# Learning without search

Geoff Webb, Zijian Zheng, Kai Ming Ting, Zhihai Wang, Fei Zheng, Janice Boughton, Houssam Salem

Monash University,

Melbourne, Australia

http://www.csse.monash.edu.au/∼webb

MONASH University

# Overview

- Most learning algorithms search a model space for a model, or a parameter space for a parametrization of a fixed model that best fits the training data.

MONASH University

# Overview

- Most learning algorithms search a model space for a model, or a parameter space for a parametrization of a fixed model that best fits the training data.
- Averaged $n$-Dependence Estimators (A$n$DE) is a family of classification learning algorithms that exemplifies an alternative paradigm
  - learner uses a fixed model to extrapolate from observed low-order probabilities to the required high-order probability

MONASH University

# So What?

# So What?

- Theoretical interest
  - alternative paradigms exist;

MONASH University

# So What?

- Theoretical interest
  - alternative paradigms exist;
    *if things aren't going right ...*

# So What?

- Theoretical interest
  - alternative paradigms exist;
    *if things aren't going right ...*

*go left!*

MONASH University

# So What?

- Theoretical interest
  - alternative paradigms exist;
    *if things aren't going right ...*

    *go left!*
  - generative learning can achieve the same low bias profile as discriminative.

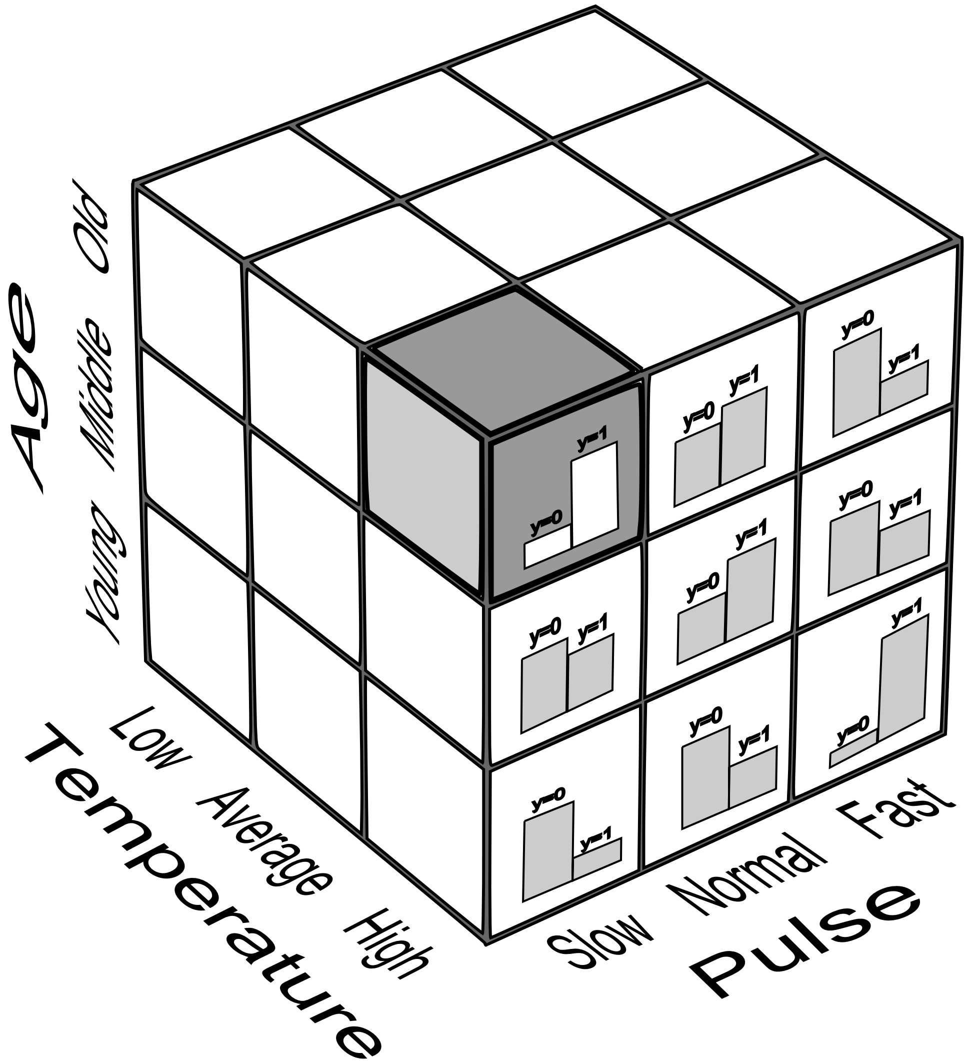
MONASH University

# So What?

- Theoretical interest
  - alternative paradigms exist;
    *if things aren't going right ...*

    *go left!*
  - generative learning can achieve the same low bias profile as discriminative.
- Unique and valuable combination of practical features

# Classification: A geometric view




MONASH University

# Learning by extrapolation

- In contrast to search paradigm, naive Bayes extrapolates to high-order conditional probabilities from lower-order probabilities.

