Classifier Generating Methods and Stochastic Discrimination

Tin Kam Ho
Bell Laboratories
Lucent Technologies

Classifier Combination Methods

- Decision Optimization:

  find consensus among a given set of classifiers

- Coverage Optimization:

  create a set of classifiers that work best with a given decision combination function

Decision Optimization

- Develop classifiers with expert knowledge
- Try to make the best use of their decisions

  via majority/plurality vote, sum/product rule, probabilistic methods, Bayesian methods, rank/confidence score combination …

- The joint capability of the classifiers set an intrinsic limit on the combined accuracy
- There is no way to handle the blind spots

Example from a word recognition problem

- rank of true class for 20 word images among a lexicon of 1091 words:

image number	classifier number 1	2	3	4	5	6	7	8	9	10
1	102	1	55	1	597	34	393	1	303	19
2	25	4	478	193	498	213	707	45	956	996
3	29	342	2	86	1	941	36	798	448	260
4	2	5	208	76	5	4	16	9	565	183
5	221	90	84	351	259	1038	838	725	819	617
6	570	130	513	33	761	274	1	351	345	105
7	77	40	91	569	11	505	250	271	366	155
8	219	8	356	218	725	72	8	18	139	19
9	41	11	1	2	19	4	16	9	59	44
10	70	82	79	30	338	3	36	26	38	100
11	3	4	214	53	6	470	4	295	420	953
12	2	68	48	188	3	87	49	29	256	718
13	3	768	149	262	34	1046	235	842	289	909
14	68	657	121	289	192	325	979	45	3	953
15	1	7	91	139	512	26	16	37	47	44
16	3	157	113	16	70	6	17	5	163	192
17	69	8	663	780	773	130	724	32	260	953
18	407	100	97	45	710	169	147	18	1057	24
19	1	10	205	189	930	386	309	296	696	699
20	2	98	721	15	852	130	771	6	3	92

Difficulties in Decision Optimization

- Reliability versus overall accuracy
- Fixed or trainable combination function
- Simple models or combinatorial estimates
- How to model complementary behavior

Coverage Optimization

- Fix a decision combination function
- Generate classifiers automatically and systematically

  via training set sub-sampling (stacking, bagging, boosting),
  subspace projection (RSM),
  superclass/subclass decomposition (ECOC),
  random perturbation of training processes, noise injection …

- Need enough classifiers to cover all blind spots
  (how many are enough?)
- What else is critical?

Difficulties in Coverage Optimization

- What kind of differences to introduce:
  - Subsamples? Subspaces? Super/Subclasses?
  - Training parameters?
  - Model geometry?
- 3-way tradeoff:
  - discrimination + diversity + generalization
- Effects of the form of component classifiers

Dilemmas and Paradoxes

- Weaken individuals for a stronger whole?
- Sacrifice known samples for unseen cases?
- Seek agreements or differences?

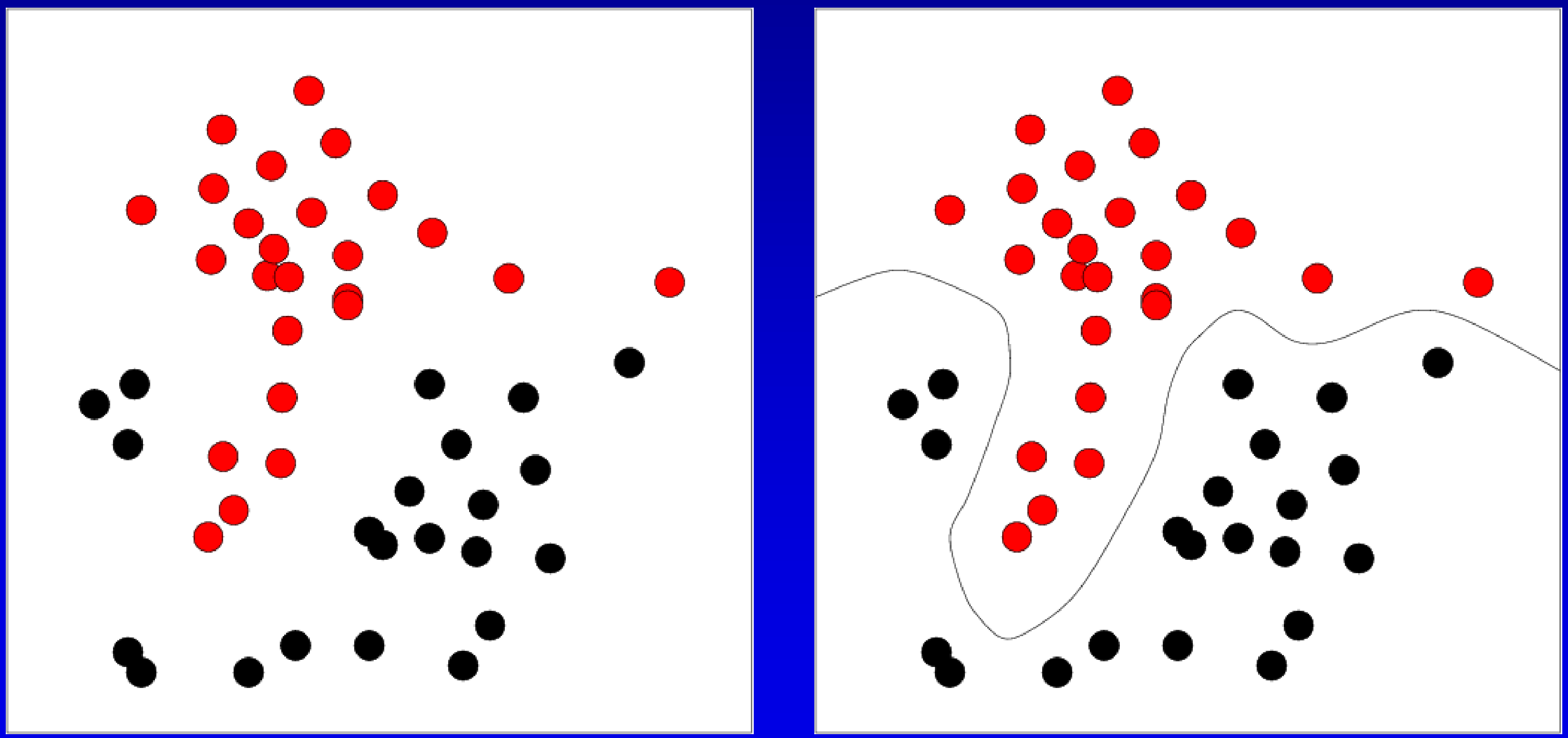
Model of Complementary Decisions

- Statistical independence of decisions: assumed or observed?
- Collective vs. point-wise error estimates
- Related estimates of neighboring samples

Stochastic Discrimination

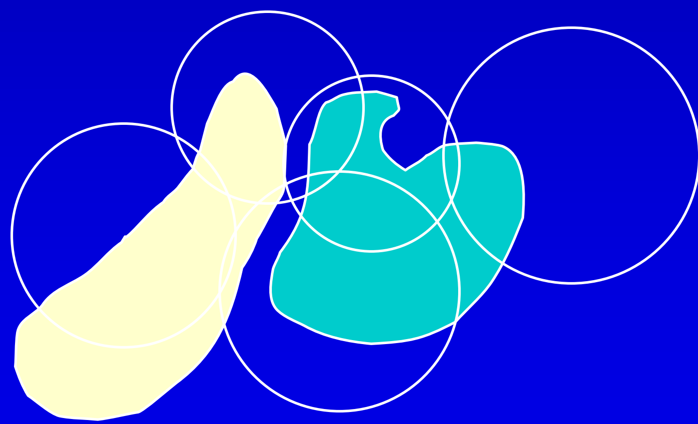
- A mathematical theory that relates several key concepts in pattern recognition:
  - Discriminative power
  - Complementary information
  - Generalization power
- It offers a way to describe complementary behavior of classifiers

Supervised Classification -- Discrimination Problems



Stochastic Discrimination

- Make random guess of class models
- Select and combine the guesses to build a classifier



History

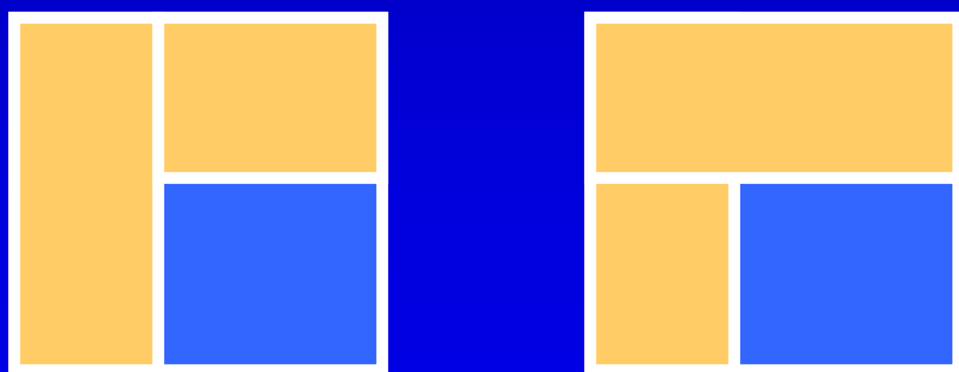
- Mathematical theory
  [Kleinberg 1990 AMAI, 1996 AoS, 2000 MCS]
- Development of theory
  [Berlind 1994 Thesis, Chen 1997 Thesis]
- Algorithm outlines [Kleinberg 2000 PAMI]
- Algorithms, experimentation, variants:
  [Kleinberg, Ho, Berlind, Bowen, Chen, Favata, Shekhawat, 1993 – 2002]

