Not so naive Bayesian classification

Geoff Webb

Monash University,

Melbourne, Australia

http://www.csse.monash.edu.au/∼webb

Overview

- Probability estimation provides a theoretically well-founded approach to classification
- Naive Bayes is efficient but suffers the attribute independence assumption
- LBR and TAN temper the naivety of naive Bayes
  - accurate, but high computational complexity
- AODE
  - relaxes the attribute independence assumption
  - increases prediction accuracy
  - retains much of naive Bayes' efficiency
  - attains LBR & TAN's accuracy with less computation
  - supports incremental, parallel and anytime classification

Classification learning

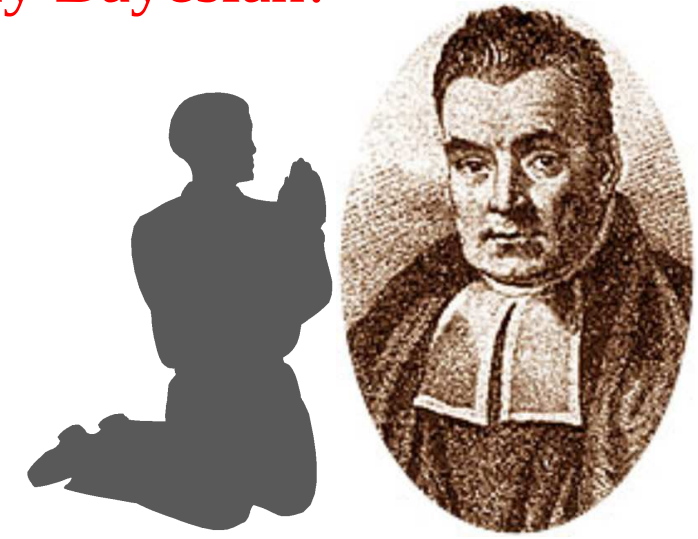
- Given a sample from XY want to select $y \in Y$ for new $\mathbf{x} = \langle x_1, \ldots, x_n \rangle \in X$
  - eg Xs = symptoms, Ys = diseases
- Error minimized by $argmax_y(P(y \mid \langle x_1, \ldots, x_n \rangle))$
  - but do not know probabilities
- Can estimate using
  - $P(W) \approx F(W)$
  - $P(W \mid Z) \approx \frac{F(W,Z)}{F(Z)}$
  - but usually too little data for accurate estimation for $P(\langle x_1, \ldots, x_n \rangle)$ or $P(y \mid \langle x_1, \ldots, x_n \rangle)$

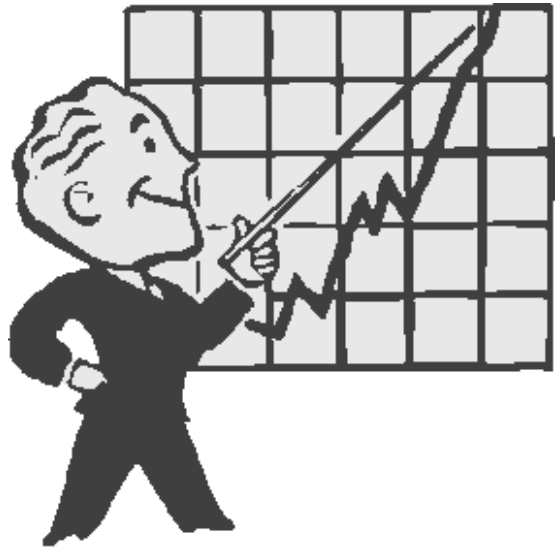
Bayes' theorem

- $P(y \mid \mathbf{x}) = \frac{P(y)P(\mathbf{x} \mid y)}{P(\mathbf{x})}$
- $P(y \mid \mathbf{x}) \propto P(y)P(\mathbf{x} \mid y)$
- can estimate $P(y)$ from data so have replaced estimating $P(y \mid \mathbf{x})$ with estimating $P(\mathbf{x} \mid y)$
- Attribute independence assumption
  - $P(\langle x_1, \ldots, x_n \rangle \mid y) = \prod_{i=1}^{n} P(x_i \mid y)$
  - eg
    $P(temp{=}high, pulse{=}high \mid ill) =$
    $P(temp{=}high \mid ill) \times P(pulse{=}high \mid ill)$

Naive Bayesian Classification

- use Bayes theorem, attribute independence assumption, and estimation of probabilities from data to select most probable class for given x
- simple, efficient, and accurate
- direct theoretical foundation
- can provide probability estimates
- not necessarily Bayesian!





