# Discovering Independence

Serafín Moral
Dpto. Ciencias de la Computación
Universidad de Granada
Spain

# Motivation

- I wanted to learn about the behaviour of different scores used for learning Bayesian networks.
- Compare independence approaches with score approaches.
- Compare with a new score based on information theory and imprecise probability.
- Useful for the interpretation of Bayesian networks.
- Useful to develop new scores or for tuning parameters.

# Outline

- The basic problem
- Basic approaches
- Upper entropy of imprecise probability procedure
- Experiments
- Changing the parameters of experiments
- The effect of equivalent sample size
- A new imprecise score
- Application to Classification Trees

# The Problem

- Two variables $X$ and $Y$ taking values on set $\{0, 1\}$.
- A joint probability distribution, $P$, $P_X$ and $P_Y$ are its marginal distributions.
- A sample of pairs of values $D = (x_1, y_1), \ldots, (x_N, y_N)$.
- Notation:

| | $Y = 0$ | $Y = 1$ | |
|---|---|---|---|
| $X = 0$ | $N(0,0)$ | $N(0,1)$ | $N_X(0)$ |
| $X = 1$ | $N(1,0)$ | $N(1,1)$ | $N_X(1)$ |
| | $N_Y(0)$ | $N_Y(1)$ | $N$ |

# The Problem

- Two variables $X$ and $Y$ taking values on set $\{0, 1\}$.
- A joint probability distribution, $P$, $P_X$ and $P_Y$ are its marginal distributions.
- A sample of pairs of values $D = (x_1, y_1), \ldots, (x_N, y_N)$.
- Notation:

| | $Y = 0$ | $Y = 1$ | |
|---|---|---|---|
| $X = 0$ | $N(0,0)$ | $N(0,1)$ | $N_X(0)$ |
| $X = 1$ | $N(1,0)$ | $N(1,1)$ | $N_X(1)$ |
| | $N_Y(0)$ | $N_Y(1)$ | $N$ |

| | $Y = 0$ | $Y = 1$ | |
|---|---|---|---|
| $X = 0$ | $\hat{P}(0,0)$ | $\hat{P}(0,1)$ | $\hat{P}_X(0)$ |
| $X = 1$ | $\hat{P}(1,0)$ | $\hat{P}(1,1)$ | $\hat{P}_X(1)$ |
| | $\hat{P}_Y(0)$ | $\hat{P}_Y(1)$ | $1.0$ |

# Three Basic Problems

- Decide whether $X$ and $Y$ are independent.

# Three Basic Problems

- Decide whether $X$ and $Y$ are independent.
- Estimate a joint probability for $X$ and $Y$
  - Assuming dependence: estimating $P^e(i,j)$.
  - Assuming independence: estimating $P^e_X(i)$ and $P^e_Y(j)$, and making $P^e(i,j) = P^e_X(i).P^e_Y(j)$

# Three Basic Problems

- Decide whether $X$ and $Y$ are independent.
- Estimate a joint probability for $X$ and $Y$
  - Assuming dependence: estimating $P^e(i,j)$.
  - Assuming independence: estimating $P^e_X(i)$ and $P^e_Y(j)$, and making $P^e(i,j) = P^e_X(i).P^e_Y(j)$
- Decide a model to classify $Y$ as a function of $X$.
  - Assuming dependence: using $P^e(j|i)$.
  - Assuming independence: using $P^e_Y(j)$

# Independence Tests

- The sample Mutual Information is computed:

$$G = \sum_{i,j} \hat{P}(i,j) \log \left( \frac{\hat{P}(i,j)}{\hat{P}_X(i)\hat{P}_Y(j)} \right)$$

- $2.N.G$ asymptotically follows a Chi-square distribution with one degree of freedom.
- First measure $CHI$ is 1 minus $p$-value of this test.
- Two decision rules, $CHI^{0.05}$ (dependence if $CHI > 0.95$) and $CHI^{0.20}$ (dependence if $CHI > 0.80$)