Key Concepts and Tools in SD

- Set-theoretic abstraction
- Symmetry of probabilities in model or feature spaces
- Enrichment / Uniformity / Projectability
- Convergence of discriminant by the law of large numbers

Set-Theoretic Abstraction

- Study classifiers by their decision regions
- Ignore all algorithmic details
- Two classifiers are equivalent if their decision regions are the same



The 0th Example

- Given a set of 3 points $S = \{a, b, c\}$
- Consider subsets of S with 2 members:

  $s_1 = \{a, b\}$ $s_2 = \{a, c\}$ $s_3 = \{b, c\}$
- Each s_i covers 2/3 of the members in S
- Let $M = \{ s_1, s_2, s_3 \}$
- Each point of S is covered by:

  $a \in s_1, s_2$ 2/3 of members in M

  $b \in s_1, s_3$ 2/3 of members in M

  $c \in s_2, s_3$ 2/3 of members in M
- 2 models / 3 models = 2 points / 3 points
- This symmetry comes from the uniformity of M:

  M is unbiased for members of S

Uniformity Implies Symmetry: The Counting Argument

Count the number of pairs (q,m) such that
"model m covers point q", call this number N

If each point is covered by the same number X of models (the collection is a uniform cover),

N = 3 point x X covering models each point
N = 2 point in each model x Y models

$3X = 2Y$ $\qquad$ $X / Y = 2 / 3$

The 1st Example

- Given a feature space F containing a set A with 10 points:

  q0 q1 q2 q3 q4 q5 q6 q7 q8 q9

- Consider all subsets m of F that cover exactly 5 points of A, e.g.,

  $m = \{q1, q2, q6, q8, q9\}$

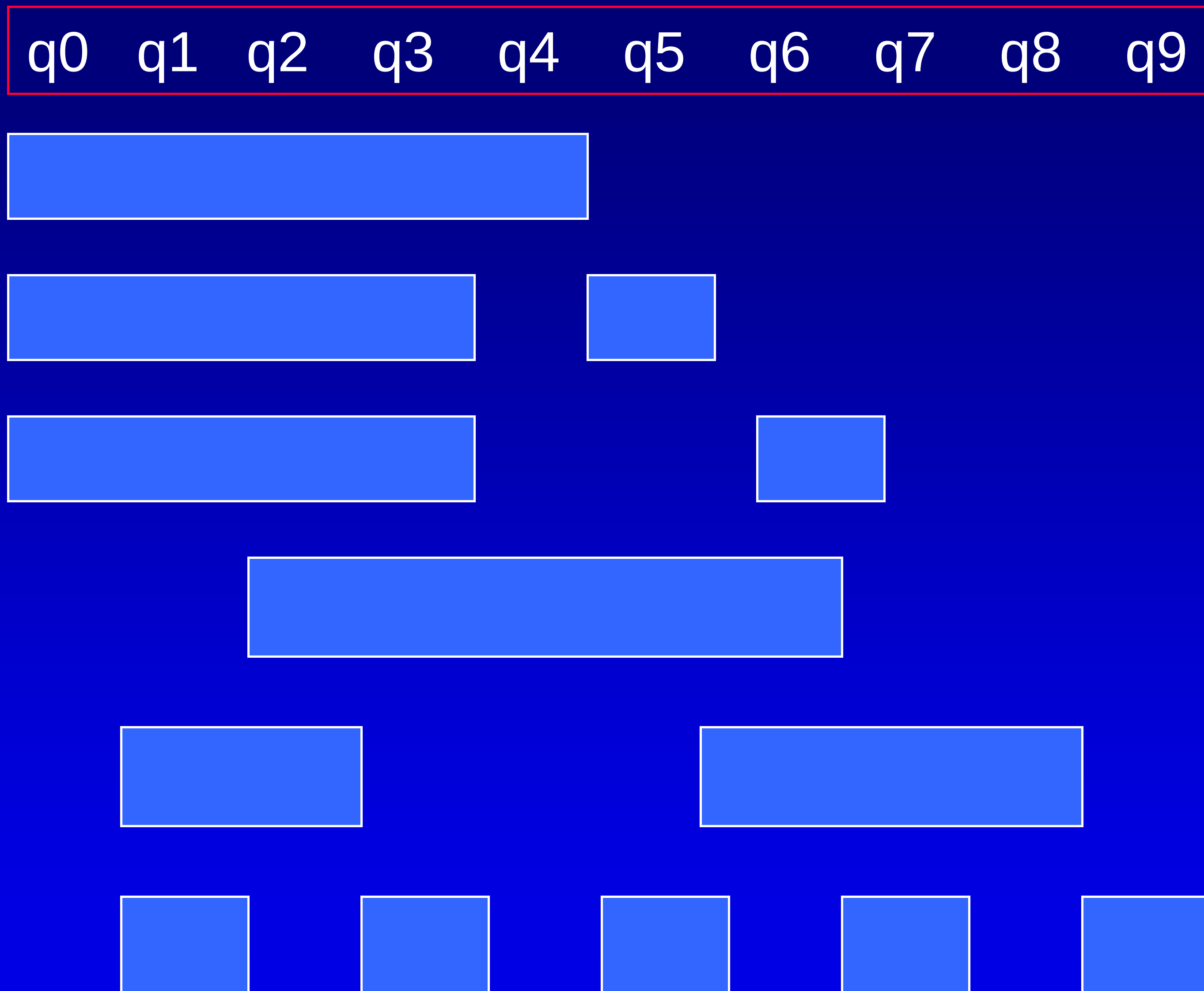
- Each model m has captured 5/10 = 0.5 of A

  $Prob_F (q \in m | q \in A) = 0.5$

- Call this set of models $M_{0.5, A}$

Some Members of M $_{0.5,\ A}$

$$\{q_0, \quad q_1, \quad q_2, \quad q_3, \quad q_4\}$$
$$\{q_0, \quad q_1, \quad q_2, \quad q_3, \quad q_5\}$$
$$\{q_0, \quad q_1, \quad q_2, \quad q_3, \quad q_6\}$$
$$\ldots$$


q0 q1 q2 q3 q4 q5 q6 q7 q8 q9

- There are C(10,5) = 252 models in $M_{0.5, A}$
- Permute this set randomly to give $m_1, m_2, \ldots, m_{252}$

Table 1: Models m_i in $M_{0.5,A}$ in the order of $M = m_1, m_2, \ldots, m_{252}$. Each model is shown with its elements denoted by the indices i of q_i in A. For example, $m_1 = \{q_3, q_5, q_6, q_8, q_9\}$.

m_i	elements	m_i	elements	m_i	elements	m_i	elements	m_i	elements	m_i	elements
m_1	35689	m_{43}	12689	m_{85}	24578	m_{127}	01469	m_{169}	02468	m_{211}	02458
m_2	01268	m_{44}	04569	m_{86}	23568	m_{128}	03679	m_{170}	35678	m_{212}	13457
m_3	04789	m_{45}	01245	m_{87}	01267	m_{129}	04579	m_{171}	03589	m_{213}	24689
m_4	25689	m_{46}	01458	m_{88}	01257	m_{130}	01237	m_{172}	34679	m_{214}	03478
m_5	02679	m_{47}	15679	m_{89}	05679	m_{131}	24789	m_{173}	12346	m_{215}	23589
m_6	34578	m_{48}	12457	m_{90}	24589	m_{132}	45689	m_{174}	12458	m_{216}	24679
m_7	13459	m_{49}	02379	m_{91}	04589	m_{133}	16789	m_{175}	35789	m_{217}	02456
m_8	01238	m_{50}	02568	m_{92}	12467	m_{134}	13479	m_{176}	02358	m_{218}	05689
m_9	12347	m_{51}	12357	m_{93}	13578	m_{135}	02349	m_{177}	35679	m_{219}	12789
m_{10}	01579	m_{52}	14678	m_{94}	02369	m_{136}	13469	m_{178}	13458	m_{220}	02346
m_{11}	34589	m_{53}	12678	m_{95}	12469	m_{137}	03678	m_{179}	01459	m_{221}	23489
m_{12}	03459	m_{54}	23567	m_{96}	04567	m_{138}	23679	m_{180}	03479	m_{222}	23467
m_{13}	23459	m_{55}	02789	m_{97}	14679	m_{139}	46789	m_{181}	14789	m_{223}	12489
m_{14}	02457	m_{56}	24567	m_{98}	13467	m_{140}	01468	m_{182}	23678	m_{224}	14589
m_{15}	02368	m_{57}	13569	m_{99}	45678	m_{141}	03689	m_{183}	03456	m_{225}	25678
m_{16}	02689	m_{58}	01259	m_{100}	03469	m_{142}	02478	m_{184}	13456	m_{226}	12579
m_{17}	01368	m_{59}	23479	m_{101}	34789	m_{143}	23457	m_{185}	01568	m_{227}	03458

First 10 Items

m_t	elements
m_1	35689
m_2	01268
m_3	04789
m_4	25689
m_5	02679
m_6	34578
m_7	13459
m_8	01238
m_9	12347
m_{10}	01579