Attribute independence assumption

- Violations of the attribute independence assumption can increase expected error.
- Some violations do not matter (Domingos & Pazzani, 1996).
- Violations that matter are frequent
  - NB is often sub-optimal

Semi-naive Bayesian classification

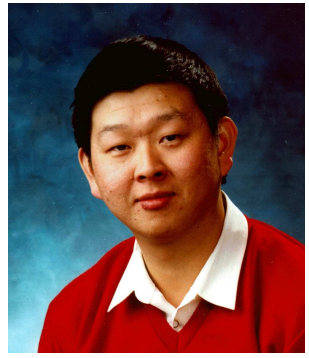
- Kononenko (1991) joins attributes
- Recursive Bayesian classifier (Langley, 1993)
- Selective naive Bayes (Langley & Sage, 1994)
- BSEJ (Pazzani, 1996)
- NBTree (Kohavi, 1996)
- Limited dependence Bayesian classifiers (Sahami, 1996)
- TAN (Friedman, Geiger & Goldszmidt, 1997)
- Adjusted probability NB (Webb & Pazzani, 1998)
- LBR [Lazy Bayesian Rules] (Zheng & Webb, 2000)
- Belief Net Classifiers (Greiner, Su, Shen & Zhou, 2005)
- PDAGs (Acid, de Campos & Castellano, 2005)
- TBMATAN (Cerquides & de Mantaras, 2005)

Tree Augmented Naive Bayes

- All attributes depend on class and at most one other attribute (Friedman, Geiger & Goldszmidt, 1997)
- $P(y \mid \langle x_1, \ldots, x_n \rangle) \propto P(y) \prod_{i=1}^{n} P(x_i \mid parent(x_i) \wedge y)$
- Parent function selected by mutual conditional information
- Keogh & Pazzani (1999) use wrapper to select parent function
  - Computationally intensive but provides considerable decrease in prediction error

LBR

- $P(y \mid \mathbf{x}', \mathbf{x}'') = \frac{P(y \mid \mathbf{x}'')P(\mathbf{x}' \mid y, \mathbf{x}'')}{P(\mathbf{x}' \mid \mathbf{x}'')}$
- $P(y \mid \mathbf{x}', \mathbf{x}'') \propto P(y \mid \mathbf{x}'')P(\mathbf{x}' \mid y, \mathbf{x}'')$
- make $\mathbf{x}'$ and $\mathbf{x}''$ a disjoint partition of $\mathbf{x}$
- defines a space of 2^n formulae all equal to $P(y \mid \mathbf{x})$
- classification task transformed to selection of one of many equivalent formulae for which probabilities can best be estimated from available data
- weakened attribute independence assumption
  - $P(\mathbf{x}' \mid y, \mathbf{x}'') = \prod_{x \in \mathbf{x}'} P(x \mid y, \mathbf{x}'')$
- wrapper used to select formula at classification time
  - lazy learning