# K2 Score- Cooper, Herskovits (1992)

- It assumes a Bayesian point of view
- To decide between independence (IND) and dependence (DEP): 'a posteriori' probabilities given the data:

$$P(DEP|D) = \frac{P(D|DEP).P(DEP)}{P(D|DEP).P(DEP) + P(D|DEP).P(DEP)},$$

$$P(IND|D) = \frac{P(D|IND).P(IND)}{P(D|DEP).P(DEP) + P(D|DEP).P(DEP)}$$

- Assuming equal 'a priori' probability then,

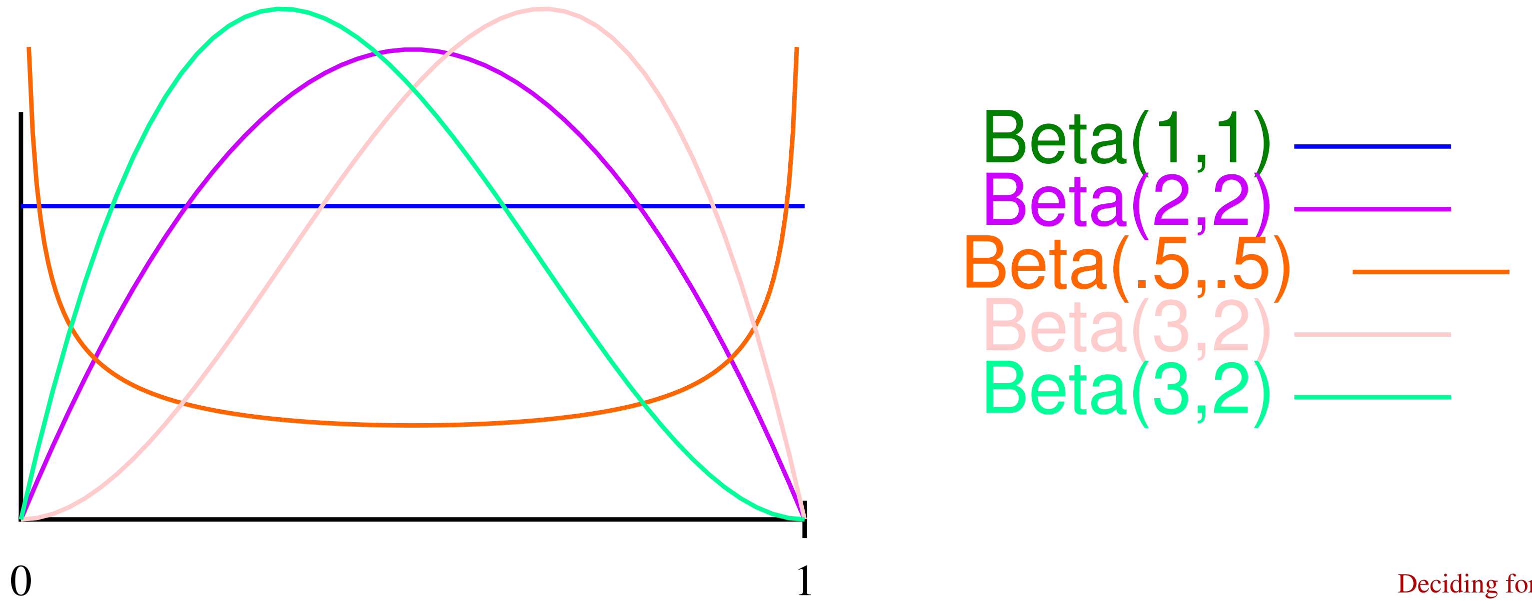
$$P(DEP|D) \propto P(D|DEP), \qquad P(IND|D) \propto P(D|IND)$$

# K2 Score - Dirichlet Distribution

- Categorical variable $Z$ taking values on set $\{z_1, \ldots, z_k\}$. Dirichlet 'a priori' distribution $Dir(\alpha_1, \ldots, \alpha_k)$

$$f(p(z_1), \ldots, p(z_k)) = \frac{\Gamma(S)}{\Gamma(\alpha_1) \ldots . \Gamma(\alpha_k)} p(z_1)^{\alpha_1 - 1} \ldots . p(z_k)^{\alpha_k - 1}$$

where $S = \sum_i \alpha_i$ is called the equivalent sample size.



# K2 Score - Dirichlet Distribution

- If we have a sample and observe $(N_1,\ldots,N_k)$ values, the 'a posteriori' probability is Dirichlet $Dir(\alpha_1+N_1,\ldots,\alpha_k+N_k)$
- The probability of observing $(N_1,\ldots,N_k)$ is

$$\frac{\Gamma(s)}{\Gamma(\alpha_1).\ldots.\Gamma(\alpha_1)}\cdot\frac{\Gamma(\alpha_1+N_1).\ldots.\Gamma(\alpha_k+N_k)}{\Gamma(s+N)}$$

- The expected 'a posteriori' probability is equal to:

$$P^e(z_i)=\frac{\alpha_i+N_i}{S+N}$$

- This probability favors large samples (for non uniform probabilities):

$$P(D_1,D_2)=P(D_1).P(D_2|D_1)>P(D_1).P(D_2)$$

# Intuitive Interpretation

The values

$$\left(\frac{\alpha_1}{S}, \ldots, \frac{\alpha_k}{S}\right)$$

represent the 'a priori' probabilities for the values of the variables based in our past experience.