MONASH University

# Naive Bayes

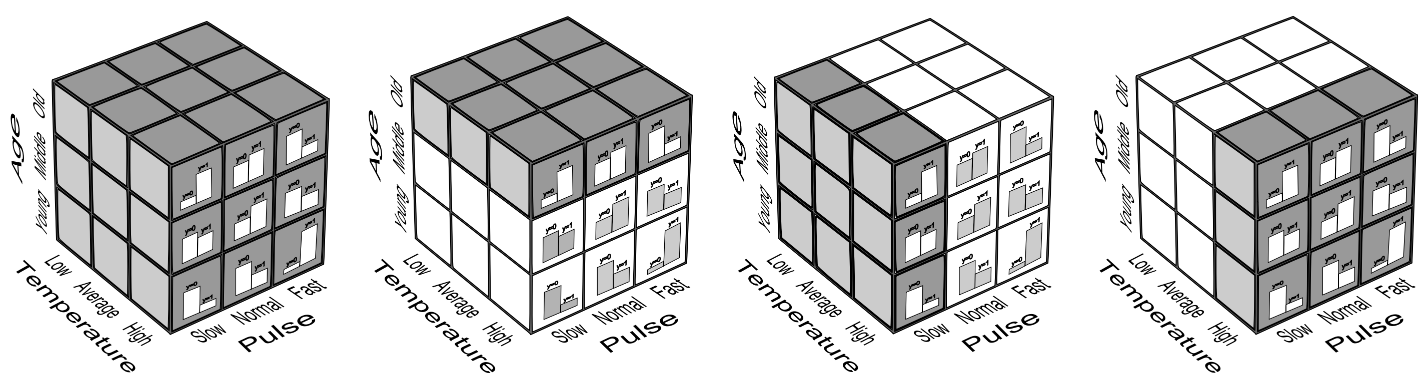
- $$\begin{aligned} \mathrm{P}(y \mid \mathbf{x}) &\propto \mathrm{P}(y, \mathbf{x}) \\ &= \mathrm{P}(y)\mathrm{P}(\mathbf{x} \mid y) \end{aligned}$$

# Naive Bayes

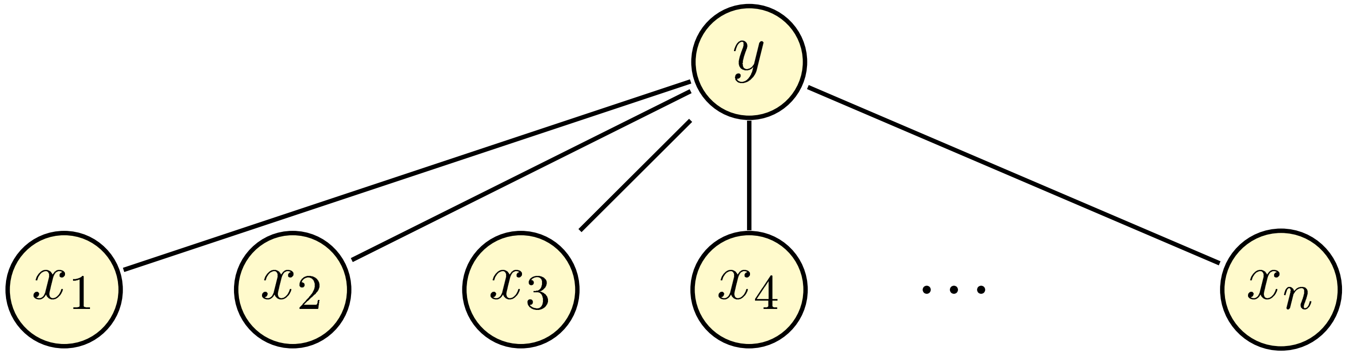
- $$\begin{aligned}\mathrm{P}(y \mid \mathbf{x}) &\propto \mathrm{P}(y, \mathbf{x}) \\ &= \mathrm{P}(y)\mathrm{P}(\mathbf{x} \mid y)\end{aligned}$$
- Attribute independence assumption
  - $$\mathrm{P}(\mathbf{x} \mid y) = \prod_{i=1}^{n} \mathrm{P}(x_i \mid y)$$

# Naive Bayes

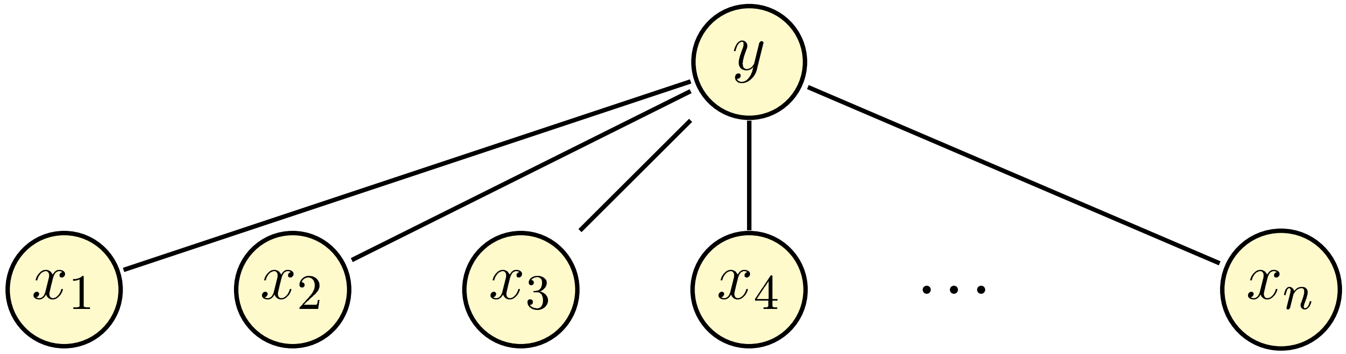
- $$\begin{aligned}\mathrm{P}(y \mid \mathbf{x}) &\propto \mathrm{P}(y, \mathbf{x}) \\ &= \mathrm{P}(y)\mathrm{P}(\mathbf{x} \mid y)\end{aligned}$$
- Attribute independence assumption
  - $$\mathrm{P}(\mathbf{x} \mid y) = \prod_{i=1}^{n} \mathrm{P}(x_i \mid y)$$
- No search
  - extrapolate high-order probabilities from low order probabilities $\mathrm{P}(y)$ and $\mathrm{P}(x_i \mid y)$