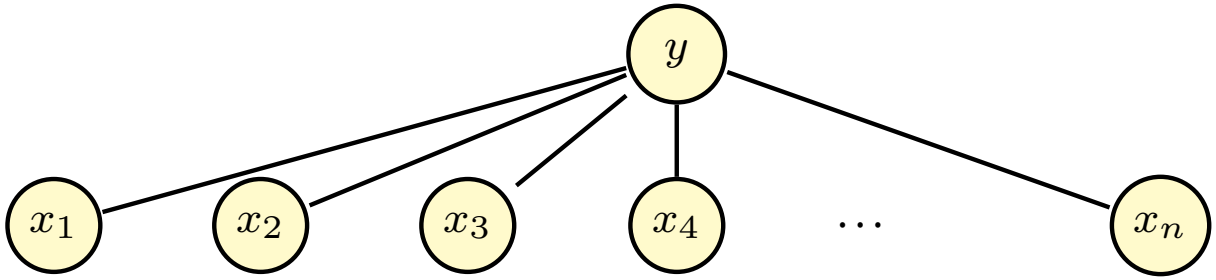


LBR and TAN performance

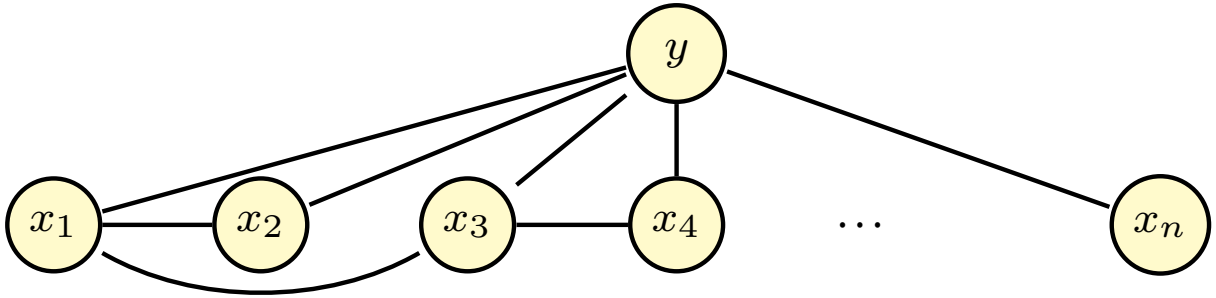
- LBR and TAN reduce the error of NB
  - by reducing bias
  - at cost of small increase in variance.
- For classification from discrete-valued data LBR has comparable error to AdaBoost, and slightly better than bagging
- LBR is very efficient for few test cases per training set
- LBR and TAN have comparable error, but different computational profiles.
- Both LBR and TAN are computationally intensive

Markov net perspective

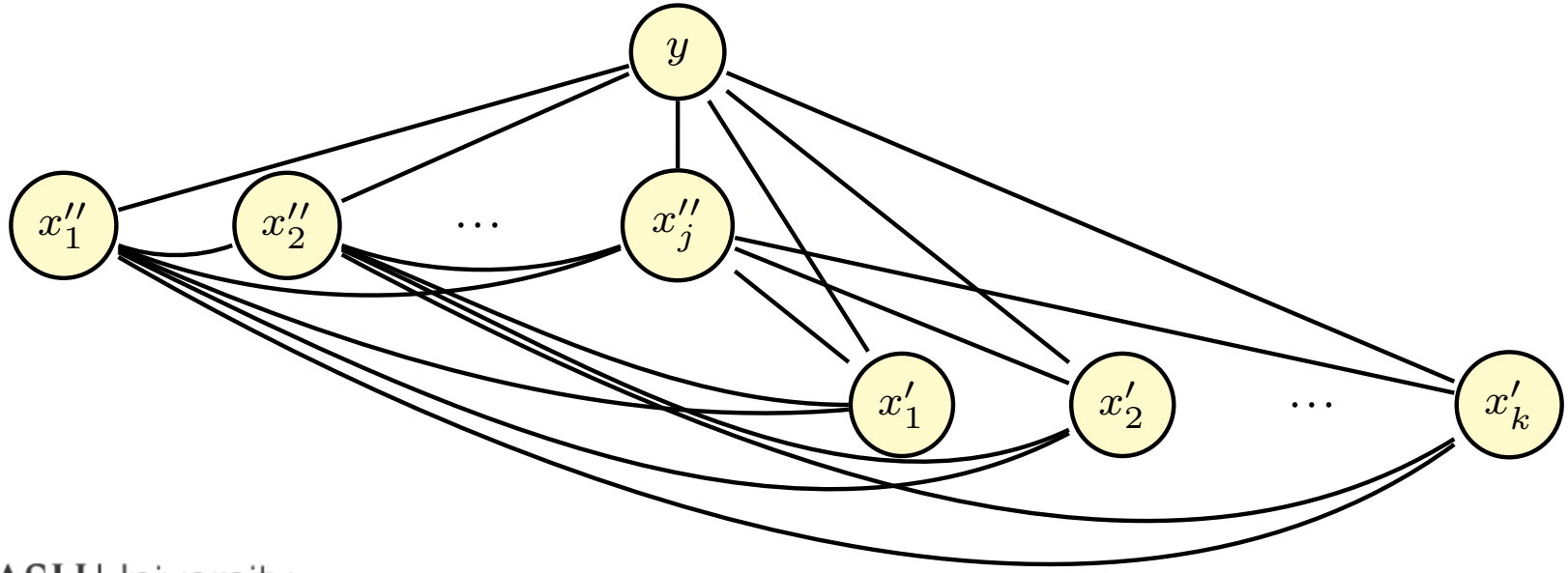
NB:



TAN:



LBR:



Weaker independence assumption

- Both TAN and LBR
  - assume independence between fewer attributes
  - independence only assumed under stronger conditional constraints
- LBR also
  - estimates fewer conditional probabilities
- So long as base probability estimates are accurate, incorrect inter-dependence assumptions should do no harm.
- Risk: base probabilities estimated from less data

Improving LBR and TAN

- Objective
  - Maintain accuracy of LBR and TAN while lowering computation
- Computation results from
  - calculation of conditional probabilities
  - selection of interdependencies
- If allow at most class + k attribute interdependencies per attribute, probabilities can be estimated from an $k + 2$ dimensional lookup table of joint frequencies
  - $P(x_i \mid y, x_j) \approx F[x_i, y, x_j]/F[x_j, y, x_j]$

AODE

- For efficiency, use 3d table, each attribute depends on class and one other attribute
  - in theory can accommodate any pair-wise attribute interdependencies
- For efficiency and to minimize variance, avoid model selection
  - use all interdependencies for which there is sufficient data for probability estimation
- Conflict: cannot represent multiple interdependencies if only one interdependency per attribute
- Solution: average all models that have a single attribute as parent to all others
- Qualification: restrict parents to frequent attribute values

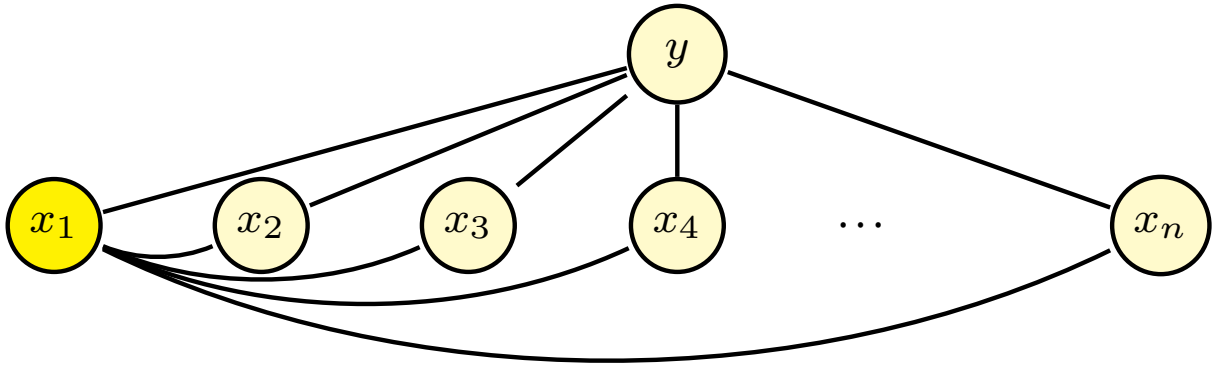
AODE (cont.)

$$P(y \mid \langle x_1, \ldots, x_n \rangle) = \frac{P(y, \langle x_1, \ldots, x_n \rangle)}{P(\langle x_1, \ldots, x_n \rangle)}$$

$$\begin{aligned} P(y, \langle x_1, \ldots, x_n \rangle) &= P(y, x_i) P(\langle x_1, \ldots, x_n \rangle \mid y, x_i) \\ &= \frac{\sum_{i:|x_i|>k} P(y, x_i) P(\langle x_1, \ldots, x_n \rangle \mid y, x_i)}{|\{i : |x_i| > k\}|} \end{aligned}$$

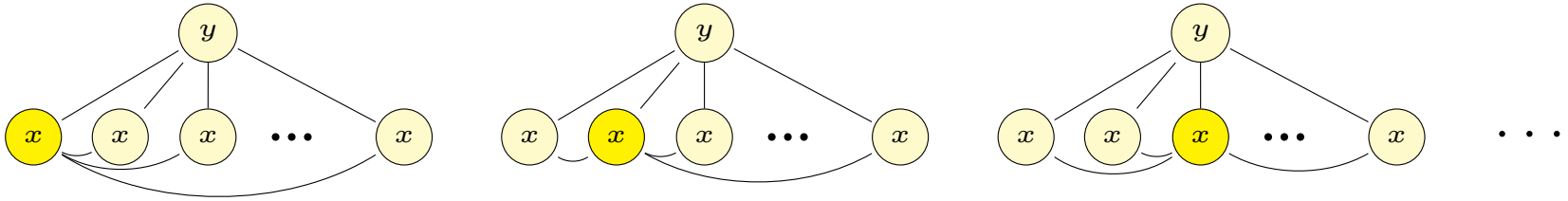
$$P(\langle x_1, \ldots, x_n \rangle \mid y, x_i) \approx \prod_{j=1}^{n} P(x_j | y, x_i)$$