The value $S = \sum_i \alpha_i$ is called the equivalent sample size measures the importance of our past experience.
Larger values make that 'a priori' probabilities have more importance.
The expected 'a posteriori' probability is equal to:

$$P^e(z_i) = \frac{\alpha_i + N_i}{S + N}$$

# K2 Score

- Independence. $(p_X(0), p_X(1))$ and $(p_Y(0), p_Y(1))$ follow independent Dirichlet distributions $Dir(1,1)$ and $p(i,j) = p_X(i).p_Y(j)$.

$$K2I = \frac{\Gamma(2)}{\Gamma(N+2)} \left( \prod_i \frac{\Gamma(N_X(i)+1)}{\Gamma(1)} \right) \cdot \frac{\Gamma(2)}{\Gamma(N+2)} \left( \prod_j \frac{\Gamma(N_Y(j)+1)}{\Gamma(1)} \right)$$

- Dependence. $(p_X(0), p_X(1))$, $(p(0|X=0), p(1|X=0))$, and $(p(0|X=1), p(1|X=1))$ follow independent Dirichlet distributions $Dir(1,1)$ and $p(i,j) = p_X(i).p(j|X=i)$.

$$K2D = \frac{\Gamma(2)}{\Gamma(N+2)} \left( \prod_i \frac{\Gamma(N_X(i)+1)}{\Gamma(1)} \right) \cdot \prod_i \frac{\Gamma(2)}{\Gamma(N_X(i)+2)} \left( \prod_j \frac{\Gamma(N(i,j)+1)}{\Gamma(1)} \right)$$

$$\boxed{K2 = K2D - K2I}$$

# Bayesian Dirichlet equivalent scores

Heckerman, Geiger, and Chickering (1995)

- K2 score is not symmetric: the global distribution is not Dirichlet. Parameters:

| | $Y=0$ | $Y=1$ | $X$ |
|---|---|---|---|
| $X=0$ | 1 | 1 | 1 |
| $X=1$ | 1 | 1 | 1 |

- Bayesian equivalent scores assume a global $S$ and then

| | $Y=0$ | $Y=1$ | $X$ |
|---|---|---|---|
| $X=0$ | $S/4$ | $S/4$ | $S/2$ |
| $X=1$ | $S/4$ | $S/4$ | $S/2$ |

# Basic Result

$P(x,y)$ is a $D(\alpha_{00},\alpha_{01},\alpha_{10},\alpha_{11})$ if and only if $P_X(x)$ follows a $D(\alpha_{00}+\alpha_{01},\alpha_{10}+\alpha_{11})$ and $P_Y(X=0)$ is $D(\alpha_{00},\alpha_{01})$ and $P_Y(X=1)$ is $D(\alpha_{10},\alpha_{11})$

The equivalent sample sizes of the conditional distributions have to be equal to the parameters of the marginal distribution.

# Bayesian Equivalent Scores

- Independence

$$BSI = \frac{\Gamma(S)}{\Gamma(N+S)} \left( \prod_i \frac{\Gamma(N_X(i) + S/2)}{\Gamma(S/2)} \right) \cdot \frac{\Gamma(S)}{\Gamma(N+S)} \left( \prod_j \frac{\Gamma(N_Y(j) + S/2)}{\Gamma(S/2)} \right)$$

- Dependence

$$BSD = \frac{\Gamma(S)}{\Gamma(N+S)} \left( \prod_i \frac{\Gamma(N_X(i) + S/2)}{\Gamma(S/2)} \right) \cdot \prod_i \frac{\Gamma(S/2)}{\Gamma(N_X(i) + S/2)} \left( \prod_j \frac{\Gamma(N(i,j) + S/4)}{\Gamma(S/4)} \right)$$

$$\boxed{BS2 = BSD - BSI \quad (S = 2), \qquad BS4 = BSD - BSI \quad (S = 4)}$$

# BIC - Bayesian Information Criterion

It tries to minimize the addition of the length of the Model + (Data|Model)

Length(Data|Model) is minus the entropy of the maximum likelihood estimation of $P$ under the given model.