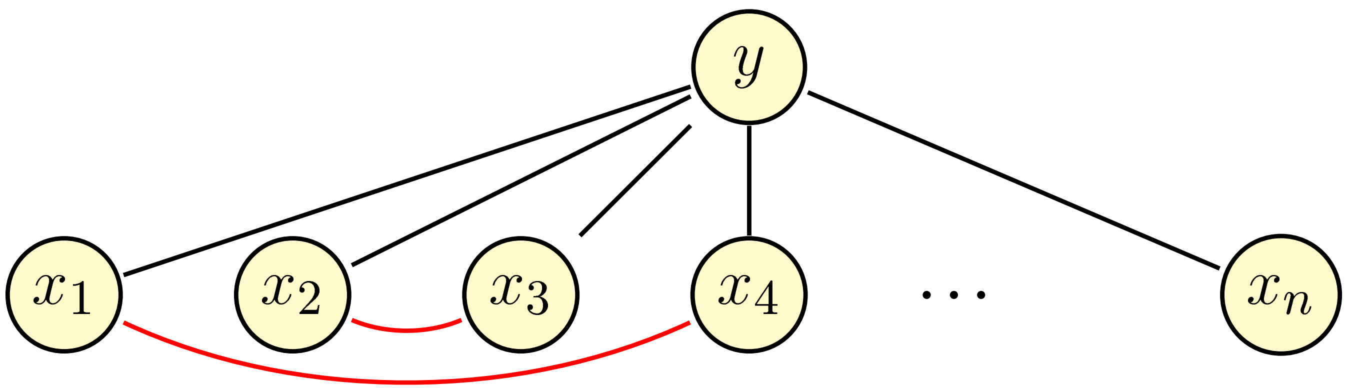
# The fixed model



# The fixed model




- Adding arbitrary links will decrease bias but increase variance



# But how to decide which links?

- Could use search
  - requires additional computation

MONASH University

# But how to decide which links?

- Could use search
  - requires additional computation
- Alternative: use all of a class of models and combine predictions

# ANDE

- Averaged $n$ Dependence Estimators

# ANDE

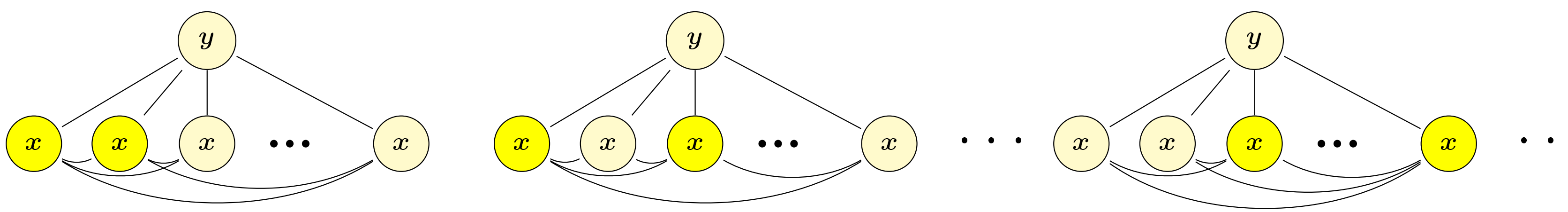
- Averaged $n$ Dependence Estimators
- Select $n$, the order of dependence

# ANDE

- Averaged $n$ Dependence Estimators
- Select $n$, the order of dependence
- Each model selects $n$ parent attributes —
  - all other attributes are independent given the class and these $n$ parents

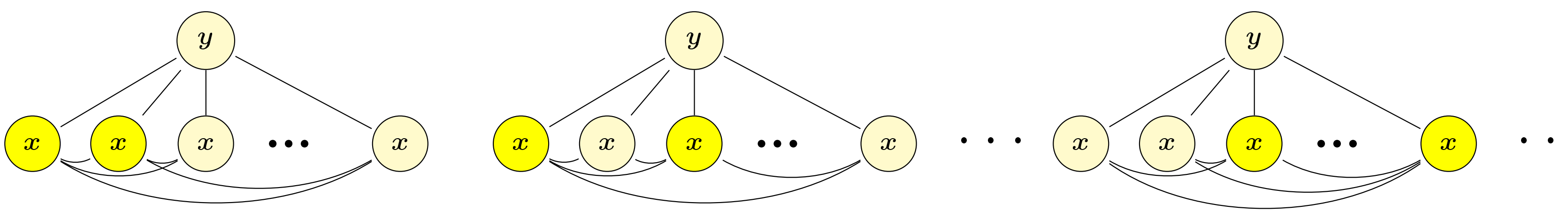
# ANDE

- Averaged $n$ Dependence Estimators
- Select $n$, the order of dependence
- Each model selects $n$ parent attributes —
  - all other attributes are independent given the class and these $n$ parents



# ANDE

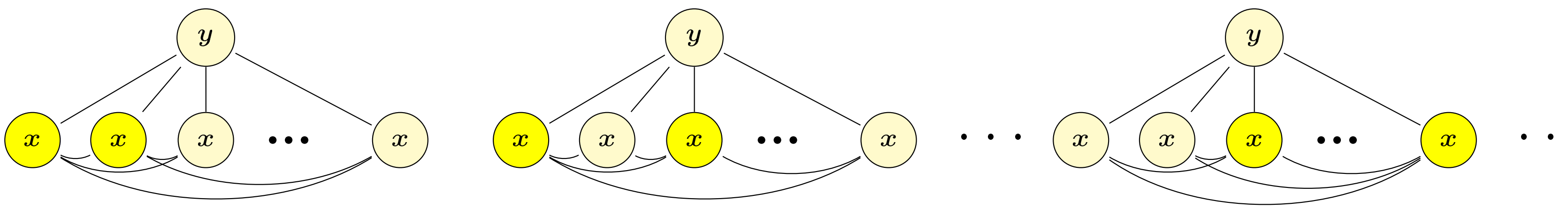
- Averaged $n$ Dependence Estimators
- Select $n$, the order of dependence
- Each model selects $n$ parent attributes —
  - all other attributes are independent given the class and these $n$ parents




- Each model has lower bias but higher variance than NB

# ANDE

- Averaged $n$ Dependence Estimators
- Select $n$, the order of dependence
- Each model selects $n$ parent attributes —
  - all other attributes are independent given the class and these $n$ parents




- Each model has lower bias but higher variance than NB
- Ensembling reduces the variance

MONASH University

# ANDE (derivation)

- ANDE aims to use

$$\mathrm{A}n\mathrm{DE}(y, \mathbf{x}) = \sum_{s \in S^n} \mathrm{P}(y, x_s)\mathrm{P}(\mathbf{x} \mid y, x_s) / \binom{a}{n}.$$

where $S^n$ indicates all subsets of size $n$ of the set $\{1, \ldots a\}$.