Markov net:



AODE interpretations

- Bayesian average over all dual parent models
  - uniform prior
- Ensemble of all dual parent models



Complexity

alg.	train time	train space	class time	class space
NB	$O(ni)$	$O(nvc)$	$O(nc)$	$O(nvc)$
AODE	$O(n^2 i)$	$O((nv)^2 c)$	$O(n^2 c)$	$O((nv)^2 c)$
TAN	$O(n^3 ci)$	$O((nv)^2 c + ni)$	$O(nc)$	$O(nv^2 c)$
LBR	$O(ni)$	$O(ni)$	$O(n^3 ci)$	$O(ni + nvc)$

n = no. of attributes
v = ave. no. attribute values
c = no. classes
i = no. training instances

Evaluation

- 37 data sets from UCI repository
  - data used in previous related research
  - minus pioneer for which we could not complete computation
- Algorithms implemented in Weka
- NB, AODE, TAN, LBR, J48, boosted J48
- MDL discretisation for NB, AODE, TAN and LBR
- Laplace estimate
- 10-fold cross-validation

Error

Mean error:

AODE	NB	TAN	LBR	J48	Boosted J48
0.209	0.223	0.214	0.212	0.229	0.206

Geometric mean error ratio:

NB	TAN	LBR	J48	Boosted J48
1.104	1.038	1.030	1.187	1.006

Win–draw–loss table with 2-tail *p*:

NB	TAN	LBR	J48	Boosted J48
21-6-10	22-2-13	18-3-16	23-0-14	20-0-17
0.0354	0.0877	0.4321	0.0939	0.3714

Compute time

- Mean training time in seconds

AODE	NB	TAN	LBR	J48	Boosted J48
3.8	3.4	516.9	4.2	26.6	390.4

- Mean testing time in seconds

AODE	NB	TAN	LBR	J48	Boosted J48
1.1	0.2	0.1	15456.1	0.1	0.6

Bias

Mean bias:

AODE	NB	TAN	LBR	J48	Boosted J48
0.148	0.164	0.148	0.145	0.130	0.111

Geometric mean ratio:

NB	TAN	LBR	J48	Boosted J48
1.136	1.005	0.978	0.952	0.741

Win–draw–loss table with 2-tail p:

NB	TAN	LBR	J48	Boosted J48
21-6-10	14-2-21	14-3-20	11-0-26	7-0-30
0.0354	0.1553	0.1958	0.0100	<0.0001

Variance

Mean variance:

