracy.