Measures of Diversity in Combining Classifiers

Part 2. Non-pairwise diversity measures

For fewer cartoons and more formulas:

http://www.bangor.ac.uk/~mas00a/publications.html

Random forest :  , x, θ_k (i.i.d, k=1,…,L), L is large

Strength and correlation:

D(x): the class label of x suggested by D

Define <u>margin function</u> for a random forest to be

$$mr(x, \omega_i) = P_\theta(D(x)=\omega_i) - \max_{t \neq i} P_\theta(D(x)=\omega_t) ,$$

and the <u>strength</u> of the set of classifiers to be

$$s = \boldsymbol{E}_{x,\omega} [mr(x, \omega)]$$

Denote $\omega_s = \mathrm{argmax}_{t \neq i} P_\theta(D(x)=\omega_t)$ and define <u>raw margin function</u> to be

$$rmr(x, \omega_i, \theta) = I(D(x)=\omega_i) - I(D(x)=\omega_s) ,$$

where $I(.)$ is an indicator function.

The probability of error of the ensemble is bounded as follows

$$PE^* \leq \boxed{\rho} \; (1 - s^2) / s^2$$

(mean) correlation
between rmr(D_i), rmr(D_k)
(averaged across all pairs
of classifiers)

strength

"Although the bound is likely to be loose, it fulfils the same suggestive function for random forests as VC-type bounds do for other types of classifiers."

The 2-class case:

$$mr(x, \omega_i) = 2\, P_\theta(D(x)=\omega_i) - 1, \quad i = 1,2$$

the strength of the set of classifiers is

$$s = \boldsymbol{E}_{x,\omega}\, [mr(x, \omega)]$$

$$\approx 2/N \left[\, \Sigma_1 P_\theta(D(x)=\omega_1) + \Sigma_2 P_\theta(D(x)=\omega_2) \,\right] - 1$$

True label ω_1

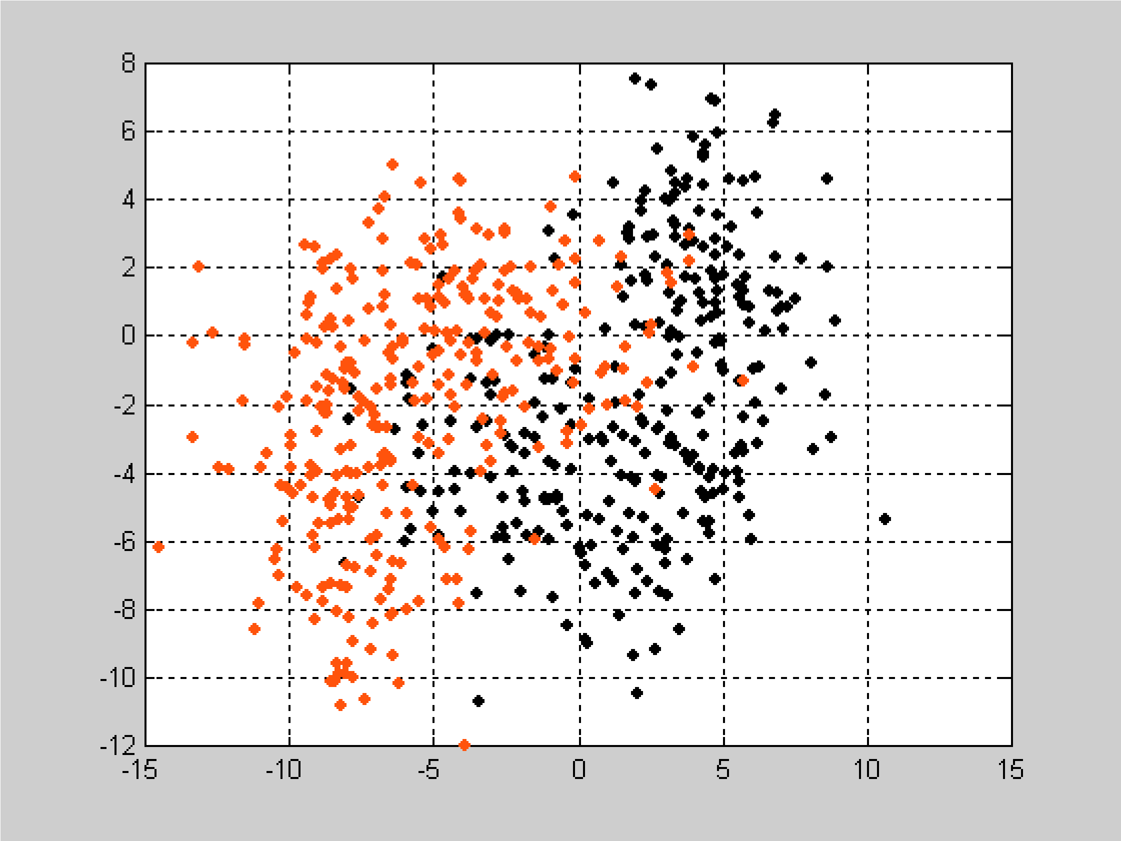
True label ω_2

The correlation ρ can be calculated as the averaged pairwise correlation between the oracle outputs

NB. Both are just estimates!

An example:

banana-shaped data (gendatb routine from Matlab toolbox PRtools)



Training

N = 600 data points

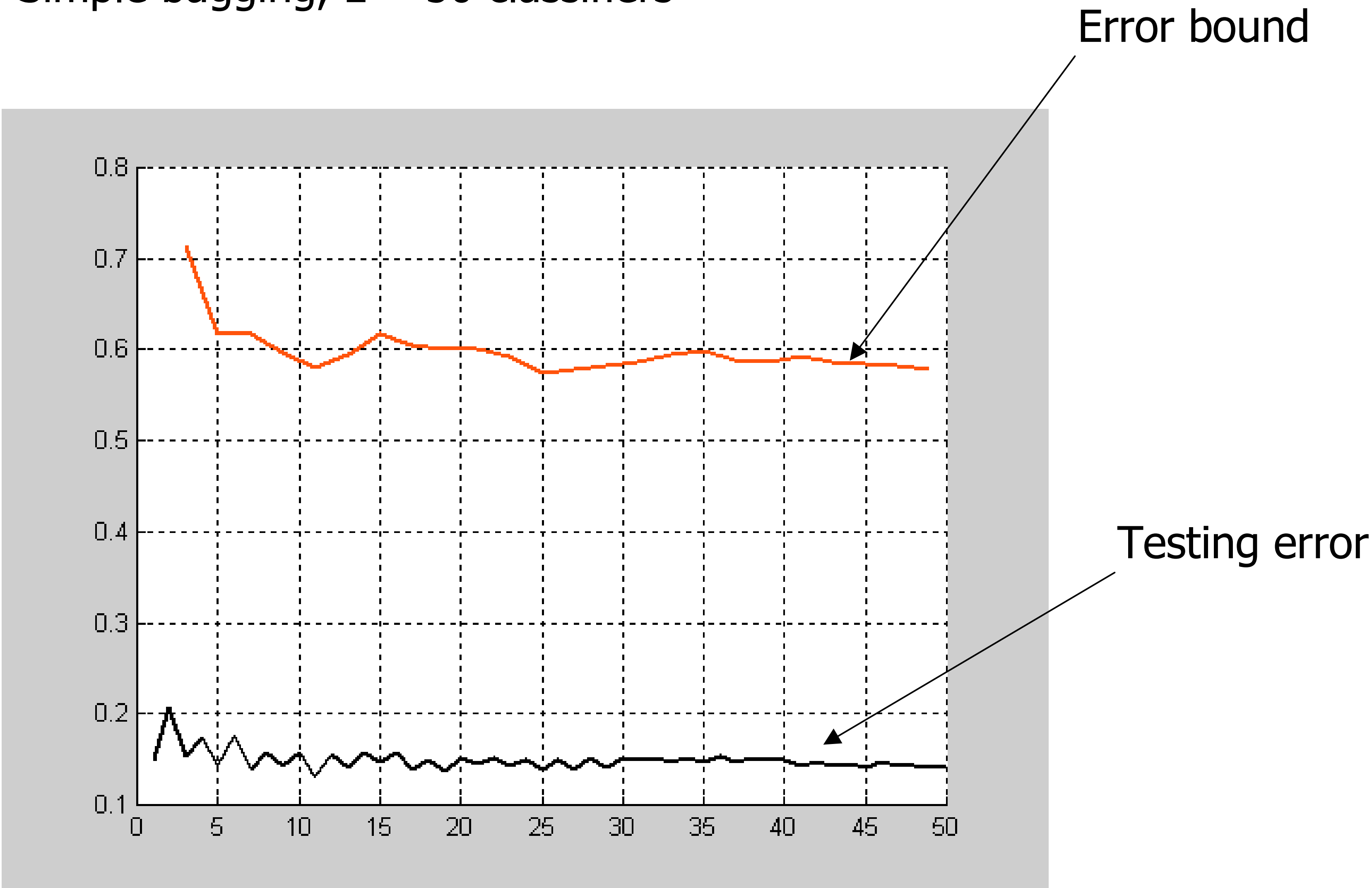
Testing

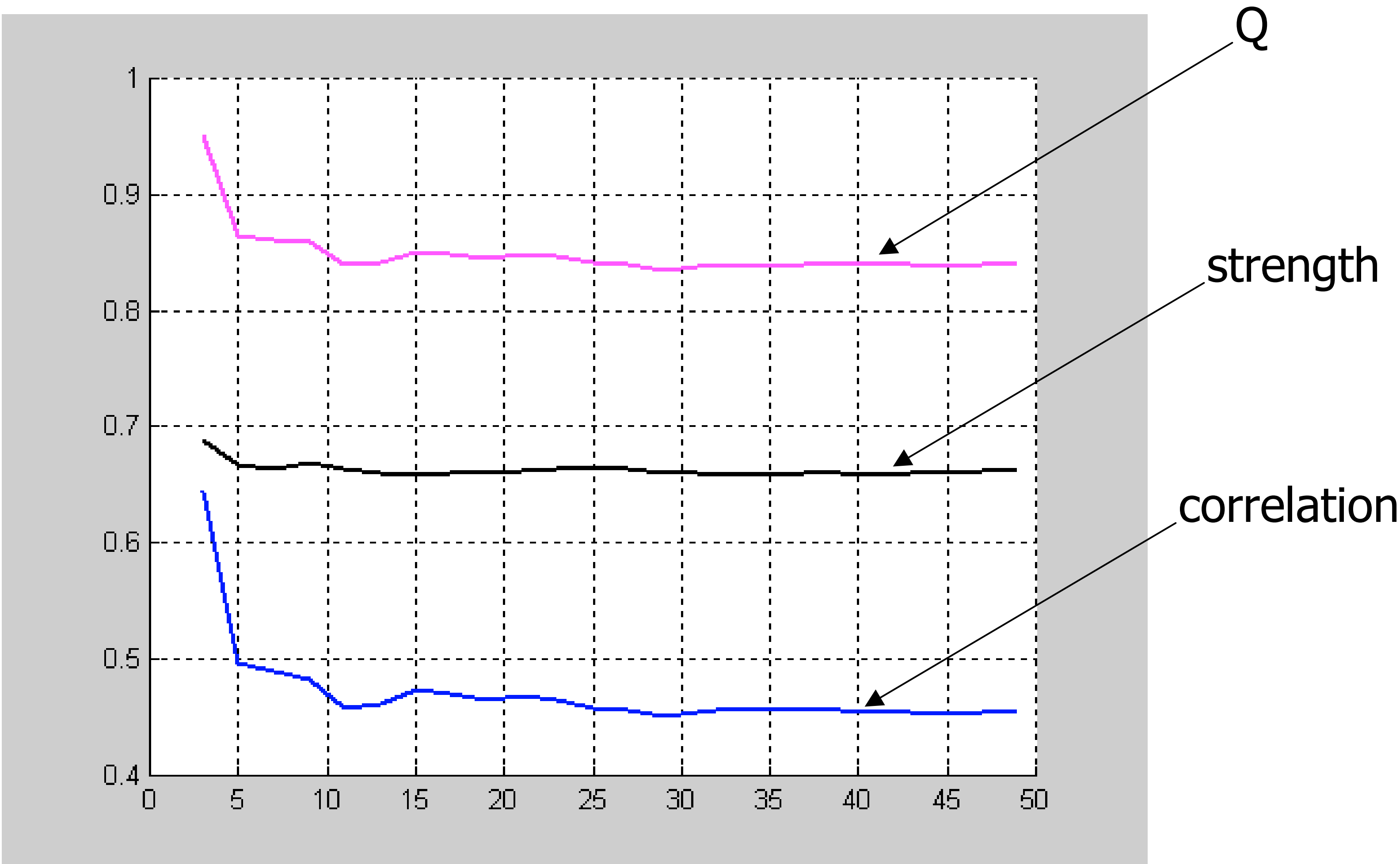
(a separate set)

N = 600 data points

The idea was to avoid using OB estimates which anyway simulate estimates on an independent testing set of the same size

Simple bagging, L = 50 classifiers

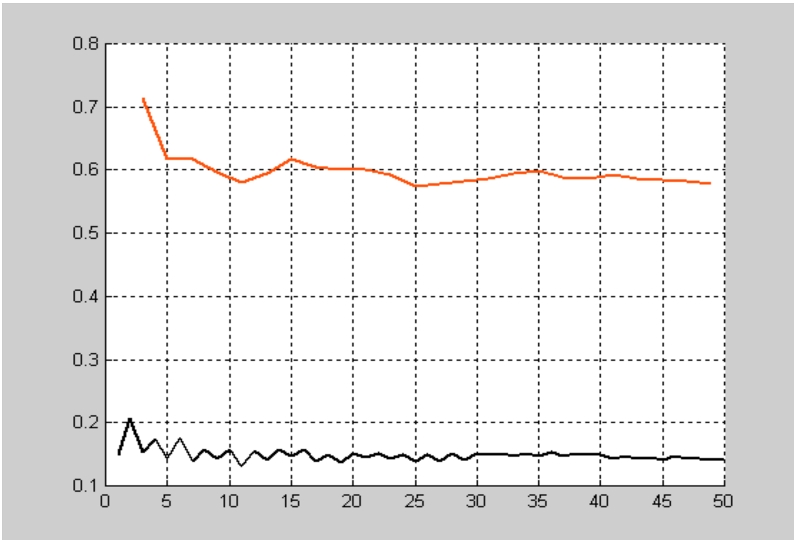


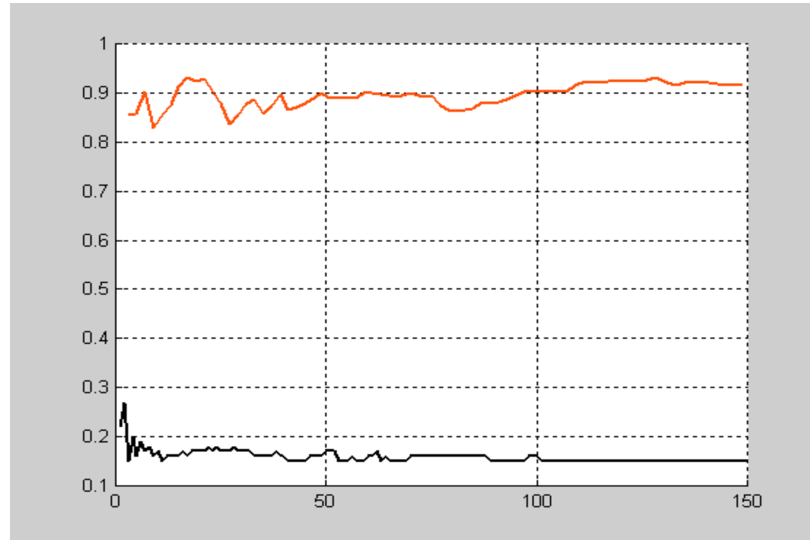


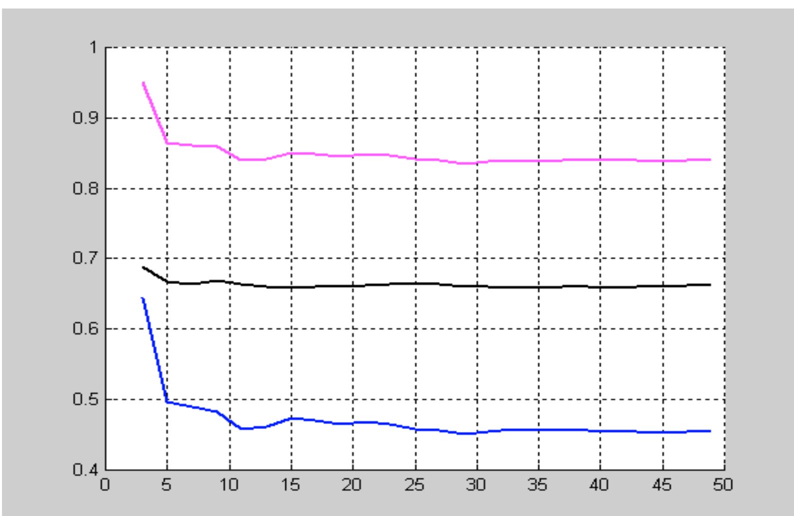
Q
strength
correlation
1
0.9
0.8
0.7
0.6
0.5
0.4
0
5
10
15
20
25
30
35
40
45
50

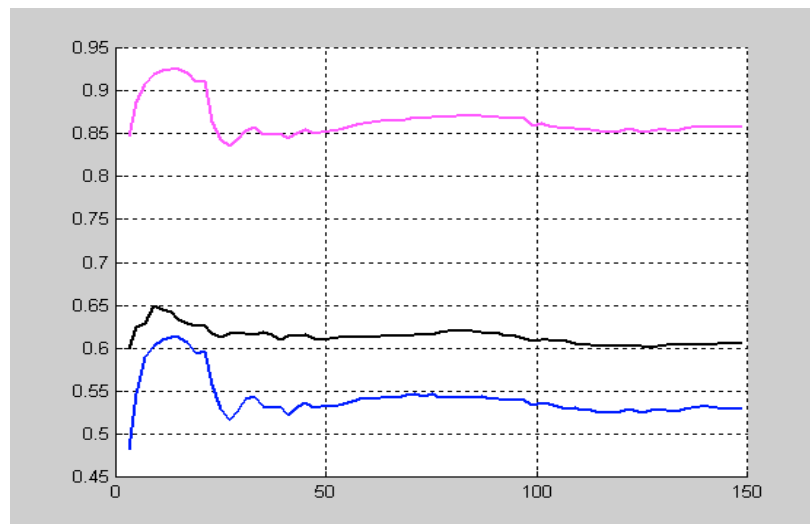
L=50, N=600

L=150, N=100









Is strength related to accuracy?

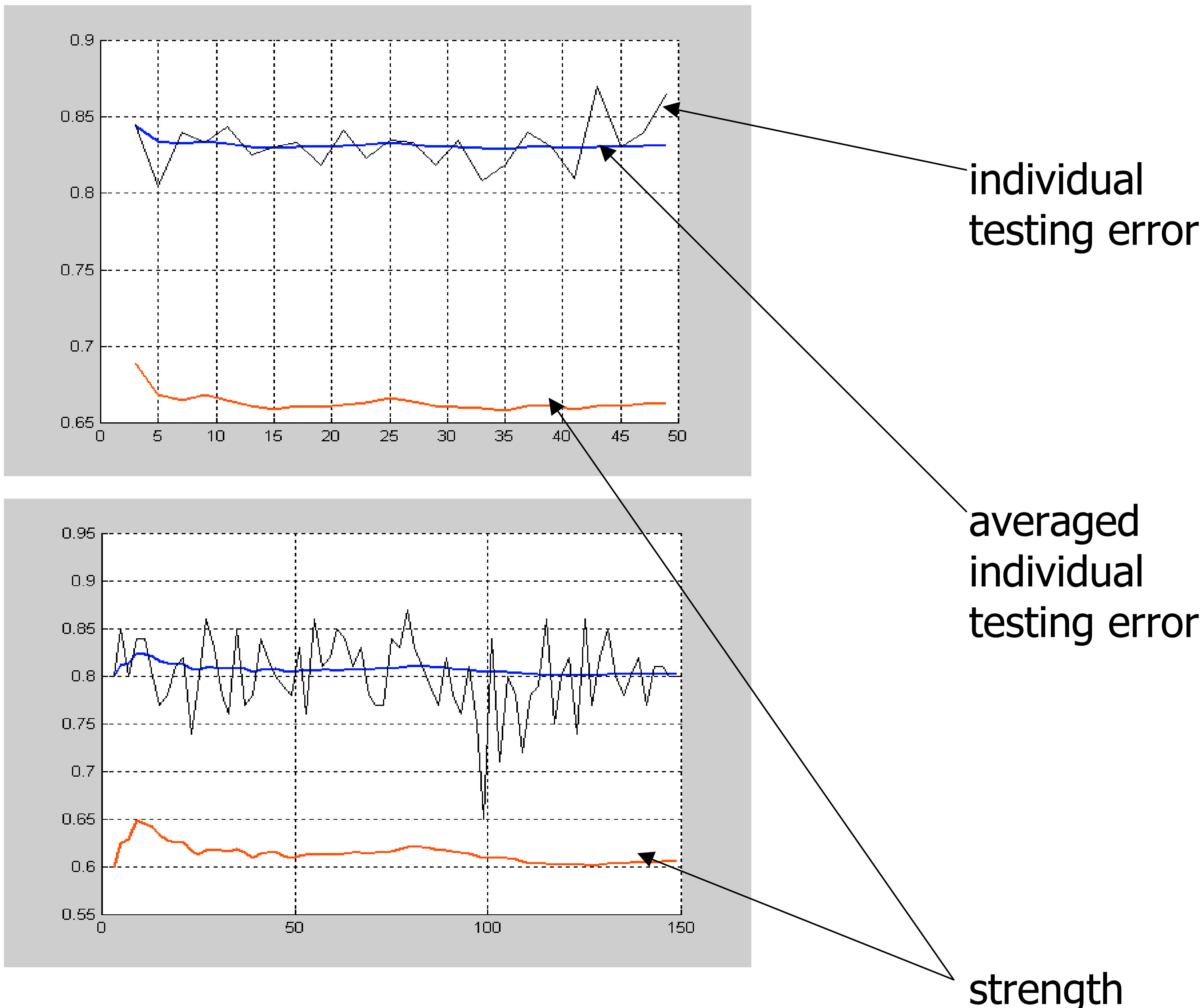
true labels

guessed labels

true labels	D_1		D_2		D_3		
1	1	1	1	1	1	1	
1	2	0	2	0	2	0	
1	1	1	1	1	1	1	$P_\theta(D(x)=\omega_1)$
1	1	1	2	1/2	1	2/3	
1	1	1	1	1	1	1	
2	1	0	2	1/2	1	1/3	
2	2	1	1	1/2	2	2/3	
2	1	0	2	1/2	2	2/3	$P_\theta(D(x)=\omega_2)$
2	2	1	2	1	1	2/3	
2	2	1	2	1	1	2/3	
	7/10		7/10		6/10		accuracy
	4/10		4/10		1/3		strength

2 x (7/10) - 1

(2/10) x (20/3) - 1


0.9
0.85
0.8
0.75
0.7
0.65
0 5 10 15 20 25 30 35 40 45 50
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0 50 100 150
individual testing error
averaged individual testing error
strength

Part 2: Non-pairwise diversity measures

0. A note on pairwise diversity (ρ) for random forests
- Measures based on a single data point + averaging (entropy, spread, KW variance)
- Interrater agreement (kappa for multiple raters)
- Measures based on difficulties of the data points
- Relationship with accuracy
- Open problems

Now we look at the whole ensemble of classifiers.

Classifier outputs

Oracle(binary)

1	0	1	1	1	0

Class labels
(abstract level)

ω_1	ω_2	ω_1	ω_1	ω_3	ω_1

Ordered list of class labels

ω_{12}	ω_8	ω_{11}	ω_{10}	ω_9	ω_{11}
ω_{11}	ω_{12}	ω_9	ω_{11}	ω_8	ω_{10}
ω_{10}	ω_2	ω_{10}	ω_{12}	ω_{10}	ω_8

Continuous-valued
(measurement level)

ω_1	ω_2	ω_3	ω_4
0.1	0.4	0.3	0.2
0.3	0.3	0.3	0.1
0.2	0.1	0.7	0.0

called “decision profile”
(remember for later)

- Measures based on a single data point (case, instance, example, object, whatever) and subsequently averaged over the whole data set.

2. Measures based on all data points.

For oracle outputs and L = 8 classifiers, are these diverse?

1	1	1	1	1	1	1	1	No-o-o-o-o-o-o!
0	0	0	0	0	0	0	0	Nope.
1	0	1	0	1	0	1	0	Yes.
1	1	1	1	0	0	0	0	Yes.

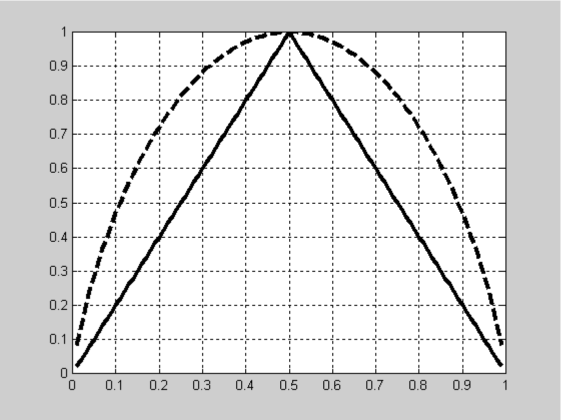
ENTROPY (oracle outputs)

How do we measure how far we are from the desired pattern of L/2 0's and L/2 1s for N objects?

$$E = \frac{1}{\lceil L-1 \rceil N} \Sigma_k \left[\min \{ \ \Sigma\, 0's\ ,\ \Sigma\, 1's \ \} \right]_k$$

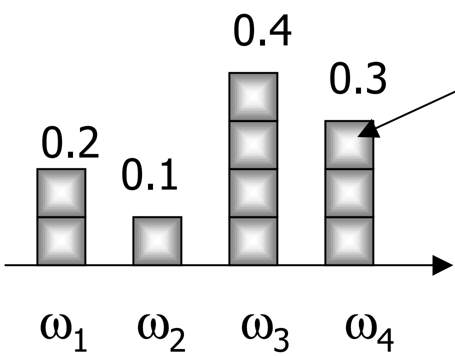
Consider the output 0 or 1 as a random variable with relative frequencies $p_0 = (\Sigma\, 0's) / L$ and $p_1 = (\Sigma\, 1's) / L$, respectively. Then the (proper) formula for the entropy of the distribution, averaged across the N data points will be

$$H = -\ \frac{1}{N} \Sigma_k \left[p_0 \log p_0 + \ p_1 \log p_1 \right]_k$$



[Cunningham Carney, 2000]

ENTROPY (label outputs)



Votes of L=10 classifiers for a single x*

$$H = -1/N \; \Sigma_k \left[\Sigma_i \, p_i \log p_i \right]_k$$

Breiman's Bias-Variance decomposition, 1996

Assume that classifier output for a given x* is a random variable with p.m.f. $P(\omega_1|x^*,D), \ldots, P(\omega_c|x^*,D)$. The classification error is

$$P(\text{error}|x^*) = 1 - \Sigma_j\ P(\omega_j|\ x^*)\ P(\omega_j|\ x^*, D)$$

$$= \ 1 - \{\ P(\omega_B|\ x^*) - P(\omega_B|\ x^*) - \Sigma_j\ P(\omega_j|\ x^*)\ P(\omega_j|\ x^*, D)\ \}$$

$$= \ [1 - P(\omega_B|\ x^*)] + \Sigma_j\ [P(\omega_B|\ x^*) - P(\omega_j|\ x^*)]\ P(\omega_j|\ x^*, D)$$

$$= \ P_B(\ x^*) + \boxed{\Sigma_j\ [P(\omega_B|\ x^*) - P(\omega_j|\ x^*)]\ P(\omega_j|\ x^*, D)}$$

$$[P(\omega_B|\ x^*) - P(\omega_s|\ x^*)]\ P(\omega_s|\ x^*, D)]$$

bias

$$+\ \Sigma_{j \neq s}\ [P(\omega_B|\ x^*) - P(\omega_j|\ x^*)]\ P(\omega_j|\ x^*, D)$$

spread

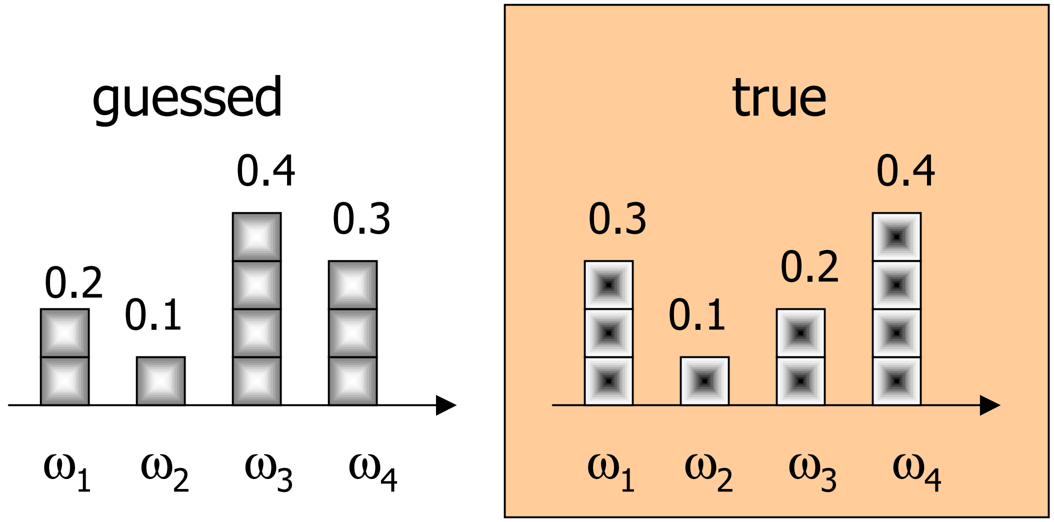
$P_B(x^*)$

$+ [P(\omega_B | x^*) - P(\omega_s | x^*)]\, P(\omega_s | x^*, D)]$ (bias)

$+ \Sigma_{j \neq s} [P(\omega_B | x^*) - P(\omega_j | x^*)]\, P(\omega_j | x^*, D)$ (spread)

Is the spread related to diversity?

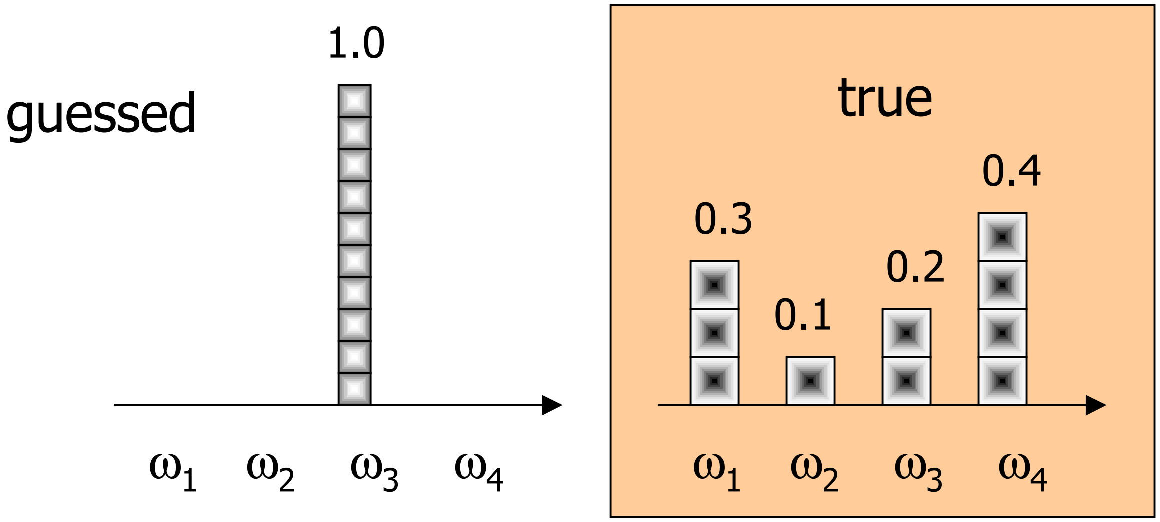
An example: If we drew a classifier at random from the distribution P_D,



P(error|x) = 0.6 + [0.4-0.2] 0.4 + [0.1x0.2+0.3x0.1] = 0.73

$P(\text{error}|x) = 0.6 + \underbrace{\boxed{[0.4-0.2]\ 0.4}}_{0.08} + \underbrace{\boxed{[0.1\text{x}0.2+0.3\text{x}0.1]}}_{0.05} = \underline{0.73}$

Take majority vote. This means "decide always ω_s for x*".



$P(\text{error}|x) = 0.6 + \underbrace{\boxed{[0.4-0.2]\ 1.0}}_{0.2} = \underline{0.80}$

KW variance (label outputs)

[Kohavi Wolpert, 1996, Bias plus variance decomposition for zero-one loss functions]

The c-class case:

$$P(\text{error}|x) = \text{bias}^2(x)+\text{variance}(x)+\text{noise}^2(x)$$

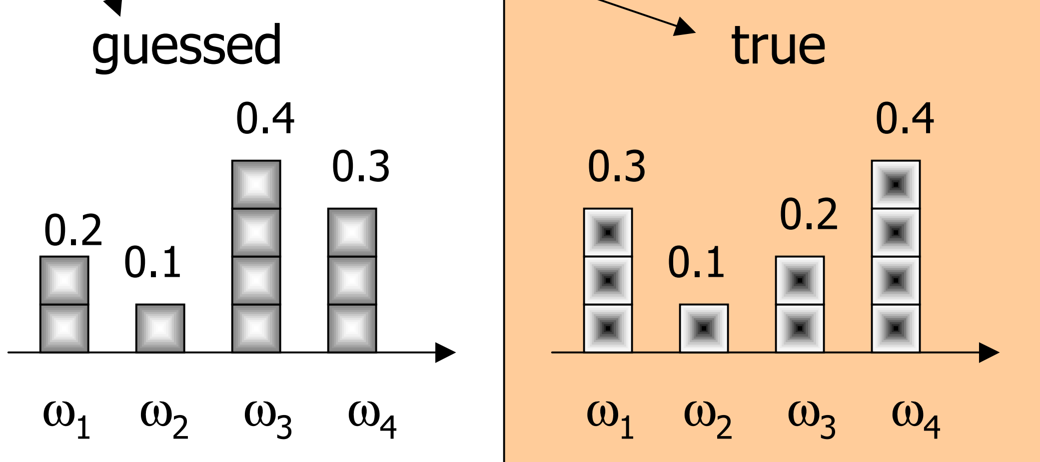
bias^2(x) $\frac{1}{2} \Sigma_{\omega} (P_{true}(\omega|x) - P_{guessed}(\omega|x))^2$

variance(x) $\frac{1}{2} (1 - \Sigma_{\omega} (P_{guessed}(\omega|x))^2$

noise^2 (x) $\frac{1}{2} (1 - \Sigma_{\omega} (P_{true}(\omega|x))^2$

$$\underline{\mathrm{bias}^2(x)} = \tfrac{1}{2} \Sigma_\omega \left(P_{true}(\omega|x) - P_{guessed}(\omega|x)\right)^2$$

$$\tfrac{1}{2}\ (0.3 - 0.2)^2 + (0.4 - 0.2)^2 + (0.3 - 0.4)^2 = \mathbf{0.03}$$



$$\underline{\mathrm{variance}(x)} = \tfrac{1}{2} \left(1 - \Sigma_\omega \left(P_{guessed}(\omega|x)\right)^2\right)$$

$$\tfrac{1}{2}\ [\ 1 - ((0.2)^2 + (0.1)^2 + (0.4)^2 + (0.3)^2\ )] = \mathbf{0.35}$$

$$\underline{\mathrm{noise}^2\ (x)} = \tfrac{1}{2} \left(1 - \Sigma_\omega \left(P_{true}(\omega|x)\right)^2\right)$$

$$\tfrac{1}{2}\ [\ 1 - ((0.3)^2 + (0.1)^2 + (0.2)^2 + (0.4)^2\ )] = \mathbf{0.35}$$

KW variance (oracle outputs)

Consider again the output 0 or 1 as a random variable with relative frequencies $p_0 = (\Sigma\ 0\text{'s}) / L$ and $p_1 = (\Sigma\ 1\text{'s}) / L$, respectively. Then the variance is

$$\underline{\text{variance}(x)} = \tfrac{1}{2}\left(1 - (p_0)^2 - (p_1)^2\right)$$

Averaging across the whole data set,

$$KW = 1/(N \times L^2)\ \Sigma_k \left[\ (\Sigma\ 0\text{'s}) \times (\Sigma\ 1\text{'s})\ \right]_k$$

Curiously, KW and the averaged pairwise disagreement measure are related through

$$KW = (L-1)/(2L)\ \ D_{av}$$

- Measures based on a single data point (case, instance, example, object, whatever) and subsequently averaged over the whole data set.

2. Measures based on all data points.

Interrater agreement, kappa, (oracle outputs)

$L \times KW$

$$k = 1 - \frac{\Sigma_k \left[\, (\Sigma\, 0's) \times (\Sigma\, 1's) \, \right]_k}{N \;\; L \; (L - 1) \;\; p \;\; (1 - p)}$$

Number of data points

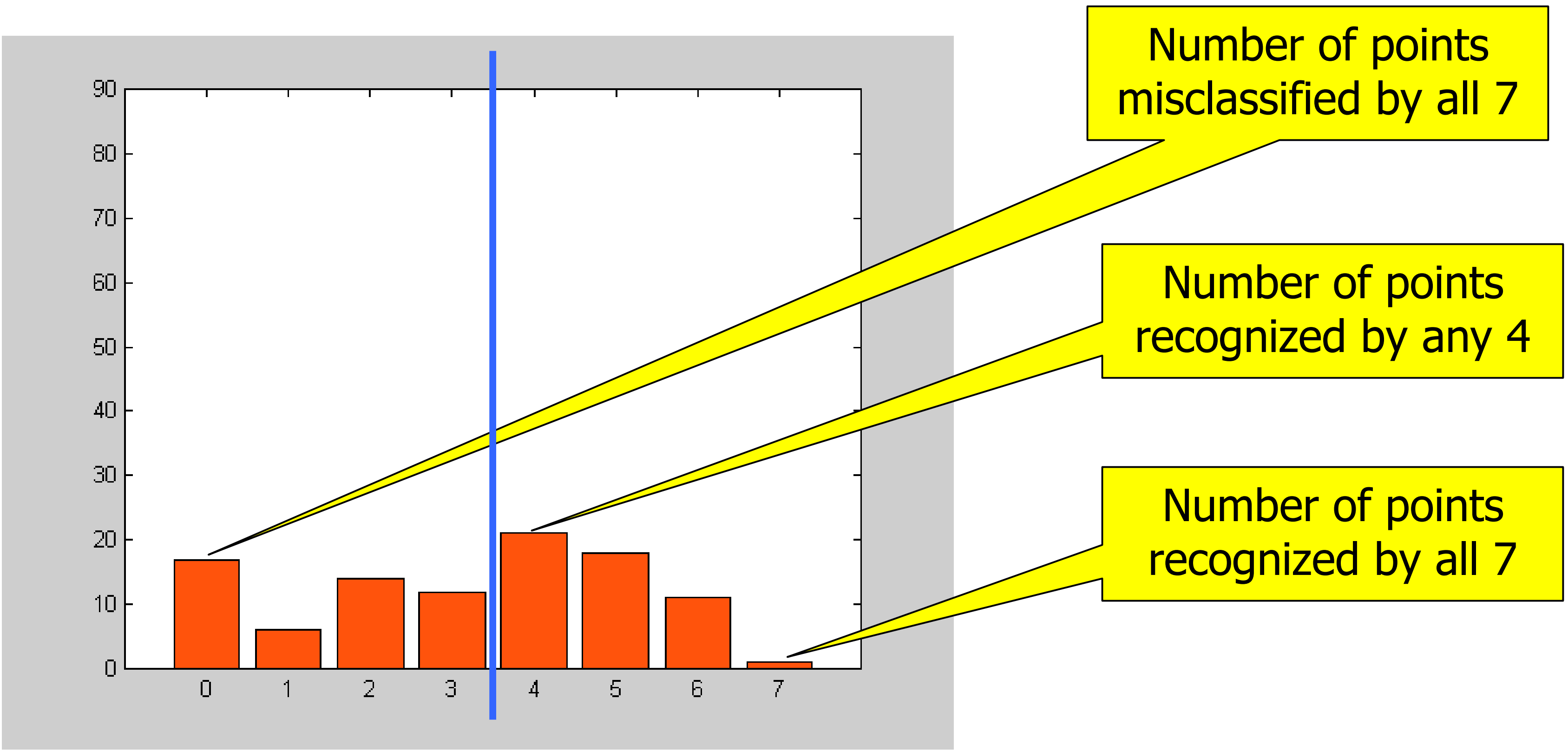
Number of classifiers

Averaged individual accuracy

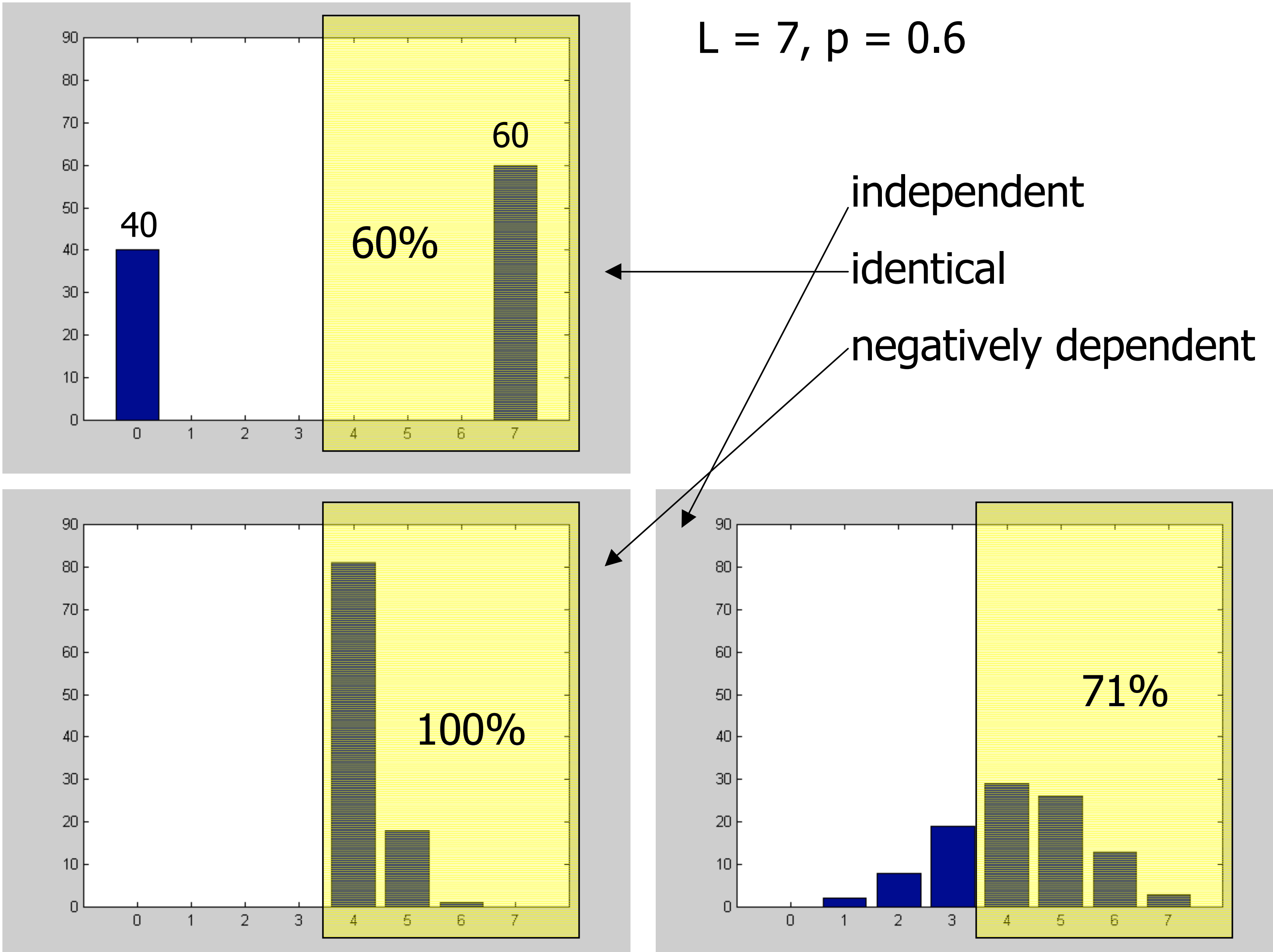
Measure of difficulty θ

[Hansen Salamon, 1990]

Define a random variable X = proportion of classifiers which correctly classify a randomly drawn sample x. Let L = 7.

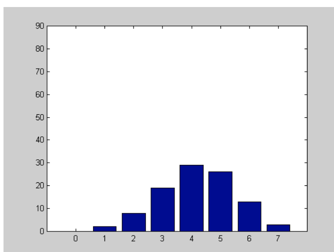


L = 7, p = 0.6


independent
identical
negatively dependent
40
60
60%
100%
71%

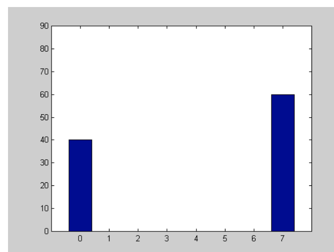
measure of diversity θ = Var(X)

independent



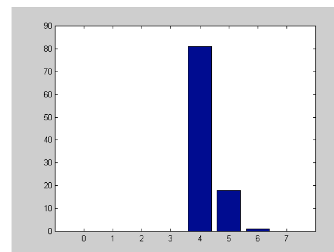
$\theta = 0.034$

identical



$\theta = 0.240$

diverse

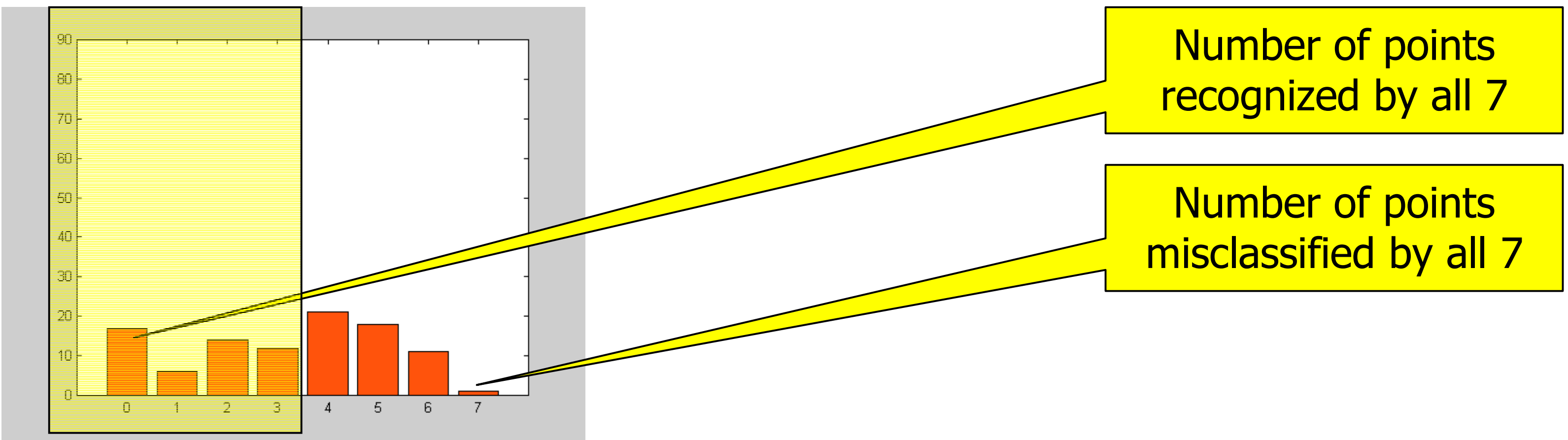


$\theta = 0.004$

Generalized diversity

[Partridge Krzanowski, 1997]

Define a random variable Y = proportion of classifiers which misclassify a randomly drawn sample x. (Y = 1 – X defined before)



Denote by p_i the probability that Y = i / L, and by p(k) the probability that k randomly chosen classifiers will fail on a randomly drawn x.

$p(1) = \Sigma_i \; p_i \times \; i \, / \, L$ (the probability of single classifier failing)

$p(2) = \Sigma_i \; p_i \times \; i \, (i - 1) / \, (L \, (L - 1))$ (the probability that two randomly chosen classifiers will fail together)

$GD = 1 - p(1)/p(2)$

Coincidence failure diversity

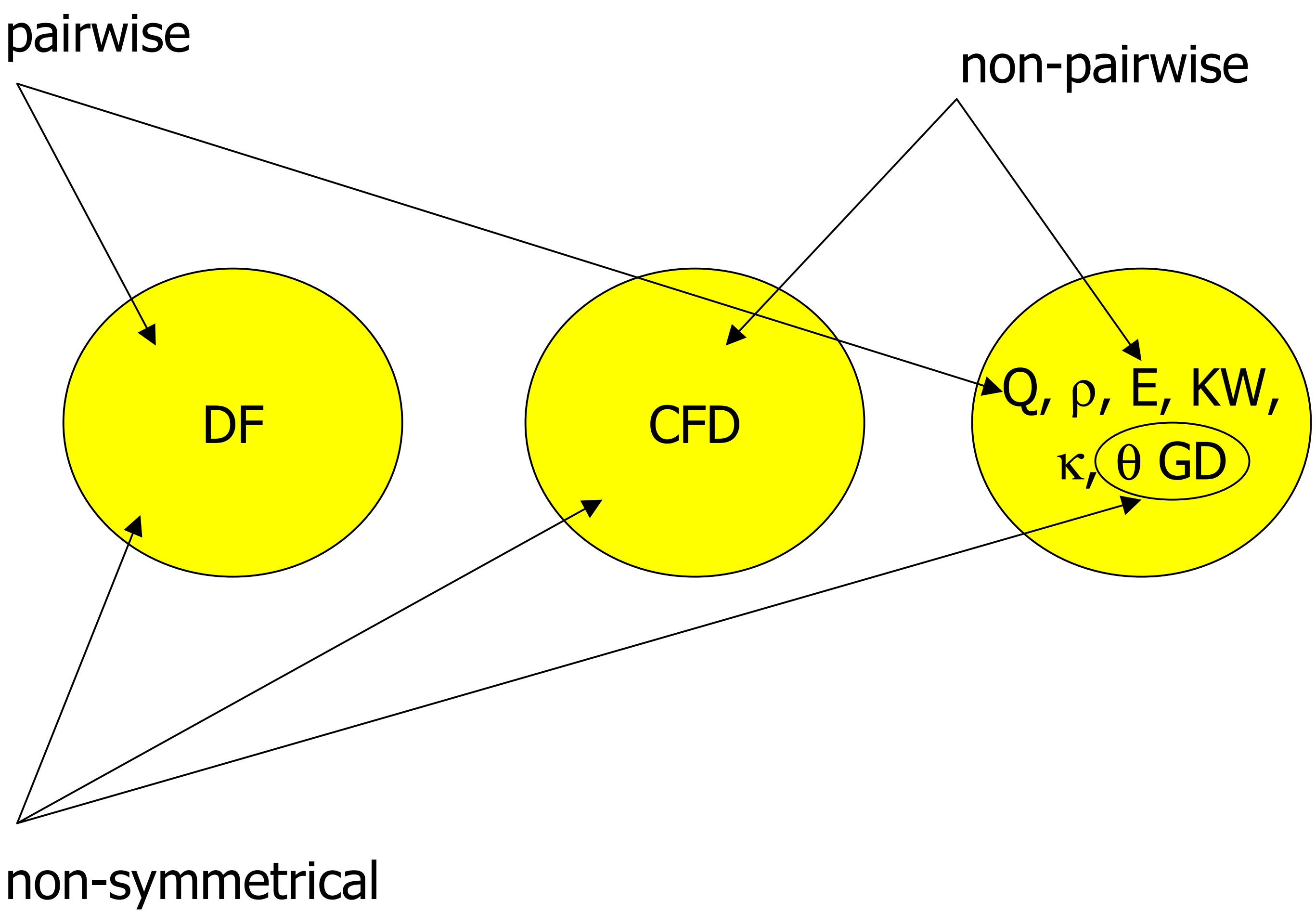
$$CFD = \begin{cases} 0, & \text{if } p_0 = 1, \\ 1/(1 - p_0) \, \Sigma_i \; p_i \times \; (L - i)/\,(L - 1), & \text{if } p_0 < 1 \end{cases}$$

Relationship between diversity and accuracy

Correlations between the improvement on the single best classifier and some diversity measures (WBC)

	Q	ρ	Dis	DF	κ	θ	GD	CFD
MAJ	-17	-21	33	18	-20	35	28	38
NB	-15	-20	32	20	-18	37	26	36
BKS	-15	-17	17	5	-15	16	18	18
WER	-15	-17	17	5	-16	17	19	18
MAX	-1	-0	20	38	0	45	7	11
AVR	-13	-15	34	33	-14	47	22	30
PRO	-11	-11	29	33	-11	44	18	24
DT	-12	-15	32	30	-14	44	22	29

Relationship between diversity measures



Open problems

- How to narrow down the study? (Use a specific methodology for building the ensemble)
- Some theory would not go amiss.
- Diversity for label outputs and continuous-valued outputs might lead somewhere.

The difficulty comes from the fact that the output of the classifiers are vectors

	ω_1	ω_2	ω_3	ω_4
D_1	0.1	0.4	0.3	0.2
D_2	0.3	0.3	0.3	0.1
D_3	0.2	0.1	0.7	0.0

similarity between distributions

(pairwise)

WAKE UP!!!

Vietri sul Mare, 27 09 02