AODE	NB	TAN	LBR	J48	Boosted J48
0.060	0.058	0.065	0.066	0.097	0.093

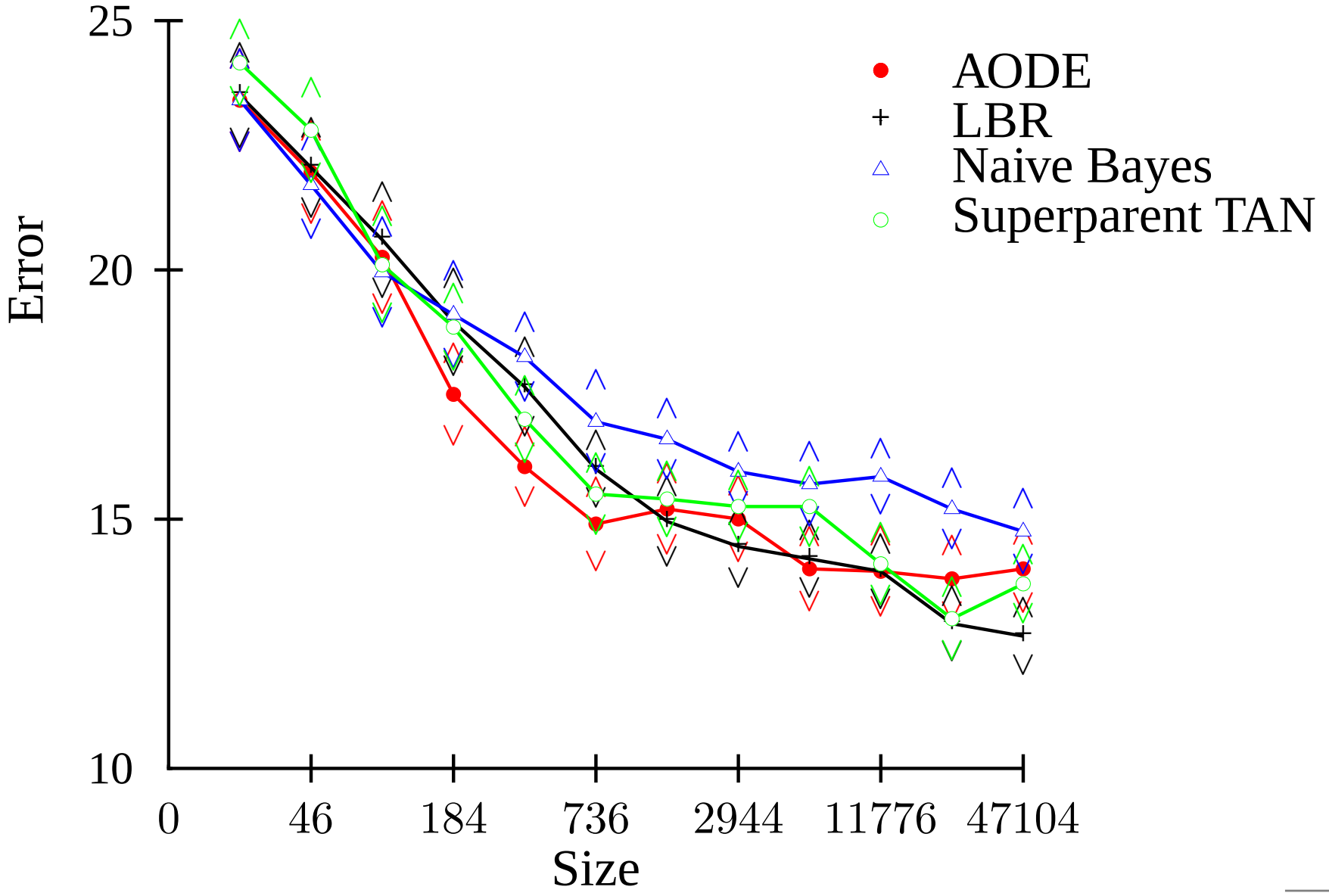
Geometric mean ratio:

NB	TAN	LBR	J48	Boosted J48
0.960	1.096	1.121	1.711	1.680

Win–draw–loss table with 1-tail p:

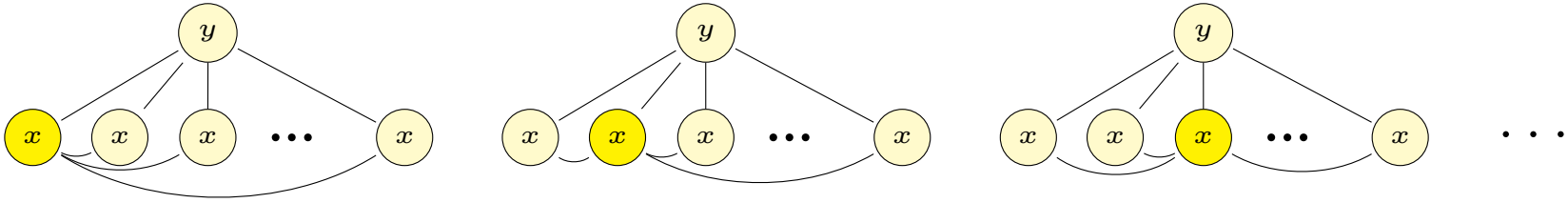
NB	TAN	LBR	J48	Boosted J48
11-6-20	24-1-12	21-3-13	31-0-6	32-0-5
0.0748	0.0326	0.1147	<0.0001	<0.0001