Listed by the indices i of q_i

$$m_1 = \{q_3, q_5, q_6, q_8, q_9\}$$

- Take collections of the members in this order

$M_1 = \{m_1\}$

$M_2 = \{m_1, m_2\}$

…

$M_{252} = \{m_1, m_2, \ldots, m_{252}\}$

- For each point q in A, count how many members of each M_t cover A
- Normalize the count by size of M_t, obtain

$$Y(q,M_t) = \frac{1}{t} \sum_{k=1}^{t} C_{mk}(q) = Prob_M (q \in m \mid m \in M_t)$$

where $C_m(q) = 1$ iff $q \in m$

	$N(M_t, q)$									
M_t	q_0	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9
M_1	0	0	0	1	0	1	1	0	1	1
M_2	1	1	1	1	0	1	2	0	2	1
M_3	2	1	1	1	1	1	2	1	3	2
M_4	2	1	2	1	1	2	3	1	4	3
M_5	3	1	3	1	1	2	4	2	4	4
M_6	3	1	3	2	2	3	4	3	5	4
M_7	3	2	3	3	3	4	4	3	5	5
M_8	4	3	4	4	3	4	4	3	6	5
M_9	4	4	5	5	4	4	4	4	6	5
M_{10}	5	5	5	5	4	5	4	5	6	6

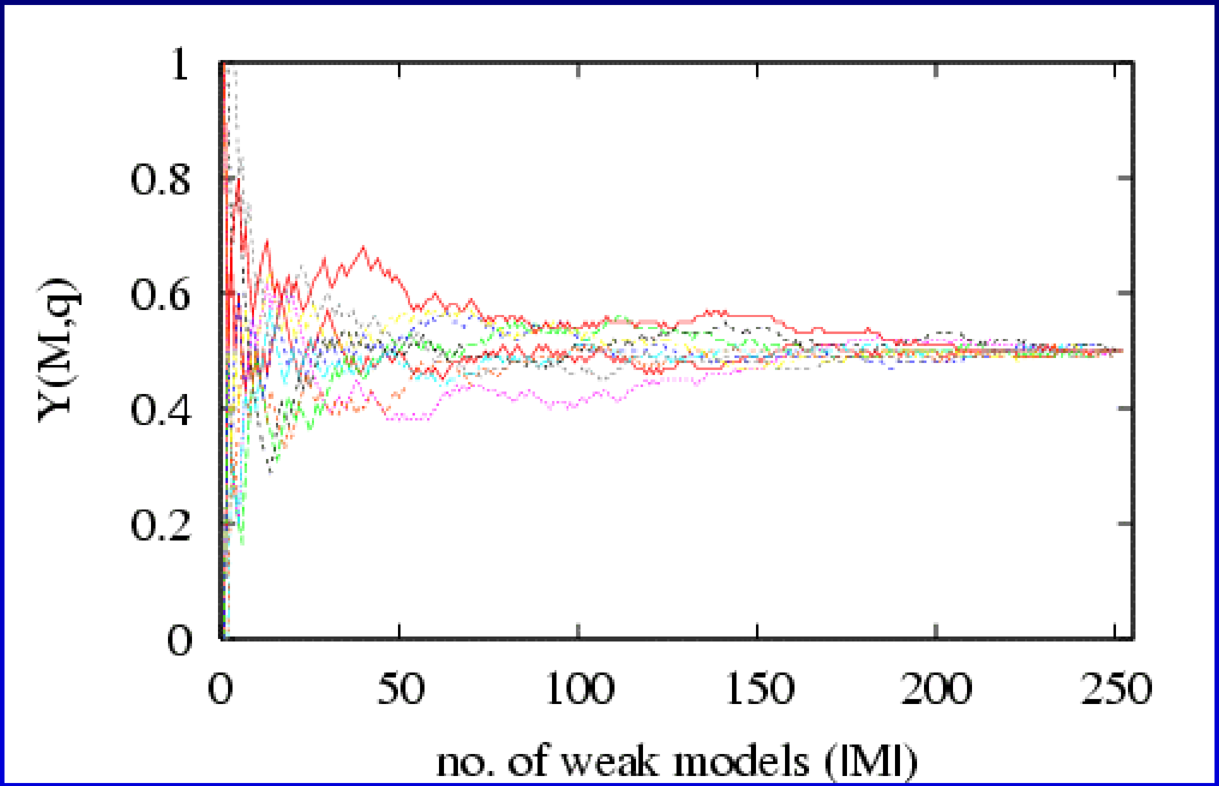
$Y(M_t, q)$									
q_0	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9
0.00	0.00	0.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00
0.50	0.50	0.50	0.50	0.00	0.50	1.00	0.00	1.00	0.50
0.67	0.33	0.33	0.33	0.33	0.33	0.67	0.33	1.00	0.67
0.50	0.25	0.50	0.25	0.25	0.50	0.75	0.25	1.00	0.75
0.60	0.20	0.60	0.20	0.20	0.40	0.80	0.40	0.80	0.80
0.50	0.17	0.50	0.33	0.33	0.50	0.67	0.50	0.83	0.67
0.43	0.29	0.43	0.43	0.43	0.57	0.57	0.43	0.71	0.71
0.50	0.38	0.50	0.50	0.38	0.50	0.50	0.38	0.75	0.62
0.44	0.44	0.56	0.56	0.44	0.44	0.44	0.44	0.67	0.56
0.50	0.50	0.50	0.50	0.40	0.50	0.40	0.50	0.60	0.60

The Y table continues …

0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
0.50	0.49	0.51	0.50	0.50	0.51	0.50	0.50	0.49	0.50
0.50	0.50	0.51	0.50	0.50	0.51	0.50	0.50	0.49	0.50
0.50	0.49	0.51	0.50	0.50	0.51	0.51	0.49	0.49	0.50
0.50	0.50	0.51	0.50	0.50	0.50	0.51	0.49	0.49	0.50
0.50	0.49	0.51	0.50	0.51	0.50	0.51	0.49	0.49	0.50
0.50	0.49	0.51	0.50	0.50	0.50	0.51	0.50	0.49	0.50
0.49	0.49	0.51	0.50	0.50	0.50	0.51	0.49	0.49	0.50
0.50	0.49	0.51	0.50	0.50	0.50	0.51	0.50	0.49	0.50
0.50	0.49	0.51	0.50	0.50	0.50	0.51	0.49	0.49	0.50
0.50	0.50	0.50	0.50	0.50	0.50	0.51	0.50	0.50	0.50
0.50	0.50	0.50	0.49	0.50	0.50	0.51	0.50	0.50	0.50
0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
0.50	0.50	0.50	0.50	0.50	0.50	0.51	0.49	0.49	0.50
0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50

As t goes to 252, Y values become …

- Trace the value of $Y(q, M_t)$ for each q as t increases



- Values of Y converge to 0.5
- They are very close to 0.5 far before t=252

- When t is large, we have

$$Y(q,M_t) = Prob_M\,(q \in m \mid m \in M_t)$$
$$= 0.5$$
$$= Prob_F\,(q \in m \mid q \in A)$$

- We have a symmetry of probabilities in two different spaces M and F
- This is due to the uniform coverage of M_t on A i.e., any two points in A are covered by the same number of models in M_t

Two-class discrimination

- Label points in A with 2 classes:

q0	q1	q2	q3	q4	q5	q6	q7	q8	q9

TR_1 = {q0, q1, q2, q7, q8}

TR_2 = {q3, q4, q5, q6, q9}

- Calculate a rating of each model m for each class:

$$r_1 = \mathrm{Prob}_F\,(q \in m \mid q \in TR_1)$$

$$r_2 = \mathrm{Prob}_F\,(q \in m \mid q \in TR_2)$$

Enriched Models

- Ratings r_1 and r_2 describe how well m is in capturing classes c_1 and c_2 as observed with TR_1 and TR_2

  $r_1(m) = Prob_F(q \in m \mid q \in TR_1)$

  $r_2(m) = Prob_F(q \in m \mid q \in TR_2)$

q0	q1	q2	q3	q4	q5	q6	q7	q8	q9

e.g. m = {q1, q2, q6, q8, q9}

$r_1(m) = 3/5$

$r_2(m) = 2/5$

enrichment degree $d_{12}(m) = r_1(m) - r_2(m) = 0.2$

The Discriminant

- Recall $C_m(q) = 1$ iff $q \in m$
- Define

$$X_{12}(q,m) = \frac{C_m(q) - r_2(m)}{r_1(m) - r_2(m)}$$

- Define a discriminant

$$Y_{12}(q,M_t) = \frac{1}{t} \sum_{k=1}^{t} X_{12}(q,m_k)$$

M_t	m_t	r_1	r_2	$r_1 - r_2$	$X_{12}(q, m_t)$ if $q \in m_t$	$X_{12}(q, m_t)$ if $q \notin m_t$
M_1	m_1	0.20	0.80	-0.60	-0.33	1.33
M_2	m_2	0.80	0.20	0.60	1.33	-0.33
M_3	m_3	0.60	0.40	0.20	3.00	-2.00
M_4	m_4	0.40	0.60	-0.20	-2.00	3.00
M_5	m_5	0.60	0.40	0.20	3.00	-2.00
M_6	m_6	0.40	0.60	-0.20	-2.00	3.00
M_7	m_7	0.20	0.80	-0.60	-0.33	1.33
M_8	m_8	0.80	0.20	0.60	1.33	-0.33
M_9	m_9	0.60	0.40	0.20	3.00	-2.00
M_{10}	m_{10}	0.60	0.40	0.20	3.00	-2.00

$Y_{12}(q, M_t)$									
q_0	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9
1.33	1.33	1.33	-0.33	1.33	-0.33	-0.33	1.33	-0.33	-0.33
1.33	1.33	1.33	-0.33	0.50	-0.33	0.50	0.50	0.50	-0.33
1.89	0.22	0.22	-0.89	1.33	-0.89	-0.33	1.33	1.33	0.78
2.17	0.92	-0.33	0.08	1.75	-1.17	-0.75	1.75	0.50	0.08
2.33	0.33	0.33	-0.33	1.00	-1.33	0.00	2.00	0.00	0.67
2.44	0.78	0.78	-0.61	0.50	-1.44	0.50	1.33	-0.33	1.06
2.28	0.62	0.86	-0.57	0.38	-1.28	0.62	1.33	-0.10	0.86
2.17	0.71	0.92	-0.33	0.29	-1.17	0.50	1.12	0.08	0.71
1.70	0.96	1.15	0.04	0.59	-1.26	0.22	1.33	-0.15	0.41
1.83	1.17	0.83	-0.17	0.33	-0.83	0.00	1.50	-0.33	0.67

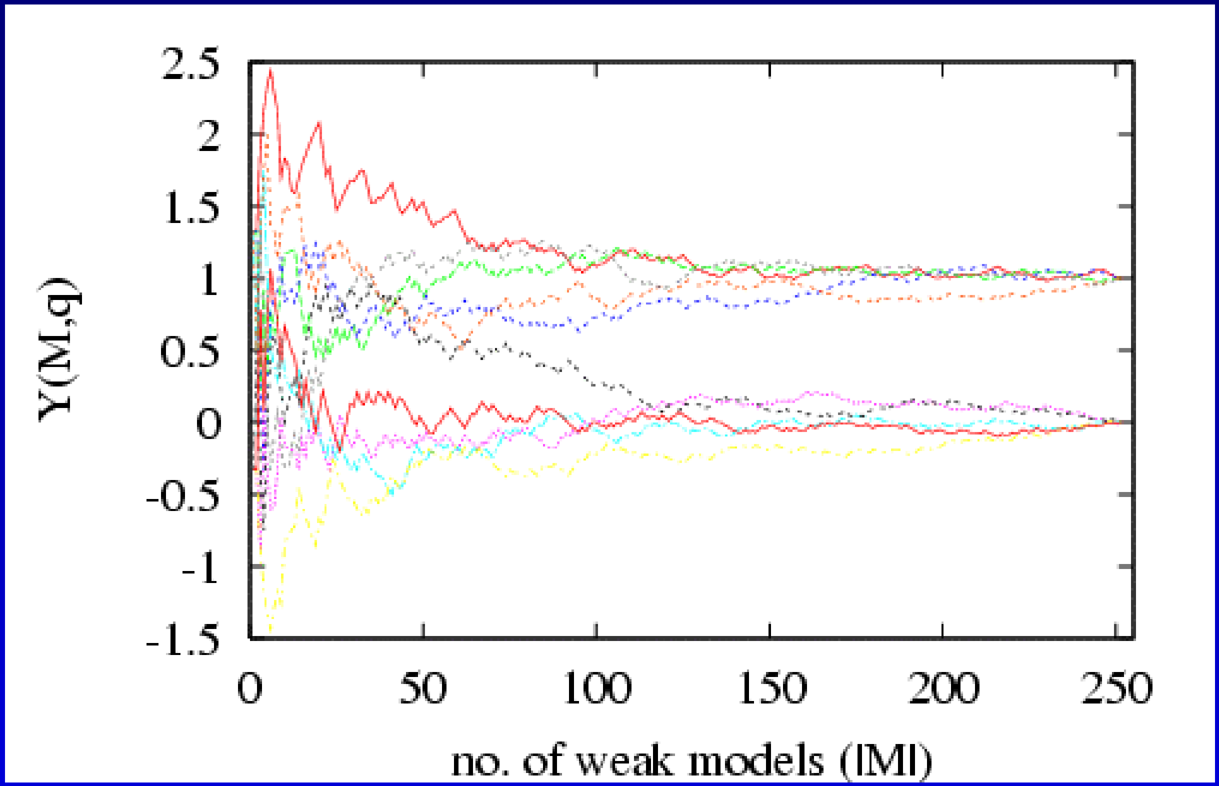
The Y table continues …

1.01	1.04	1.04	0.09	-0.04	-0.05	0.06	0.92	0.99	-0.05
1.01	1.04	1.04	0.09	-0.04	-0.05	0.05	0.92	0.99	-0.05
1.00	1.05	1.05	0.08	-0.05	-0.04	0.05	0.91	1.00	-0.04
1.01	1.03	1.03	0.07	-0.03	-0.05	0.06	0.92	1.01	-0.05
1.01	1.02	1.04	0.06	-0.04	-0.03	0.07	0.93	0.99	-0.06
1.02	1.01	1.03	0.06	-0.03	-0.04	0.06	0.94	1.00	-0.04
1.01	1.02	1.02	0.07	-0.04	-0.03	0.05	0.95	1.01	-0.05
1.02	1.00	1.02	0.06	-0.05	-0.02	0.04	0.96	1.00	-0.04
1.03	0.99	1.03	0.05	-0.03	-0.03	0.04	0.96	0.98	-0.03
1.03	1.00	1.04	0.04	-0.02	-0.01	0.03	0.95	0.97	-0.03
1.04	0.99	1.05	0.03	-0.01	-0.02	0.02	0.96	0.96	-0.02
1.05	0.98	1.06	0.03	0.00	-0.03	0.03	0.97	0.95	-0.03
1.06	0.98	1.04	0.02	0.00	-0.02	0.02	0.96	0.96	-0.02
1.06	0.98	1.04	0.02	0.00	-0.02	0.02	0.96	0.96	-0.02
1.05	0.99	1.03	0.01	-0.01	-0.01	0.01	0.97	0.97	-0.01
1.03	0.98	1.03	0.00	-0.01	0.01	0.03	0.97	0.97	-0.01
1.02	0.99	1.02	-0.01	0.00	0.00	0.02	0.98	0.98	0.00
1.01	1.00	1.01	-0.02	0.01	0.01	0.01	0.99	0.99	-0.01
1.00	1.00	1.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00
1.00	1.00	1.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00

q0 q1 q2 q3 q4 q5 q6 q7 q8 q9

As t goes to 252, Y values become …

- Trace the value of $Y(q, M_t)$ for each q as t increases



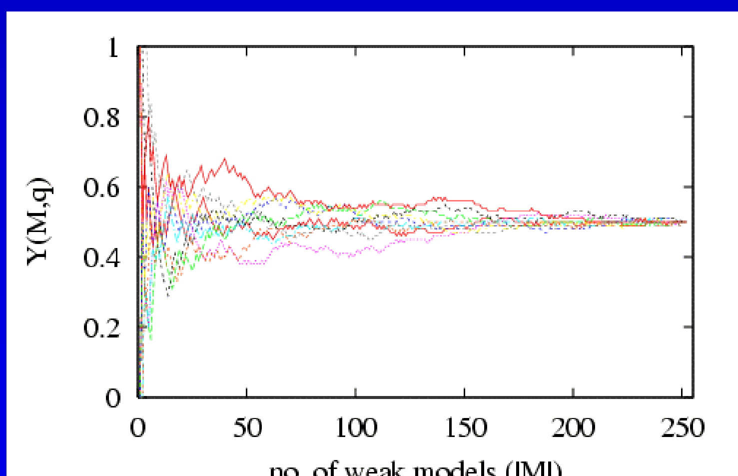
- Values of Y converge to 1 or 0 (1 for TR_1, 0 for TR_2)
- They are very close to 1 or 0 far before t=252

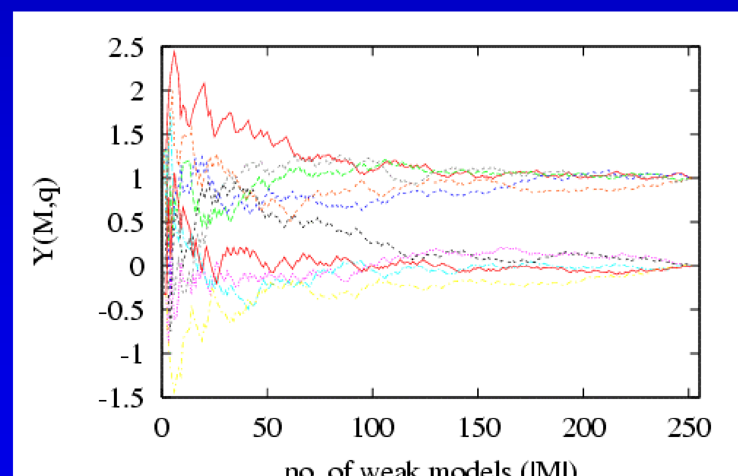
Why?

$$X_{12}(q,m) = C_m(q)$$

$$X_{12}(q,m) = \frac{C_m(q) - r_2(m)}{r_1(m) - r_2(m)}$$

$$Y_{12}(q,M_t) = \frac{1}{t} \sum_{k=1}^{t} X_{12}(q,m_k)$$





Profile of Coverage

- Find the fraction of models of each rating that cover a fixed point q

$$f_{M_t, r1, TR1}(q) \quad \text{and} \quad f_{M_t, r2, TR2}(q)$$

- Since M_t is expanded in a uniform way, as t increases, for all x,

$$f_{M_t, x, TRi}(q) \rightarrow x$$

Ratings of m in M_t

no. of points from TR_1	0	1	2	3	4	5
no. of points from TR_2	5	4	3	2	1	0
r_1	0.0	0.2	0.4	0.6	0.8	1.0
r_2	1.0	0.8	0.6	0.4	0.2	0.0

We have models of 6 different "types"

Profile of Coverage of q_0 at t=10

q_0 at $t = 10$ is only covered by 5 models $(m_2, m_3, m_5, m_8, m_{10})$ in M_{10},

m_3, m_5, m_{10}

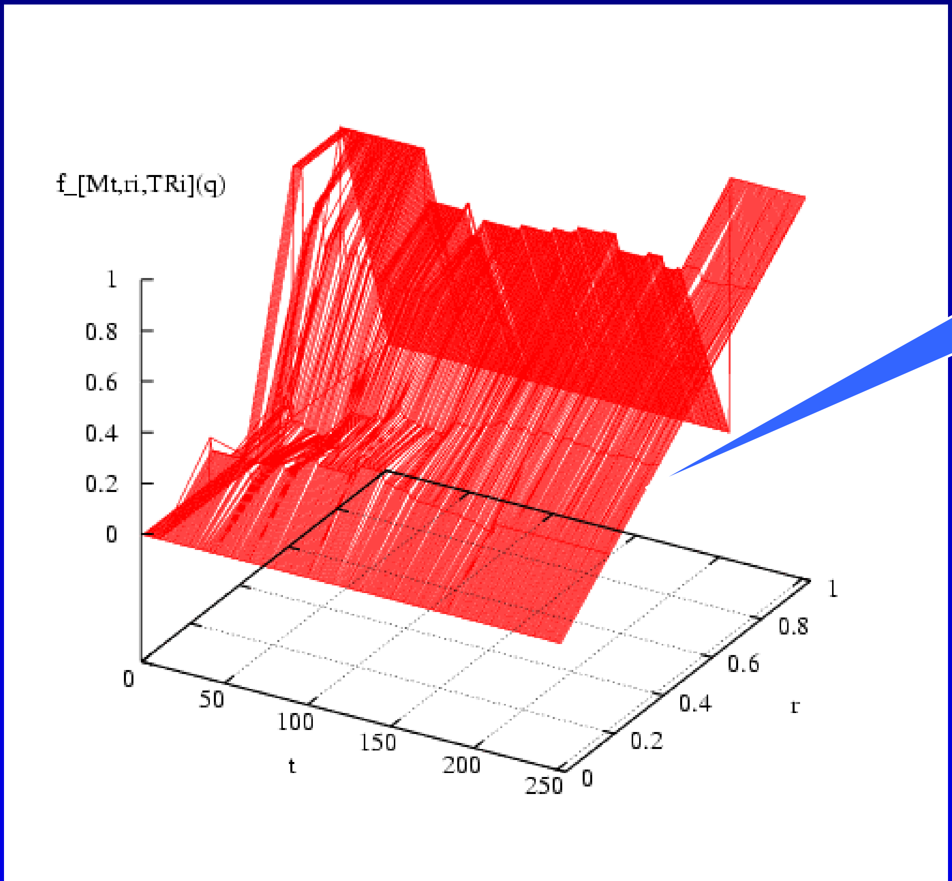
m_2, m_8

r_1	0.0	0.2	0.4	0.6	0.8	1.0
no. of models in M_{10} with r_1	0	2	2	4	2	0
$N_{M_{10},r_1,TR_1}(q_0)$	0	0	0	3	2	0
$f_{M_{10},r_1,TR_1}(q_0)$	0	0	0	0.75	1.0	0
r_2	0.0	0.2	0.4	0.6	0.8	1.0
no. of models in M_{10} with r_2	0	2	4	2	2	0
$N_{M_{10},r_2,TR_2}(q_0)$	0	2	3	0	0	0
$f_{M_{10},r_2,TR_2}(q_0)$	0	1.0	0.75	0	0	0

Ratings of m
(repeated for reference)

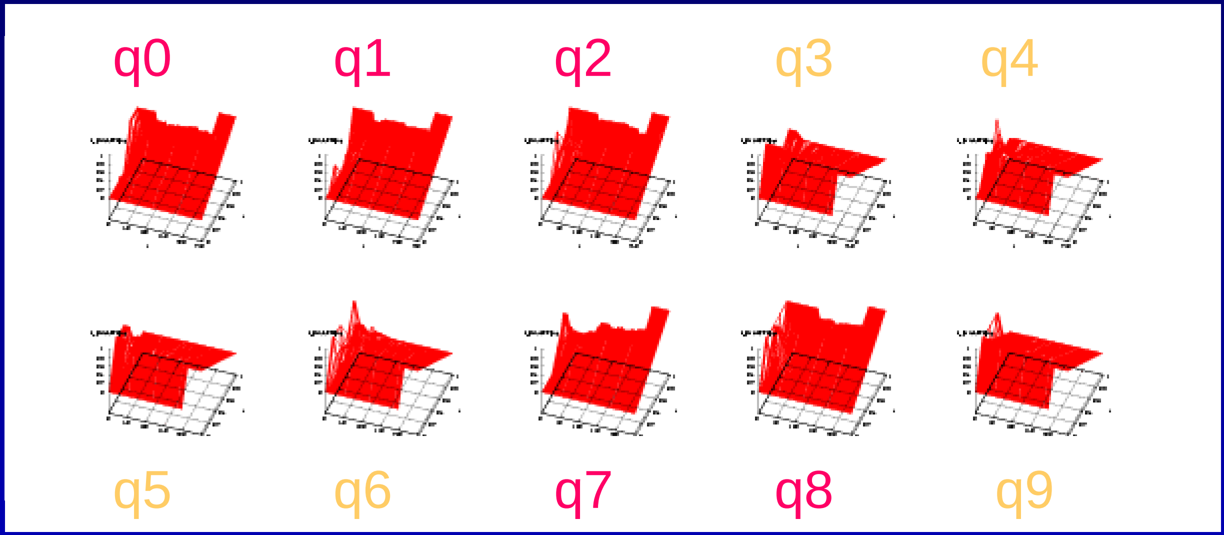
M_t	m_t	r_1	r_2	$r_1 - r_2$	$X_{12}(q, m_t)$ if $q \in m_t$	$X_{12}(q, m_t)$ if $q \notin m_t$
M_1	m_1	0.20	0.80	-0.60	-0.33	1.33
M_2	m_2	0.80	0.20	0.60	1.33	-0.33
M_3	m_3	0.60	0.40	0.20	3.00	-2.00
M_4	m_4	0.40	0.60	-0.20	-2.00	3.00
M_5	m_5	0.60	0.40	0.20	3.00	-2.00
M_6	m_6	0.40	0.60	-0.20	-2.00	3.00
M_7	m_7	0.20	0.80	-0.60	-0.33	1.33
M_8	m_8	0.80	0.20	0.60	1.33	-0.33
M_9	m_9	0.60	0.40	0.20	3.00	-2.00
M_{10}	m_{10}	0.60	0.40	0.20	3.00	-2.00

Profile of Coverage for a fixed point q in TR_i

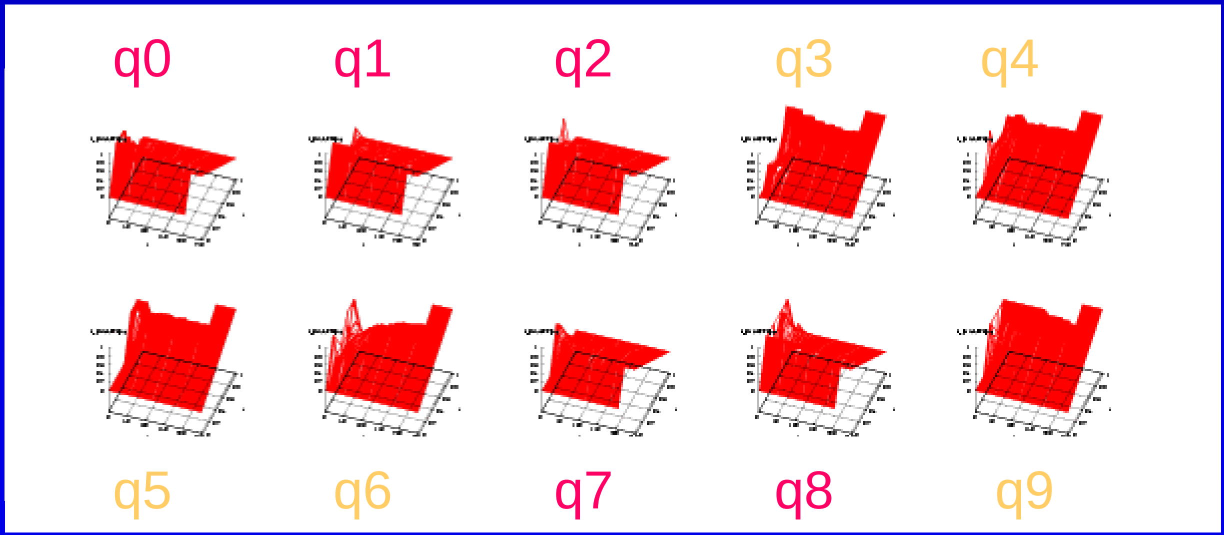


$f_{Mt,\, x,\, TRi}(q) = x$

Profile of coverage as a function of r1: $f_{Mt,\ r1,\ TR1}(q)$



Profile of coverage as a function of r2: $f_{Mt,\ r2,\ TR2}(q)$



Decomposition of Y

$$Y_{12}(q,M_t) = \frac{t_{0.0}}{t}[\frac{1}{t_{0.0}}\sum_{k_{0.0}=1}^{t_{0.0}} X_{12}(q,m_{k_{0.0}})] + \frac{t_{0.2}}{t}[\frac{1}{t_{0.2}}\sum_{k_{0.2}=1}^{t_{0.2}} X_{12}(q,m_{k_{0.2}})] + \frac{t_{0.4}}{t}[\frac{1}{t_{0.4}}\sum_{k_{0.4}=1}^{t_{0.4}} X_{12}(q,m_{k_{0.4}})] + \frac{t_{0.6}}{t}[\frac{1}{t_{0.6}}\sum_{k_{0.6}=1}^{t_{0.6}} X_{12}(q,m_{k_{0.6}})] + \frac{t_{0.8}}{t}[\frac{1}{t_{0.8}}\sum_{k_{0.8}=1}^{t_{0.8}} X_{12}(q,m_{k_{0.8}})] + \frac{t_{1.0}}{t}[\frac{1}{t_{1.0}}\sum_{k_{1.0}=1}^{t_{1.0}} X_{12}(q,m_{k_{1.0}})].$$

$$E(X_{12}(q,m_x)) = E(\frac{C_{m_x}(q)-y}{x-y}) = \frac{E(C_{m_x}(q))-y}{x-y} = \frac{x-y}{x-y} = 1.$$

$$Y_{12}(q,M_t) = \frac{t_{0.0}+t_{0.2}+t_{0.4}+t_{0.6}+t_{0.8}+t_{1.0}}{t} = 1.$$

Duality due to uniformity

Can be shown to be 0 for $q \in TR_2$ in a similar way.

Projectability of Models

- If F has more than the training points q:

q0,p0	q1,p1	q2,p2	q3,p3	q4,p4	q5,p5	q6,p6	q7,p7	q8,p8	q9,p9

- If the models m are larger – not only including the q points but also their neighboring p, the same discriminant Y_{12} can be used to classify the p points
- The points p and q are M_t-indiscernible

Example Definition of a Model

$$
\begin{aligned}
m_1 = \quad & \{q \mid \frac{v(q_2)+v(q_3)}{2} < v(q) < \frac{v(q_3)+v(q_4)}{2}\} \cup \\
& \{q \mid \frac{v(q_4)+v(q_5)}{2} < v(q) < \frac{v(q_6)+v(q_7)}{2}\} \cup \\
& \{q \mid \frac{v(q_7)+v(q_8)}{2} < v(q)\}.
\end{aligned}
$$

q0,p0 q1,p1 q2,p2 $\boxed{\text{q3,p3}}$ q4,p4 $\boxed{\text{q5,p5 q6,p6}}$ q7,p7 $\boxed{\text{q8,p8 q9,p9}}$

$\cup$ $\cup$

Points within m are m-indiscernible

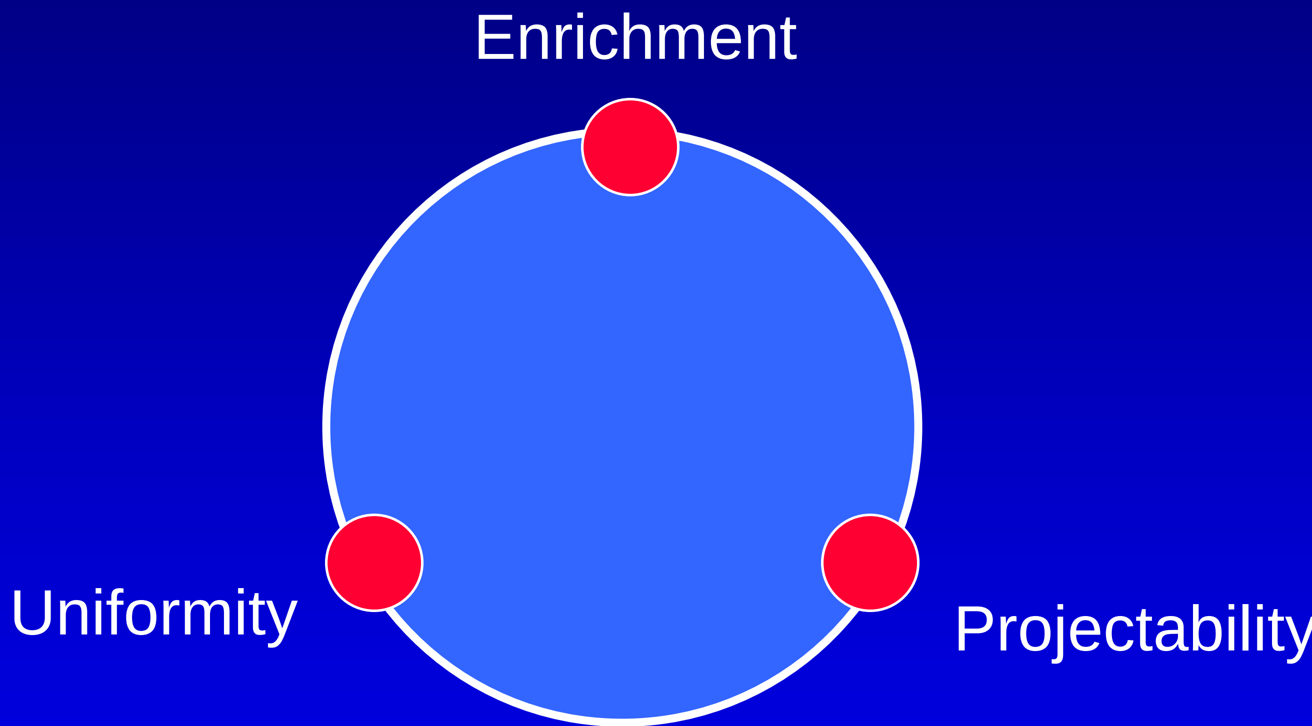
Model Size and Projectability

- Points within the same model share the same interpretation
- Larger models -> more stable the ratings are w.r.t. sampling differences -> more similar classification between TR and TE
- Tradeoff with easiness to achieve uniformity

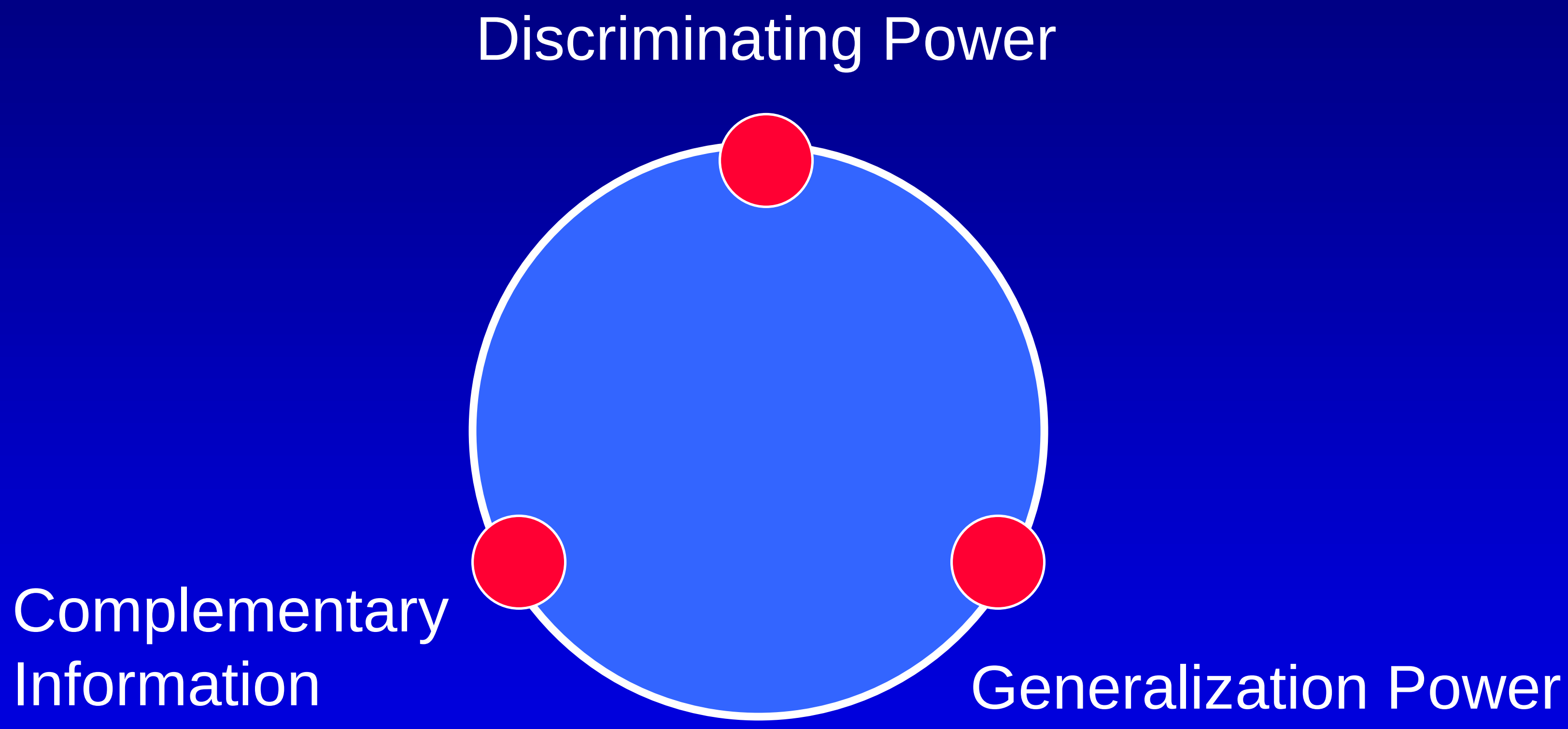
Enrichment and Convergence

- Larger enrichment degree -> smaller variance in X -> Y converges faster
- Models with large enrichment degree are more difficult to obtain
- Thus more difficult to achieve uniformity