MONASH University

# ANDE (derivation)

- ANDE aims to use

$$\mathrm{A}n\mathrm{DE}(y, \mathbf{x}) = \sum_{s \in S^n} \mathrm{P}(y, x_s)\mathrm{P}(\mathbf{x} \,|\, y, x_s) / \binom{a}{n}.$$

where $S^n$ indicates all subsets of size $n$ of the set $\{1, \ldots a\}$.

- In practice we use

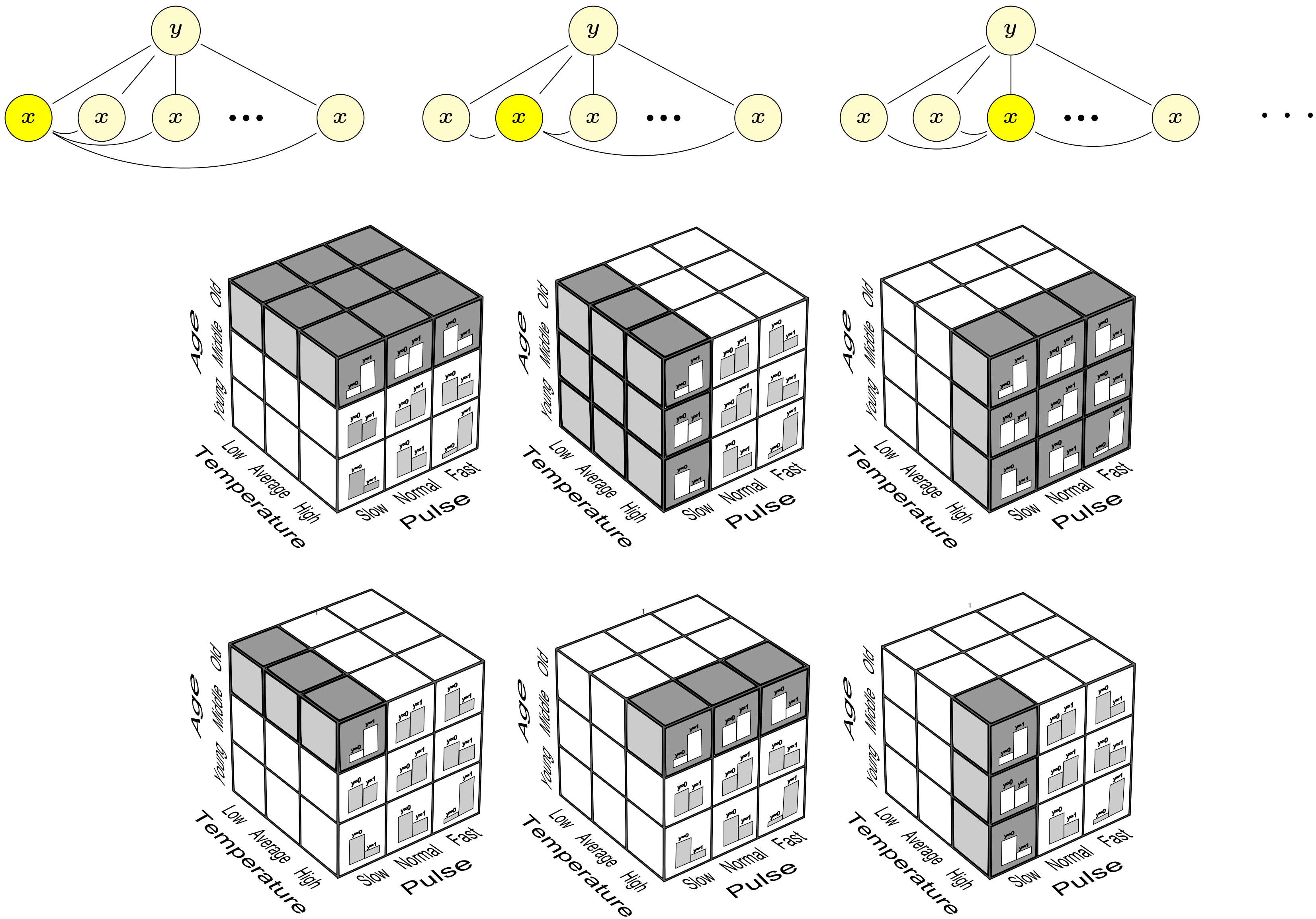
$$\mathrm{A}n\mathrm{DE}(y, \mathbf{x}) = \begin{cases} \dfrac{\sum_{s \in S^n} \delta(s)\mathrm{P}(y, x_s)\mathrm{P}(\mathbf{x} \,|\, y, x_s)}{\sum_{s \in S^n} \delta(s)} & : \sum_{s \in S^n} \delta(s) > 0 \\ \mathrm{A}(n{-}1)\mathrm{DE}(y, \mathbf{x}) & : \text{otherwise} \end{cases}$$