Learning curves for adult

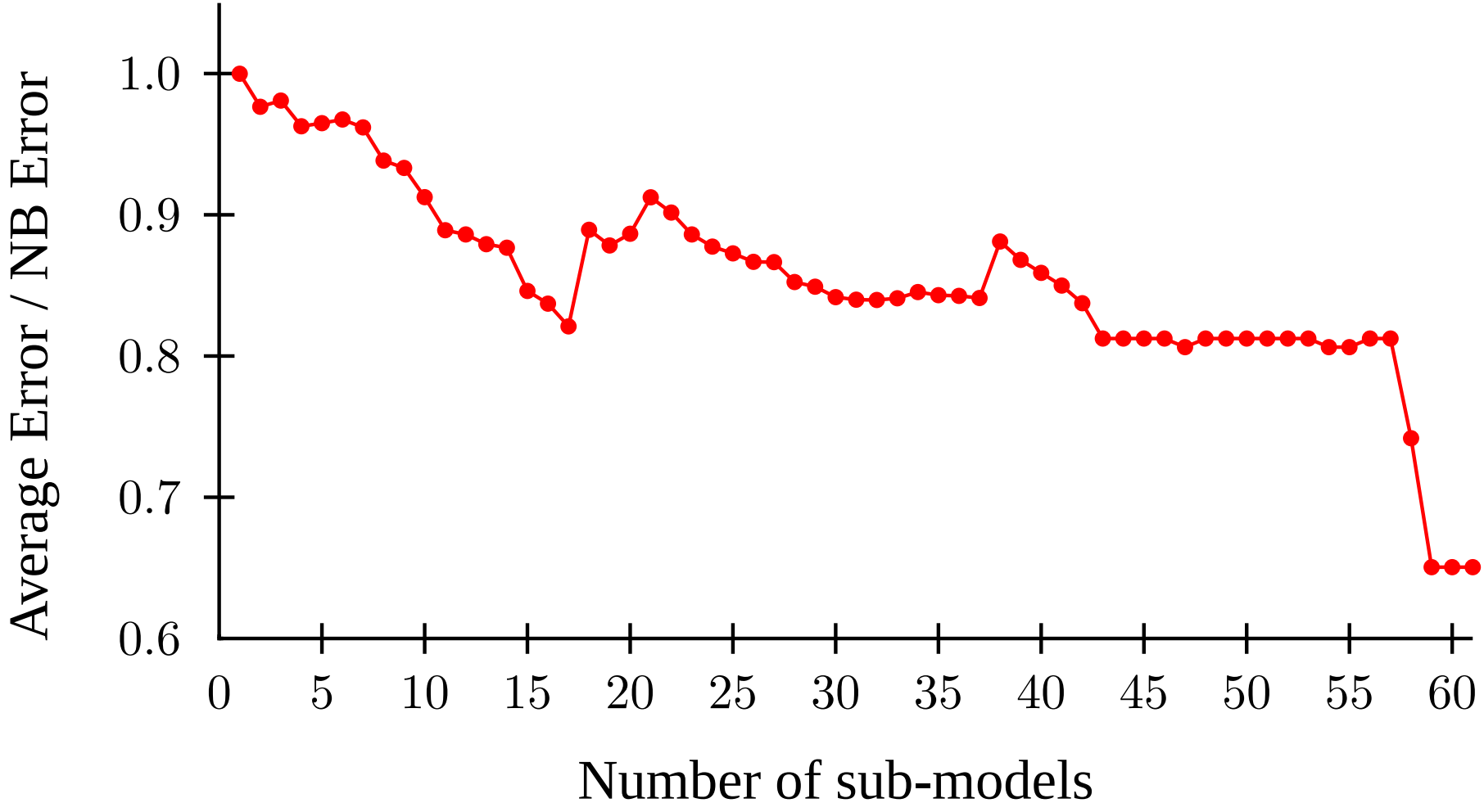


Further features

- Incremental
- Parallelizable
- Anytime classification



Anytime classification



Cerquides and de Màntaras, 2005

- Minimum frequency = 1 outperforms minimum frequency = 30
- MAPLMG, learns maximum a posteriori weights.
  - substantial reduction in error at substantial computational cost

Lazy Elimination

- Delete x_i values such that for some x_j, $P(X_i \mid x_j) = 1.0$
- Eg, $P(y \mid \text{pregnant}) = P(y \mid \text{pregnant}, \text{female})$

pregnant	female
not-pregnant	female
not-pregnant	male

- Substantial reduction in error

Conclusions

- AODE classifies by conditional probability estimation
- Averages over all single-parent one-dependence models
- Computationally efficient learning at some cost in classification time
  - learning time is linear on number of training objects
- Error appears comparable to LBR and TAN
  - slightly higher bias but lower variance
- Error appears comparable to Boosted J48 for small data sets
  - substantially lower variance
- Supports incremental, parallel and anytime classification