$$BIC = N.\sum_{i,j} \hat{P}(i,j) \log\left(\frac{\hat{P}(i,j)}{\hat{P}_X(i).\hat{P}_Y(j)}\right) - (1/2)\log(N)$$

# Akaike Criterion

It penalizes complexity by a constant factor:

$$AKA = N.\sum_{i,j}\hat{P}(i,j)\log\left(\frac{\hat{P}(i,j)}{\hat{P}_X(i).\hat{P}_Y(j)}\right) - (1/2)$$

# Upper Entropy of Imprecise Estimation

- If we have an 'a priori' Dirichlet $Dir(\alpha_1, \ldots, \alpha_k)$ and observe $(N_1, \ldots, N_k)$ values, the 'a posteriori' expectation of $P(z_i)$ is:

$$P^e(z_i) = \frac{N_i + \alpha_i}{N + S}$$

- The Imprecise Dirichlet Model (Walley, 1996) assumes only a sample size $S$ and all $(\alpha_1, \ldots, \alpha_k)$ such that $\sum_i \alpha_i = S, \alpha_i > 0$.

$$P^e(z_i) \in \left[ \frac{N_i}{N+S}, \frac{N_i + S}{N+S} \right]$$

- Upper entropy: the supremum of the entropies of the probabilities verifying the intervals $\overline{H}(Z)$.

# Example

- Assume two possible values and $S = 1$.

# Example

- Assume two possible values and $S = 1$.
- If we observe:

| $N_i$ | 1 | 2 |
|---|---|---|
| Int. | [1/4,2/4] | [2/4,3/4] |

Entropy: $\log(2)$

# Example

- Assume two possible values and $S = 1$.
- If we observe:

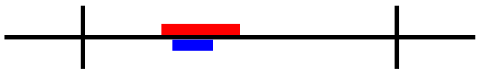
| $N_i$ | 1 | 2 |
|---|---|---|
| Int. | [1/4,2/4] | [2/4,3/4] |

Entropy: $\log(2)$

- If we observe:

| $N_i$ | 2 | 4 |
|---|---|---|
| Int. | [2/7,3/7] | [4/7,5/7] |

Entropy: $-3/7\log(3/7) - 4/7\log(4/7)$



# Deciding for Independence

- *Independence.* Apply the imprecise Dirichlet model to $X$ and $Y$ and compute

$$IPI = \overline{H}(X) + \overline{H}(Y)$$

- *Dependence.* Apply the imprecise Dirichlet model to $X$ and the conditional probabilities of $Y$ and compute

$$IPD = \overline{H}(X) + \sum_i \hat{P}(i).\overline{H}(Y|X = i)$$

$$\boxed{IMP = IPI - IPD}$$

A non-symmetrical score.

# Experiments

- We generate 10000 cases of independent variables and 10000 cases of dependent variables.
- If we are in the case of independence, we generate the probabilities $(P_X(0), P_X(1))$ and $(P_Y(0), P_Y(1))$ with Dirichlet distribution $Dir(1,1)$
- If we are in the case of dependence, we generate the probabilities $(P_X(0), P_X(1))$ with Dirichlet distribution $Dir(1,1)$ and the conditional probabilities of $Y|X = i$ with Dirichlet $Dir(0.5, 0.5)$.
- Of each distribution, we obtain samples of sizes $s = 3, 5, 10, 20, 50, 100, 1000, 10000$
- We try to determine the correct model from the samples using the different scores.

# Measuring Errors

- Deciding Independence: Number of errors
- Estimating the Joint Probability: The opposite of the expected log likelihood.

$$-\sum_{i,j} P(i,j) \log P^e(i,j) = KL(P,P^e) + H(P)$$

where $P^e(i,j) = \frac{N(i,j)+0.5}{2}$ in the case of dependence (Bayesian estimation with $s = 2$).

- Classify $Y$: The opposite of the expected log likelihood of the conditional probability of $Y$

$$-\sum_{i,j} P(i,j) \log P^e(j|X=i)$$

# Results

Dependence

| Size | $CHI^{0.05}$ | $CHI^{0.20}$ | K2 | BD2 | BD4 | BIC | AKA | IMP |
|---|---|---|---|---|---|---|---|---|
| 3 | 10000 | 8119 | 8119 | 5008 | 5008 | 8119 | 6879 | 8119 |
| 5 | 8292 | 6787 | 6787 | 5149 | 5149 | 6787 | 5848 | 6787 |
| 10 | 7306 | 5635 | 5521 | 4893 | 4667 | 6001 | 4582 | 5868 |
| 20 | 5788 | 4234 | 4627 | 4294 | 3927 | 5406 | 3525 | 4609 |
| 50 | 4222 | 3021 | 3693 | 3489 | 3141 | 4256 | 2502 | 3400 |
| 100 | 3180 | 2213 | 2920 | 2853 | 2527 | 3445 | 1790 | 2550 |
| 1000 | 1138 | 763 | 1307 | 1238 | 1123 | 1515 | 610 | 1000 |
| 10000 | 346 | 224 | 480 | 465 | 429 | 542 | 174 | 324 |