The 3-way Tension



The 3-way Tension



Review
Key Concepts and Tools in SD

- Set-theoretic abstraction
- Symmetry of probabilities in model or feature spaces
- Enrichment / Uniformity / Projectability
- Convergence of discriminant by the law of large numbers

Weak Models

- A weak model *m* is a subset of the feature space F

$$m \in 2^{F}$$

- It contains points sharing the same interpretation

- It should have a simple form, easy-to-compute membership function
- It should have a minimum size
- It may be cheaply produced by a stochastic process

Enriched Weak Models

- Rate a weak model by how well it captures points of each class

$$r(m, A) = \frac{|m \cap A|}{|A|}$$

- Degree of enrichment is how much the model is biased between two classes

$$d_{ij}(m) = r(m, TR_i) - r(m, TR_j)$$

- A weak model is enriched if $d_{ij}(m) > 0$

The Stochastic Discriminant

- For point *q* and model *m*, classes *i* and *j*:

$$X_{ij}(q,m) = \begin{cases} 2\left(\dfrac{C(q,m) - r(m,TR_j)}{r(m,TR_i) - r(m,TR_j)}\right) - 1 & \text{if } r(m,TR_i) \neq r(m,TR_j) \\ 0 & \text{if } r(m,TR_i) = r(m,TR_j) \end{cases}$$

$$\text{where} \quad C(q,m) = 1 \Leftrightarrow q \in m$$

- For a collection of *t* weak models $\mathbf{M}^t = \{m_1, m_2, \ldots, m_t\}$:

$$Y_{ij}^{t}(q,\mathbf{M}^t) = \frac{\sum_{k=1}^{t} X_{ij}(q,m_k)}{t}$$

A Uniform Cover

- The collection of models should cover the space uniformly – any two points of the same class should fall equally likely in models of a specific rating

- **M** is *A*-uniform if for every $x = r(m,A)$ such that $\mathbf{M}_{x,A}$ is nonempty, and for any two points *p*,*q* in *A*:

$$P_{2^F}(p \in m \mid m \in \mathbf{M}_{x,A}) = P_{2^F}(q \in m \mid m \in \mathbf{M}_{x,A})$$

- We need a collection of models that is both TR_i-uniform and TR_j -uniform.

Symmetry between Probabilities w.r.t. F and 2^F

- If **M** is *A*-uniform, then for all *q* in *A*,

$$\boxed{P_{2^F}(q \in m \mid m \in \mathbf{M}_{x,A}) = x}$$

- But by definition, for all *m* in $\mathbf{M}_{x,A}$

$$\boxed{P_F(q \in m \mid q \in A) = x}$$

Duality between Distributions of

$\lambda q C(q,m)$ and $\lambda m C(q,m)$

$\lambda q \left[Y_{ij}\,(q, \tilde{\mathsf{M}}^{t}) \right]$ and $\lambda \mathsf{M}^{t} \left[Y_{ij}(q, \mathsf{M}^{t}) \right]$

Convergence of the Discriminant

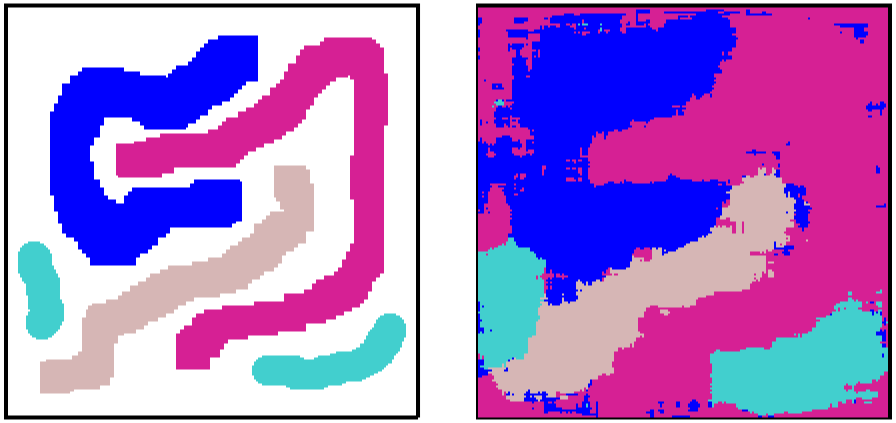
- With enriched weak models, values of X_{ij} are distributed around

  +1 for points of class i and

  –1 for points of class j

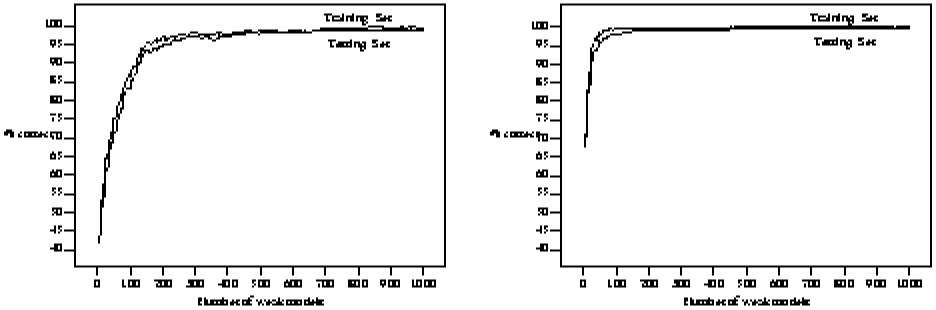
- Y_{ij} converges to $E(x_{ij})$ with variance 1/t that of X_{ij} according to the law of large numbers

- Classifier obtainable within time proportional to 1/u (u = upper bound on error) and $1/d_{ij}^2$ (d_{ij} = enrichment degree)



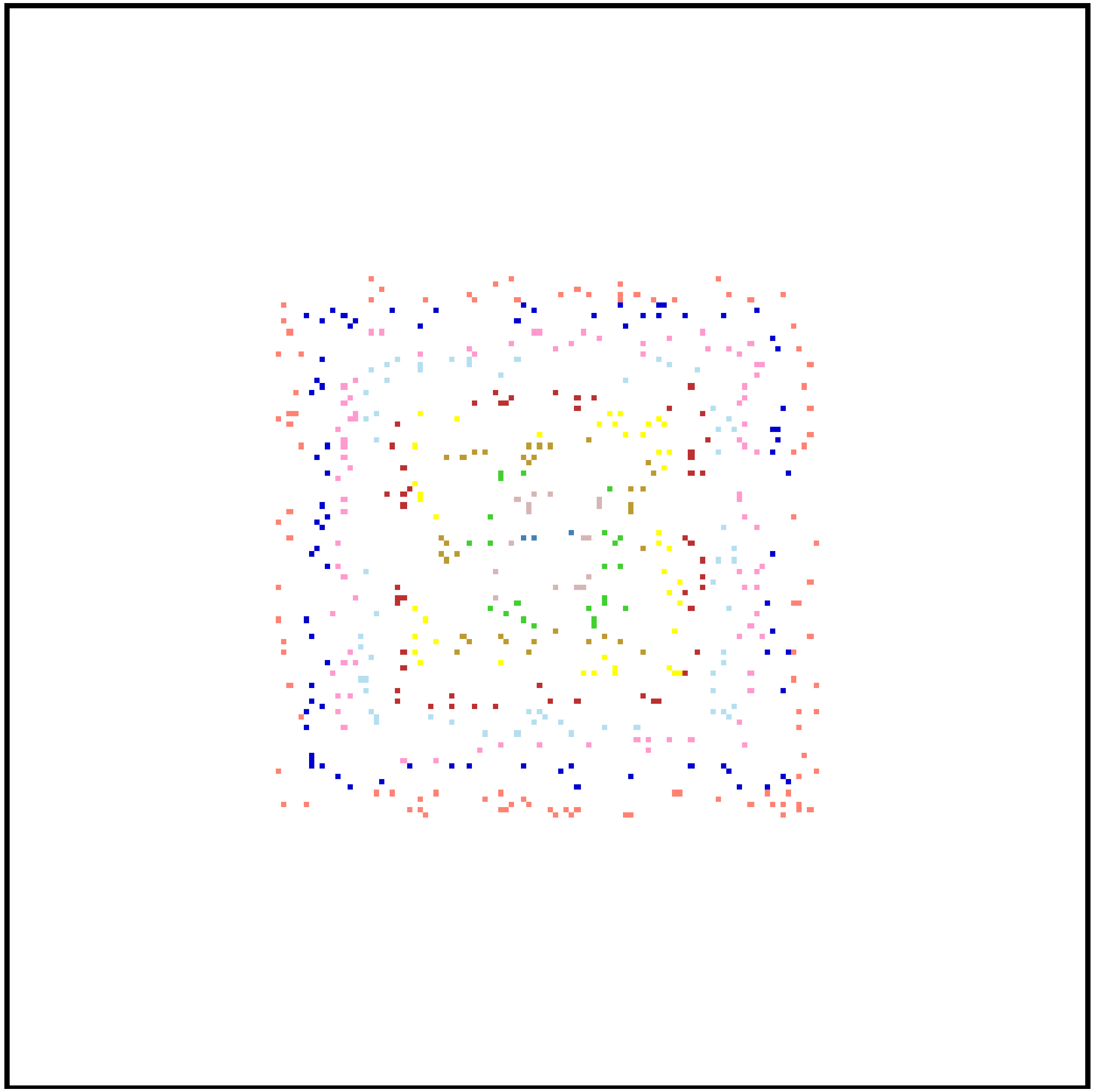
(a) (b)