MONASH University

# ANDE Equivalences

- $\mathrm{A0DE} = \mathrm{NB}$
- $\mathrm{A1DE} = \mathrm{AODE}$

# AODE



y
x
...
Age
Old
Middle
Young
Temperature
Low
Average
High
Pulse
Slow
Normal
Fast

# Popular


1. Affendey, L.S., Paris, I.H.M. Mustapha, N. Sulaiman, M.N., Muda, Z.: Ranking of influencing factors in predicting students academic performance. *Inform. Technol. J.,* 9 (2010) 832-837.
2. Birzele, F., Kramer, S.: A new representation for protein secondary structure prediction based on frequent patterns. *Bioinformatics* **22**(21) (2006) 2628–2634.
3. Camporelli, M.: Using a Bayesian Classifier for Probability Estimation: Analysis of the AMIS Score for Risk Stratification in Myocardial Infarction. Diploma Thesis, Dept. Informatics, U. Zurich (2006).
4. Eduardo, AL., Iakes E., Beatriz, G., Alfonso, V., David, J.: EcID. A database for the inference of functional interactions in E. coli. *Nucleic Acids Research* (2008) doi:10.1093/nar/gkn853
5. Ferrari, L.D., Aitken, S.: Mining housekeeping genes with a naive Bayes classifier. *BMC Genomics* **7**(1) (2006) 277.
6. Flikka, K., Martens, L., Vandekerckhove, J., Gevaert, K., Eidhammer, I.: Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. *Proteomics* **6**(7) (2006) 2086–2094.
7. Garcia, B., Aler, R., Ledezma, A., Sanchis, A.: Protein-protein functional Assoc. prediction using genetic programming. In: *Proc. Tenth Annual Conf. Genetic and Evolutionary Computation*, ACM. (2008) 347–348.
8. García-Jiménez B, Juan D, Ezkurdia I, Andrés-León E, Valencia A.: Inference of Functional Relations in Predicted Protein Networks with a Machine Learning Approach. *PLoS ONE* (2010) 5(4): e9969. doi:10.1371/journal.pone.0009969
9. Hopfgartner, F., Urruty, T., Lopez, P.B., Villa, R., Jose, J.M: Simulated evaluation of faceted browsing based on feature selection. *Multimedia Tools and Applications* 47(3) (2010) 631-662.
10. Hunt, K.: *Evaluation of Novel Algorithms to Optimize Risk Stratification Scores in Myocardial Infarction*. PhD thesis, Dept. Informatics, University of Zurich (2006).
11. Kunchevaa, L.I., Vilas, V.J.D.R., Rodr´ıguezc, J.J.: Diagnosing scrapie in sheep: A classification experiment. *Computers in Biology and Medicine* **37**(8) (2007) 1194–1202.
12. Kurz, D, Bernstein, A, Hunt, K, Radovanovic, D, Erne, P, Siudak, Z, Bertel, O: Simple point-of-care risk stratification in acute coronary syndromes: the AMIS model. *British Medical Journal* **95**(8) (2009) 662.
13. Lasko, T.A., Atlas, S.J., Barry, M.J., Chueh, K.H.C.: Automated identification of a physician's primary patients. *Journal of the American Medical Informatics Assoc.* **13**(1) (2006) 74–79.
14. Lau, Q.P., Hsu, W., Lee, M.L., Mao, Y., Chen, L.: Prediction of cerebral aneurysm rupture. In: *Proc. 19th IEEE International Conf. Tools with Artificial Intelligence* (2007) 350–357.
15. Leon, A., et al.: EcID. A database for the inference of functional interactions in E. coli. *Nucleic Acids Research* **37**(Database issue) (2009) D629.
16. Liew, CY., Ma, XH., Yap, CW.: Consensus model for identification of novel PI3K inhibitors in large chemical library. *Journal of Computer-Aided Molecular Design*. **24**(2) (2010) 131-141.
17. Masegosa, AR., Joho, H., Jose JM.: Evaluating Query-Independent Object Features for Relevancy Prediction. In *Advances in Information Retrieval*. Springer Berlin. (2007) 283-294.
18. Nikora, A.P.: Classifying requirements: Towards a more rigorous analysis of natural-language specifications. In: *Proc. Sixteenth IEEE International Symp.Software Reliability Engineering* (2005) 291–300.
19. Orhan, Z., Altan, Z.: Impact of feature selection for corpus-based WSD in Turkish. In: *Proc. 5th Mexican International Conf. Artificial Intelligence* (2006) 868–878.
20. Shahri, SH., Jamil, H.: An Extendable Meta-learning Algorithm for Ontology Mapping. In *Flexible Query Answering Systems*, Springer (2009) 418-430.
21. Simpson, M., Demner-Fushman, D., Sneiderman, C., Antani, S., Thoma, G.: Using non-lexical features to identify effective indexing terms for biomedical illustrations. In: *Proc. 12th Conf. European Chapter of the Assoc. Computational Linguistics* (2009) 737–744.
22. Tian, Y., Chen, C., Zhang, C: AODE for Source Code Metrics for Improved Software Maintainability. *Fourth International Conf. Semantics, Knowledge and Grid* (2008) pp.330-335.
23. Wang, H., Klinginsmith, J., Dong, X., Lee, A., Guha, R., Wu, Y., Crippen, G., Wild, D.: Chemical data mining of the NCI human tumor cell line database. *Journal of Chemical Information and Modeling* **47**(6) (2007) 2063–2076.

# A$a$DE

- $S^a = \{\{1, \ldots a\}\}$ and hence when $n = a$, $x_s = \mathbf{x}$

MONASH University

# A$a$DE

- $S^a = \{\{1, \ldots a\}\}$ and hence when $n = a$, $x_s = \mathbf{x}$
- A$a$DE seeks to classify using

$$\mathrm{A}a\mathrm{DE}(y, \mathbf{x}) = \mathrm{P}(y, \mathbf{x})\mathrm{P}(\mathbf{x} \,|\, y, \mathbf{x}) / \binom{a}{a}$$

MONASH University

# A$a$DE

- $S^a = \{\{1, \ldots a\}\}$ and hence when $n = a$, $x_s = \mathbf{x}$
- A$a$DE seeks to classify using

$$\mathrm{A}a\mathrm{DE}(y, \mathbf{x}) = \mathrm{P}(y, \mathbf{x})\mathrm{P}(\mathbf{x} \,|\, y, \mathbf{x}) / \binom{a}{a}$$