# Results

Independence

| Size | $CHI^{0.05}$ | $CHI^{0.20}$ | K2 | BD2 | BD4 | BIC | AKA | IMP |
|---|---|---|---|---|---|---|---|---|
| 3 | 0 | 841 | 841 | 3322 | 3322 | 841 | 2485 | 841 |
| 5 | 335 | 1548 | 1548 | 2622 | 2622 | 1548 | 2920 | 1548 |
| 10 | 389 | 1698 | 1912 | 2030 | 2300 | 1304 | 3223 | 2082 |
| 20 | 499 | 2057 | 1643 | 1559 | 1992 | 773 | 3290 | 2441 |
| 50 | 515 | 2100 | 1199 | 1053 | 1585 | 497 | 3310 | 2644 |
| 100 | 522 | 2099 | 904 | 758 | 1208 | 318 | 3303 | 2678 |
| 1000 | 536 | 2098 | 299 | 236 | 399 | 90 | 3259 | 2754 |
| 10000 | 517 | 1968 | 89 | 64 | 117 | 21 | 3182 | 2705 |

# Results

## Total Errors

| Size | $CHI^{0.05}$ | $CHI^{0.20}$ | K2 | BD2 | BD4 | BIC | AKA | IMP |
|---|---|---|---|---|---|---|---|---|
| 3 | 10000 | 8960 | 8960 | 8330 | 8330 | 8960 | 9364 | 8960 |
| 5 | 8627 | 8335 | 8335 | 7771 | 7771 | 8335 | 8768 | 8335 |
| 10 | 7695 | 7333 | 7433 | 6923 | 6967 | 7305 | 7805 | 7950 |
| 20 | 6287 | 6291 | 6270 | 5853 | 5919 | 6179 | 6815 | 7050 |
| 50 | 4737 | 5121 | 4892 | 4542 | 4726 | 4753 | 5812 | 6044 |
| 100 | 3702 | 4312 | 3824 | 3611 | 3735 | 3763 | 5093 | 5228 |
| 1000 | 1674 | 2861 | 1606 | 1474 | 1522 | 1605 | 3869 | 3754 |
| 10000 | 863 | 2192 | 569 | 529 | 546 | 563 | 3356 | 3029 |

# Results

## Statistical Tests

| Size | $CHI^{0.05}$ | $CHI^{0.20}$ | K2 | BD2 | BD4 | BIC | AKA | IMP |
|---|---|---|---|---|---|---|---|---|
| 3 | 10000 | 8960 | 8960 | 8330 | 8330 | 8960 | 9364 | 8960 |
| 5 | 8627 | 8335 | 8335 | 7771 | 7771 | 8335 | 8768 | 8335 |
| 10 | 7695 | 7333 | 7433 | 6923 | 6967 | 7305 | 7805 | 7950 |
| 20 | 6287 | 6291 | 6270 | 5853 | 5919 | 6179 | 6815 | 7050 |
| 50 | 4737 | 5121 | 4892 | 4542 | 4726 | 4753 | 5812 | 6044 |
| 100 | 3702 | 4312 | 3824 | 3611 | 3735 | 3763 | 5093 | 5228 |
| 1000 | 1674 | 2861 | 1606 | 1474 | 1522 | 1605 | 3869 | 3754 |
| 10000 | 863 | 2192 | 569 | 529 | 546 | 563 | 3356 | 3029 |

*Statistical tests should decrease significance level with the sample size in order to minimize the total number of errors.*