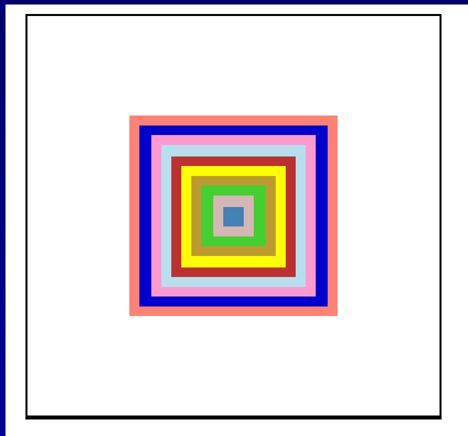
Figure 2. (a) True distributions and (b) classification with 500 weak models.

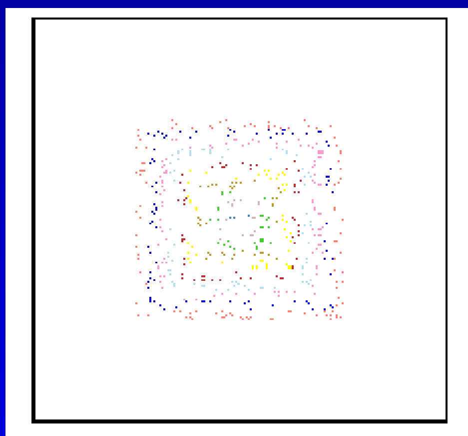


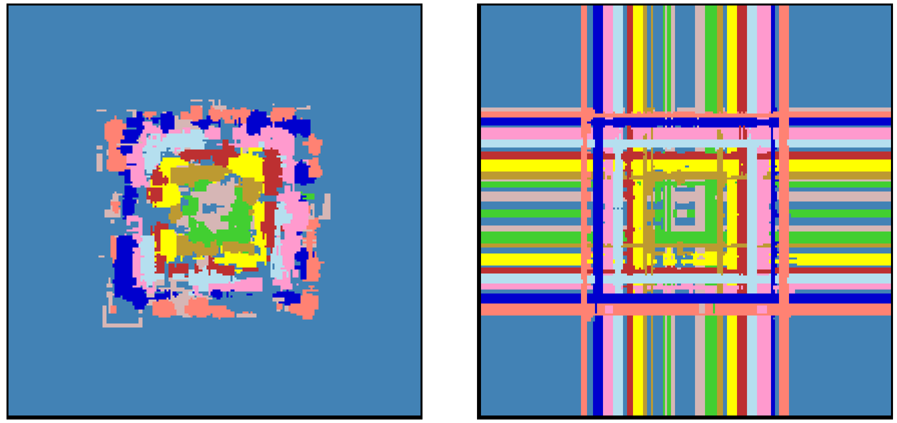
(a) (b)

Figure 3. Accuracy (a) without and (b) with uniformity promotion.



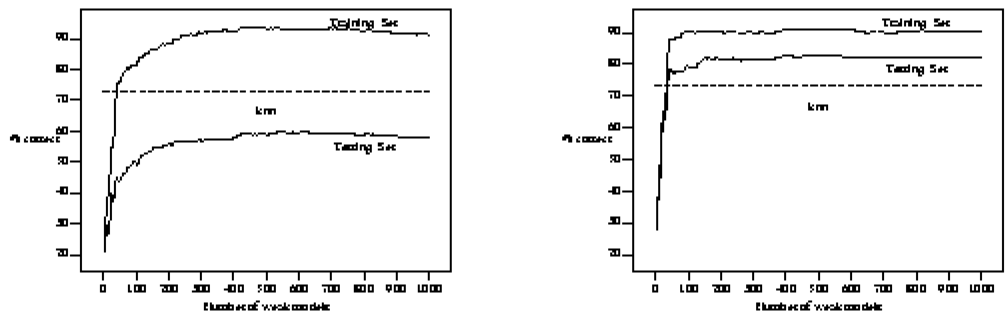






(a) (b)

Figure 5. Classification of space with 100 of (a) type 1 and (b) type 2 models.



(a) (b)

Figure 6. Accuracies with (a) type 1 and (b) type 2 models.

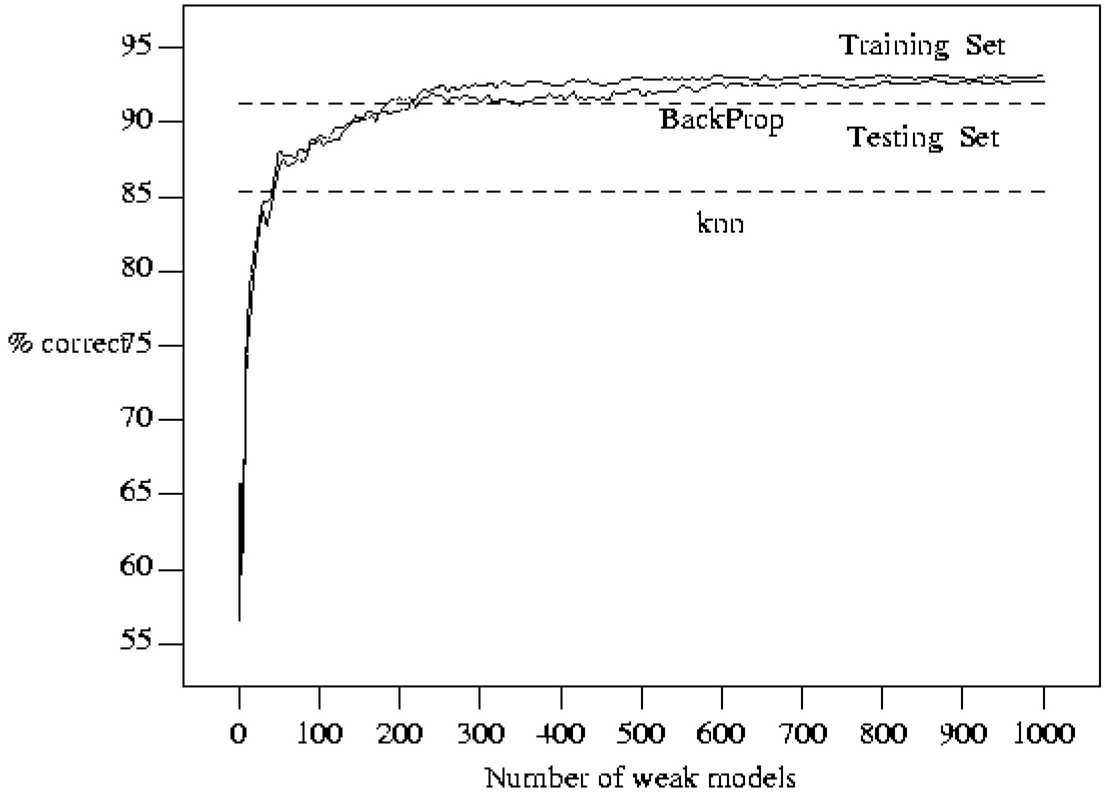


Figure 7. Accuracies on the DNA data.

Open Problems in Stochastic Discrimination

- Algorithm for uniformity enforcement
- Desirable form of weak models

  Fewer, more sophisticated classifiers?
- Other ways to address the 3-way trade-off

  Enrichment / Uniformity / Projectability

Random Decision Forest

- [Ho 1995, 1998]
- A structured way to create models

  fully split a tree, use leaves as models

- Perfect enrichment and uniformity for TR
- Promote projectability by subspace projection

Compact Distribution Maps

- [Ho & Baird 1993, 1997]
- Another structured way to create models
- Start with projectable models by coarse quantization of feature values
- Seek enrichment and uniformity

Alternative Discriminants

- [Berlind 1994]
- Different discriminants for N-class problems
- Additional condition on symmetry
- Approximate uniformity
- Hierarchy of indiscernibility

Estimates of Classification Accuracies

- [Chen 1997]
- Statistical estimate of classification accuracy under weaker conditions:

  Approximate uniformity

  Approximate indiscernibility

Stochastic Discrimination

- A family of mathematical theories that relate several key concepts in pattern recognition:
  - Discriminative power … enrichment
  - Complementary information … uniformity
  - Generalization power … projectability
- It offers a way to describe complementary behavior of classifiers
- It offers guidelines to design multiple classifier systems

Homework

- Read http://www.cs.bell-labs.com/who/tkh/talks/example.ps
- Reproduce this example and all tables
- Try a different random permutation of $M_{0.5,A}$
- Try changing size of m …
- Try changing size of TR1, TR2, …
  see what happens with X and Y
- Try the ideas on other data …
- Visit http://kappa.math.buffalo.edu/sd
- \