- $\mathrm{P}(y, \mathbf{x})$ is estimated directly from $\mathcal{D}$

# A$a$DE

- $S^a = \{\{1, \ldots a\}\}$ and hence when $n = a$, $x_s = \mathbf{x}$
- A$a$DE seeks to classify using

$$\mathrm{A}a\mathrm{DE}(y, \mathbf{x}) = \mathrm{P}(y, \mathbf{x})\mathrm{P}(\mathbf{x} \,|\, y, \mathbf{x}) / \binom{a}{a}$$

- $\mathrm{P}(y, \mathbf{x})$ is estimated directly from $\mathcal{D}$
- $\mathrm{P}(\mathbf{x} \,|\, y, \mathbf{x})$ and $\binom{a}{a}$ both equal 1.0

# A$a$DE

- $S^a = \{\{1, \dots a\}\}$ and hence when $n = a$, $x_s = \mathbf{x}$
- A$a$DE seeks to classify using

$$\mathrm{A}a\mathrm{DE}(y, \mathbf{x}) = \mathrm{P}(y, \mathbf{x})\mathrm{P}(\mathbf{x} \,|\, y, \mathbf{x}) / \binom{a}{a}$$

- $\mathrm{P}(y, \mathbf{x})$ is estimated directly from $\mathcal{D}$
- $\mathrm{P}(\mathbf{x} \,|\, y, \mathbf{x})$ and $\binom{a}{a}$ both equal 1.0
- seeks to classify using $\mathrm{P}(y, \mathbf{x})$ estimated directly from $\mathcal{D}$, cascading to ever lower dependence estimators when the combination of attribute-values is not be present in $\mathcal{D}$.

# A$a$DE

- $S^a = \{\{1, \ldots a\}\}$ and hence when $n = a$, $x_s = \mathbf{x}$
- A$a$DE seeks to classify using

$$\mathrm{A}a\mathrm{DE}(y, \mathbf{x}) = \mathrm{P}(y, \mathbf{x})\mathrm{P}(\mathbf{x} \,|\, y, \mathbf{x}) / \binom{a}{a}$$

- $\mathrm{P}(y, \mathbf{x})$ is estimated directly from $\mathcal{D}$
- $\mathrm{P}(\mathbf{x} \,|\, y, \mathbf{x})$ and $\binom{a}{a}$ both equal 1.0
- seeks to classify using $\mathrm{P}(y, \mathbf{x})$ estimated directly from $\mathcal{D}$, cascading to ever lower dependence estimators when the combination of attribute-values is not be present in $\mathcal{D}$.
- has asymptotic error of the Bayes optimal classifier!

# A$a$DE

- $S^a = \{\{1, \ldots a\}\}$ and hence when $n = a$, $x_s = \mathbf{x}$
- A$a$DE seeks to classify using

$$\mathrm{A}a\mathrm{DE}(y, \mathbf{x}) = \mathrm{P}(y, \mathbf{x})\mathrm{P}(\mathbf{x} \,|\, y, \mathbf{x}) / \binom{a}{a}$$

- $\mathrm{P}(y, \mathbf{x})$ is estimated directly from $\mathcal{D}$
- $\mathrm{P}(\mathbf{x} \,|\, y, \mathbf{x})$ and $\binom{a}{a}$ both equal 1.0
- seeks to classify using $\mathrm{P}(y, \mathbf{x})$ estimated directly from $\mathcal{D}$, cascading to ever lower dependence estimators when the combination of attribute-values is not be present in $\mathcal{D}$.
- has asymptotic error of the Bayes optimal classifier!
- has computational complexity of at least $\mathrm{O}(k \prod_{i=1}^{a} v_i)$

# Computational Complexity

- Space: $\mathrm{O}\left(k\binom{a}{n+1}v^{n+1}\right)$

# Computational Complexity

- Space: $\mathrm{O}\left(k\binom{a}{n+1}v^{n+1}\right)$
- Training Time: $\mathrm{O}\left(t\binom{a}{n+1}\right)$

MONASH University

# Computational Complexity

- Space: $\mathrm{O}\left(k\binom{a}{n+1}v^{n+1}\right)$
- Training Time: $\mathrm{O}\left(t\binom{a}{n+1}\right)$
- Testing Time: $\mathrm{O}\left(ka\binom{a}{n}\right)$

MONASH University

# Computational Complexity

- Space: $\mathrm{O}\left(k\binom{a}{n+1}v^{n+1}\right)$
- Training Time: $\mathrm{O}\left(t\binom{a}{n+1}\right)$
- Testing Time: $\mathrm{O}\left(ka\binom{a}{n}\right)$
- In practice our Weka implementation of $\mathrm{A3DE}$ is defeated by high-dimensional data