# Results - K-L distance

Dependence

| Size | $CHI^{0.05}$ | $CHI^{0.20}$ | K2 | BD2 | BD4 | BIC | AKA | IMP |
|---|---|---|---|---|---|---|---|---|
| 3 | 0.28794 | 0.25469 | 0.25469 | 0.23437 | 0.23437 | 0.25469 | 0.25134 | 0.25469 |
| 5 | 0.19904 | 0.18584 | 0.18584 | 0.17455 | 0.17455 | 0.18584 | 0.18319 | 0.18584 |
| 10 | 0.12943 | 0.11676 | 0.11620 | 0.11400 | 0.11287 | 0.11865 | 0.11402 | 0.12070 |
| 20 | 0.07050 | 0.06412 | 0.06520 | 0.06409 | 0.06327 | 0.06850 | 0.06279 | 0.06666 |
| 50 | 0.03032 | 0.02778 | 0.02898 | 0.02858 | 0.02816 | 0.03045 | 0.02724 | 0.02904 |
| 100 | 0.01551 | 0.01438 | 0.01510 | 0.01514 | 0.01472 | 0.01603 | 0.01417 | 0.01503 |
| 1000 | 0.00155 | 0.00150 | 0.00160 | 0.00157 | 0.00155 | 0.00165 | 0.00149 | 0.00155 |
| 10000 | 0.00015 | 0.00015 | 0.00016 | 0.00016 | 0.00016 | 0.00016 | 0.00015 | 0.00015 |

*Again Akaike is very good (if 'a priori' dependence)*

# Results - K-L distance

Independence

| Size | $CHI^{0.05}$ | $CHI^{0.20}$ | K2 | BD2 | BD4 | BIC | AKA | IMP |
|---|---|---|---|---|---|---|---|---|
| 3 | 0.17470 | 0.19164 | 0.19164 | 0.20577 | 0.20577 | 0.19164 | 0.19723 | 0.19164 |
| 5 | 0.13717 | 0.14940 | 0.14940 | 0.15553 | 0.15553 | 0.14940 | 0.15339 | 0.14940 |
| 10 | 0.08367 | 0.09341 | 0.09398 | 0.09312 | 0.09440 | 0.09127 | 0.09709 | 0.09175 |
| 20 | 0.04820 | 0.05427 | 0.05312 | 0.05227 | 0.05327 | 0.04982 | 0.05653 | 0.05330 |
| 50 | 0.02145 | 0.02449 | 0.02297 | 0.02253 | 0.02324 | 0.02139 | 0.02560 | 0.02397 |
| 100 | 0.01104 | 0.01271 | 0.01150 | 0.01122 | 0.01163 | 0.01065 | 0.01333 | 0.01241 |
| 1000 | 0.00115 | 0.00134 | 0.00109 | 0.00108 | 0.00111 | 0.00104 | 0.00141 | 0.00131 |
| 10000 | 0.00011 | 0.00013 | 0.00010 | 0.00010 | 0.00010 | 0.00010 | 0.00014 | 0.00013 |

# Results - K-L distance

Total Error

| Size | $CHI^{0.05}$ | $CHI^{0.20}$ | K2 | BD2 | BD4 | BIC | AKA | IMP |
|---|---|---|---|---|---|---|---|---|
| 3 | 0.46264 | 0.44632 | 0.44633 | 0.440141 | 0.440141 | 0.446327 | 0.448566 | 0.446327 |
| 5 | 0.33621 | 0.33523 | 0.33524 | 0.330078 | 0.330078 | 0.335237 | 0.336579 | 0.335237 |
| 10 | 0.21309 | 0.21016 | 0.21018 | 0.207127 | 0.207272 | 0.209917 | 0.211113 | 0.212449 |
| 20 | 0.11870 | 0.11838 | 0.11831 | 0.116357 | 0.116538 | 0.118318 | 0.119318 | 0.119957 |
| 50 | 0.05177 | 0.05226 | 0.05195 | 0.051115 | 0.051396 | 0.051842 | 0.052843 | 0.05301 |
| $10^2$ | 0.02655 | 0.02709 | 0.02660 | 0.026362 | 0.026347 | 0.026679 | 0.027503 | 0.027445 |
| $10^3$ | 0.0027 | 0.00283 | 0.00269 | 0.002655 | 0.002663 | 0.002688 | 0.002897 | 0.002864 |
| $10^4$ | 0.00026 | 0.00028 | 0.00026 | 0.000257 | 0.000256 | 0.000259 | 0.000291 | 0.000284 |

*BD2 is the best, and IMP is bad for intermediate samples, AKA bad for large samples*

# Changing the Conditions of Experiments

- We generate 10000 cases of independent variables and 10000 cases of dependent variables.
- If we are in the case of independence, we generate the probabilities $(p_X(0), p_X(1))$ and $(p_Y(0), p_Y(1))$ with Dirichlet distribution $Dir(2,2)$
- If we are in the case of dependence, we generate the probabilities $(p_X(0), p_X(1))$ with Dirichlet distribution $Dir(2,2)$ and the conditional probabilities of $Y|X=i$ with Dirichlet $Dir(2,2)$.
- Of each distribution, we obtain samples of sizes $s = 3, 5, 10, 20, 50, 100, 1000$
- We try to determine the correct model from the samples using the different scores.

# Results - Expected Log Likelihood

| Size | $CHI^{0.05}$ | $CHI^{0.20}$ | K2 | BD2 | BD4 | BIC | AKA | IMP |
|---|---|---|---|---|---|---|---|---|
| 3 | 2.678415 | 2.707347 | 2.707347 | 2.717153 | 2.717153 | 2.707347 | 2.719971 | 2.707347 |
| 5 | 2.626271 | 2.648085 | 2.648085 | 2.650484 | 2.650484 | 2.648085 | 2.656433 | 2.648085 |
| 10 | 2.536459 | 2.547588 | 2.547812 | 2.543038 | 2.543934 | 2.544948 | 2.553794 | 2.545218 |
| 20 | 2.456497 | 2.460718 | 2.458458 | 2.457833 | 2.457705 | 2.457229 | 2.462325 | 2.459232 |
| 50 | 2.388114 | 2.388964 | 2.388058 | 2.388221 | 2.388003 | 2.388089 | 2.389774 | 2.388677 |
| 100 | 2.360719 | 2.361192 | 2.360589 | 2.360927 | 2.360656 | 2.360996 | 2.36151 | 2.361035 |
| 1000 | 2.335361 | 2.335469 | 2.33535 | 2.335391 | 2.335366 | 2.33541 | 2.33553 | 2.335496 |

*The performance of BD2 is not as good as before. IMP is good (small samples)*

# Changing the Conditions of Experiments

- We generate 10000 cases of independent variables and 10000 cases of dependent variables.
- If we are in the case of independence, we generate the probabilities $(P_X(0), P_X(1))$ and $(P_Y(0), P_Y(1))$ with Dirichlet distribution $Dir(0.5, 0.5)$
- If we are in the case of dependence, we generate the probabilities $(P_X(0), P_X(1))$ with Dirichlet distribution $Dir(0.5, 0.5)$ and the conditional probabilities of $Y|X = i$ with Dirichlet $Dir(0.5, 0.5)$.
- Of each distribution, we obtain samples of sizes $s = 3, 5, 10, 20, 50, 100, 1000$
- We try to determine the correct model from the samples using the different scores.

# Results - Number of errors

| Size | $CHI^{0.05}$ | $CHI^{0.20}$ | K2 | BD2 | BD4 | BIC | AKA | IMP |
|---|---|---|---|---|---|---|---|---|
| 3 | 10000 | 9243 | 9243 | 9285 | 9285 | 9243 | 9239 | 9243 |
| 5 | 8994 | 8762 | 8762 | 8666 | 8666 | 8762 | 8787 | 8762 |
| 10 | 8422 | 8066 | 8089 | 8592 | 8562 | 8099 | 8087 | 8259 |
| 20 | 7603 | 7479 | 7466 | 7973 | 8150 | 7517 | 7574 | 7711 |
| 50 | 6650 | 6565 | 6603 | 7138 | 7637 | 6650 | 6794 | 6902 |
| 100 | 5870 | 5937 | 5861 | 6496 | 7133 | 5957 | 6259 | 6253 |
| 1000 | 3871 | 4378 | 3950 | 4660 | 5288 | 4054 | 5021 | 4570 |

*IMP is better than BD2*

# The effect of $S$

Same original conditions (100000 distributions).

Independence

| Size | $S = 0.02$ | $S = 0.2$ | $S = 1.0$ | $S = 1.6$ | $S = 2$ | $s = 4$ | $s = 16$ |
|---|---|---|---|---|---|---|---|
| 3 | 33397 | 33397 | 33397 | 33397 | 33397 | 33397 | 33397 |
| 5 | 26910 | 26910 | 26910 | 26910 | 26910 | 26910 | 49022 |
| 10 | 13148 | 13148 | 13669 | 14227 | 20864 | 23653 | 36833 |
| 20 | 7533 | 8394 | 10000 | 12776 | 15858 | 20147 | 38829 |
| 50 | 4454 | 4836 | 7047 | 8952 | 10467 | 15830 | 34116 |
| 100 | 2874 | 3319 | 5030 | 6534 | 7448 | 11873 | 29641 |
| 1000 | 495 | 672 | 1440 | 1958 | 2337 | 4007 | 12431 |