MONASH University

# Evaluation

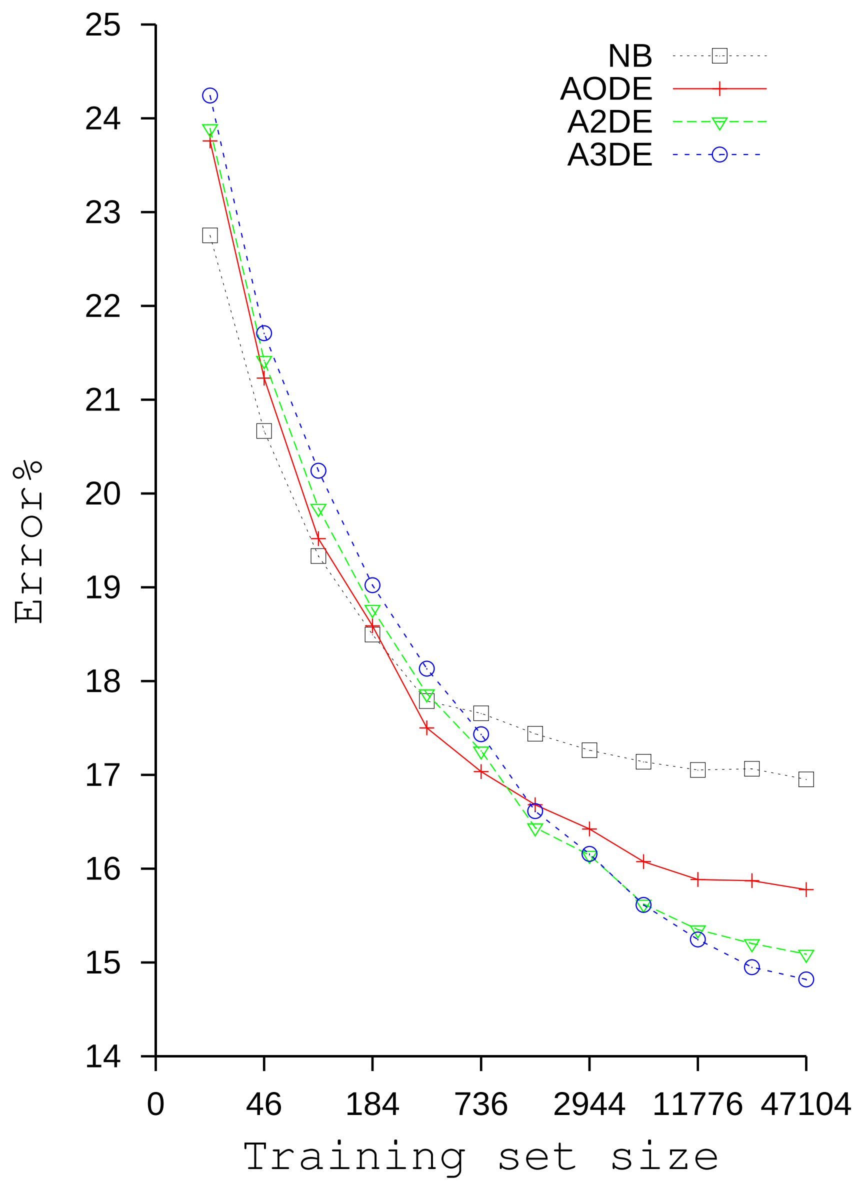
- 62 UCI data sets used previously in related research
- Use fifty runs of two-fold cross validation to estimate bias, variance, 0-1 loss and RMSE.

# ANDE, $n$ = 0, 1 and 2

Win/Draw/Loss

| | A2DE vs AODE | | A2DE vs NB | | AODE vs NB | |
|---|---|---|---|---|---|---|
| | W/D/L | $p$ | W/D/L | $p$ | W/D/L | $p$ |
| Bias | 47/0/15 | <0.001 | 49/2/11 | <0.001 | 48/0/14 | <0.001 |
| Variance | 19/1/42 | <0.001 | 15/0/47 | <0.001 | 20/1/41 | 0.005 |
| 0-1 loss | 33/2/27 | 0.259 | 42/1/19 | 0.002 | 44/1/17 | <0.001 |
| RMSE | 35/1/26 | 0.153 | 45/0/17 | <0.001 | 49/1/12 | <0.001 |

# Error as function of training set size




MONASH University

# Higher-order probabilities vs search

- AODE & A2DE vs TAN
- Win/Draw/Loss

| | A2DE vs TAN | | AODE vs TAN | |
|---|---|---|---|---|
| | W/D/L | $p$ | W/D/L | $p$ |
| Bias | 34/0/28 | 0.263 | 20/1/41 | 0.005 |
| Variance | 48/0/14 | $<$0.001 | 52/1/9 | $<$0.001 |
| Zero-one loss | 48/0/14 | $<$0.001 | 43/1/18 | 0.001 |
| RMSE | 43/1/18 | 0.001 | 40/1/21 | 0.010 |

MONASH University

# Higher-order probabilities vs search

- A2DE & AODE vs MAPLMG
- Win/Draw/Loss

| | A2DE vs MAPLMG | | AODE vs MAPLMG | |
|---|---|---|---|---|
| | W/D/L | $p$ | W/D/L | $p$ |
| Bias | 40/0/22 | 0.015 | 17/4/41 | 0.001 |
| Variance | 19/1/42 | 0.002 | 36/5/21 | 0.031 |
| Zero-one loss | 30/1/31 | 0.500 | 22/4/36 | 0.043 |
| RMSE | 34/1/28 | 0.263 | 19/0/39 | 0.006 |

- Win/Draw/Loss, A2DE vs MAPLMG on 10 largest data sets
  - 10/0/0, $p = 0.001$

MONASH University

# No search vs state-of-the-art

- A2DE vs RF10, RF100
- Win/Draw/Loss

| | A2DE vs RF10 | | A2DE vs RF100 | |
|---|---|---|---|---|
| | W/D/L | $p$ | W/D/L | $p$ |
| Bias | 14/1/47 | $<0.001$ | 20/2/40 | 0.007 |
| Variance | 56/1/5 | $<0.001$ | 46/1/15 | $<0.001$ |
| 0-1 loss | 40/1/21 | 0.010 | 34/2/26 | 0.399 |
| RMSE | 39/0/23 | 0.028 | 34/0/28 | 0.263 |

MONASH University

# Times

- Average training/test time per instance, excluding Census Income

| | NB | AODE | A2DE | TAN | MAPLMG | RF10 | RF100 |
|---|---|---|---|---|---|---|---|
| Train | 0.0005 | 0.0007 | 0.0413 | 0.0022 | 0.1290 | 0.0177 | 0.1645 |
| Test | 0.0001 | 0.0022 | 0.0552 | 0.0002 | 0.0025 | 0.0001 | 0.0017 |

# Scalability: training quantity

- On 10 smallest data sets

| | NB | AODE | A2DE | TAN | MAPLMG | RF10 | RF100 |
|---|---|---|---|---|---|---|---|
| Train | 0.0020 | 0.0020 | 0.0920 | 0.0064 | 0.1339 | 0.0114 | 0.0844 |

- On 10 largest data sets, excluding Census Income

| | NB | AODE | A2DE | TAN | MAPLMG | RF10 | RF100 |
|---|---|---|---|---|---|---|---|
| Train | 0.0001 | 0.0002 | 0.0077 | 0.0004 | 0.1360 | 0.0229 | 0.2016 |

MONASH University

# Scalability: dimensionality

- On 10 lowest dimensional data sets (4-8 atts)

| | NB | AODE | A2DE | TAN | MAPLMG | RF10 | RF100 |
|---|---|---|---|---|---|---|---|
| Train | 0.0010 | 0.0009 | 0.0011 | 0.0018 | 0.0448 | 0.0046 | 0.0311 |
| Test | 0.0001 | 0.0002 | 0.0002 | 0.0001 | 0.0003 | 0.0001 | 0.0004 |

- On 10 highest dimensional data sets (43-70 atts)

| | NB | AODE | A2DE | TAN | MAPLMG | RF10 | RF100 |
|---|---|---|---|---|---|---|---|
| Train | 0.0008 | 0.0017 | 0.1870 | 0.0067 | 0.5025 | 0.0435 | 0.4125 |
| Test | 0.0002 | 0.0097 | 0.2976 | 0.0005 | 0.0097 | 0.0002 | 0.0033 |

MONASH University

# Feating

- Feating uses the all combinations of attribute-values approach to ensemble local classifiers.

MONASH University

# Feating

- Feating uses the all combinations of attribute-values approach to ensemble local classifiers.
  - first generic ensemble method that is effective for low variance learners such as SVM

# Feating

- Feating uses the all combinations of attribute-values approach to ensemble local classifiers.
  - first generic ensemble method that is effective for low variance learners such as SVM
- Feating uses the mode of the posterior class predictions while ANDE uses the mean of the joint probability estimates

# Feating

- Feating uses the all combinations of attribute-values approach to ensemble local classifiers.
  - first generic ensemble method that is effective for low variance learners such as SVM
- Feating uses the mode of the posterior class predictions while ANDE uses the mean of the joint probability estimates
- ANDE has lower bias but higher variance than Feating NB

# Future Research

- Alternative classes of models

# Future Research

- Alternative classes of models
- Approximation of extrapolated values

# Future Research

- Alternative classes of models
- Approximation of extrapolated values
- Extension to numeric data

# Future Research

- Alternative classes of models
- Approximation of extrapolated values
- Extension to numeric data
- Extension to high-dimensional data

# Future Research

- Alternative classes of models
- Approximation of extrapolated values
- Extension to numeric data
- Extension to high-dimensional data
- Weighting, parent selection, child selection

# Future Research

- Alternative classes of models
- Approximation of extrapolated values
- Extension to numeric data
- Extension to high-dimensional data
- Weighting, parent selection, child selection
- Understand why ensembling joint probabilities results in lower bias than ensembling posteriors

# Conclusions

- ANDE demonstrates that there is an alternative to the search paradigm that delivers accuracy competitive with the state-of-the-art

MONASH University

# Conclusions

- ANDE demonstrates that there is an alternative to the search paradigm that delivers accuracy competitive with the state-of-the-art
- A2DE is a practical algorithm with:

MONASH University

# Conclusions

- ANDE demonstrates that there is an alternative to the search paradigm that delivers accuracy competitive with the state-of-the-art
- A2DE is a practical algorithm with:
  - computational complexity linear wrt number of training examples;

# Conclusions

- ANDE demonstrates that there is an alternative to the search paradigm that delivers accuracy competitive with the state-of-the-art
- A2DE is a practical algorithm with:
  - computational complexity linear wrt number of training examples;
  - direct prediction of class probabilities;

MONASH University

# Conclusions

- ANDE demonstrates that there is an alternative to the search paradigm that delivers accuracy competitive with the state-of-the-art
- A2DE is a practical algorithm with:
  - computational complexity linear wrt number of training examples;
  - direct prediction of class probabilities;
  - integrated handling of missing values;

# Conclusions

- ANDE demonstrates that there is an alternative to the search paradigm that delivers accuracy competitive with the state-of-the-art
- A2DE is a practical algorithm with:
  - computational complexity linear wrt number of training examples;
  - direct prediction of class probabilities;
  - integrated handling of missing values;
  - robustness in the face of noise;

# Conclusions

- ANDE demonstrates that there is an alternative to the search paradigm that delivers accuracy competitive with the state-of-the-art
- A2DE is a practical algorithm with:
  - computational complexity linear wrt number of training examples;
  - direct prediction of class probabilities;
  - integrated handling of missing values;
  - robustness in the face of noise;
  - non-reliance on tuneable parameters;

# Conclusions

- ANDE demonstrates that there is an alternative to the search paradigm that delivers accuracy competitive with the state-of-the-art
- A2DE is a practical algorithm with:
  - computational complexity linear wrt number of training examples;
  - direct prediction of class probabilities;
  - integrated handling of missing values;
  - robustness in the face of noise;
  - non-reliance on tuneable parameters;
  - simple mechanism to control bias/variance trade-off;

# Conclusions

- ANDE demonstrates that there is an alternative to the search paradigm that delivers accuracy competitive with the state-of-the-art
- A2DE is a practical algorithm with:
  - computational complexity linear wrt number of training examples;
  - direct prediction of class probabilities;
  - integrated handling of missing values;
  - robustness in the face of noise;
  - non-reliance on tuneable parameters;
  - simple mechanism to control bias/variance trade-off;
  - incremental, parallel and anytime classification; and

# Conclusions

- ANDE demonstrates that there is an alternative to the search paradigm that delivers accuracy competitive with the state-of-the-art
- A2DE is a practical algorithm with:
  - computational complexity linear wrt number of training examples;
  - direct prediction of class probabilities;
  - integrated handling of missing values;
  - robustness in the face of noise;
  - non-reliance on tuneable parameters;
  - simple mechanism to control bias/variance trade-off;
  - incremental, parallel and anytime classification; and
  - direct theoretical basis (Bayes optimal prediction except insofar as clearly specified assumptions are violated).