*Small values of $S$ are in favor of independence.*

# The effect of $S$

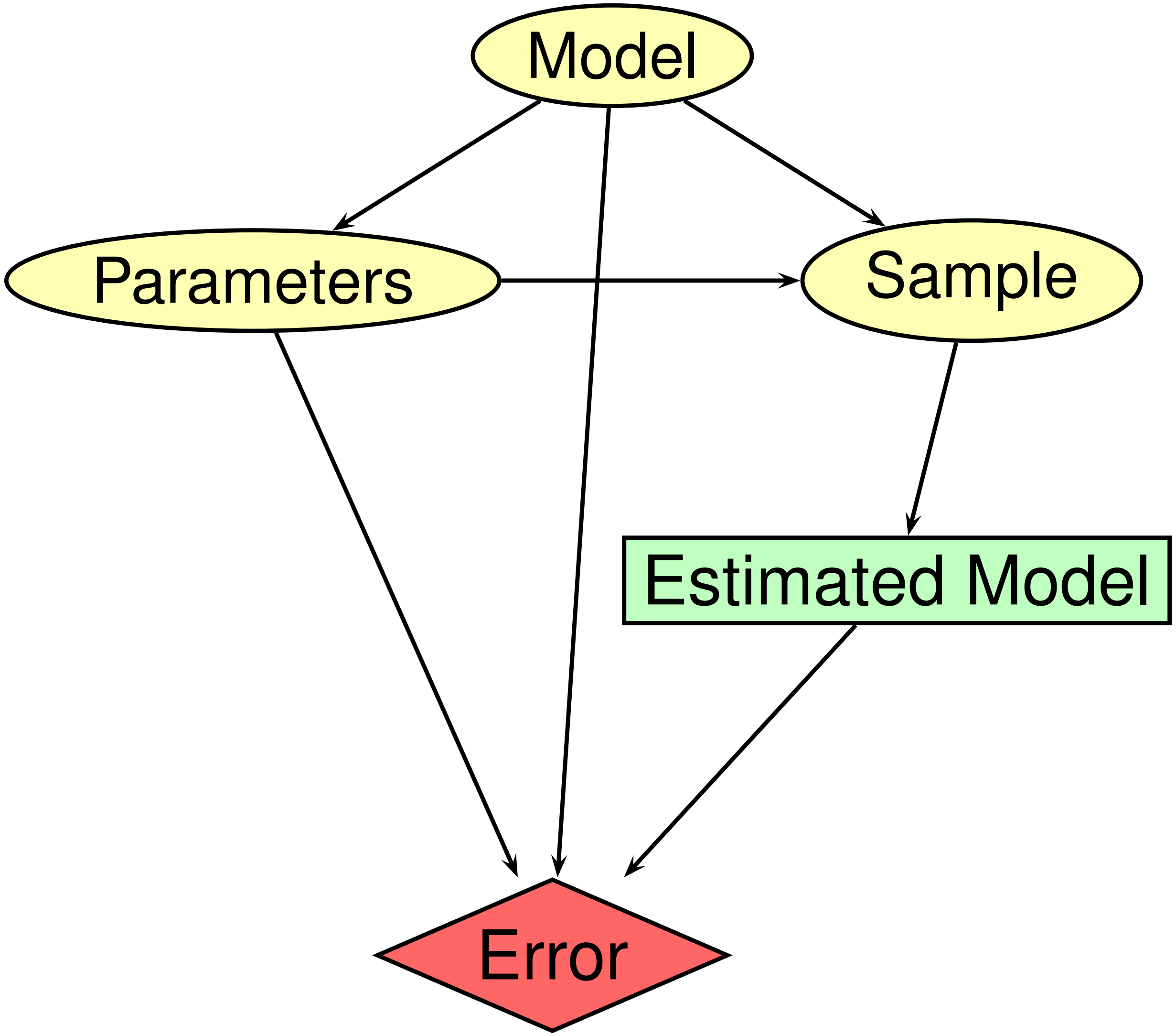
Same original conditions (100000 distributions).

Dependence

| Size | $s = 0.02$ | $s = 0.2$ | $s = 1.0$ | $s = 1.6$ | $s = 2$ | $s = 4$ | $s = 16$ |
|---|---|---|---|---|---|---|---|
| 3 | 50006 | 50006 | 50006 | 50006 | 50006 | 50006 | 50006 |
| 5 | 51470 | 51470 | 51470 | 51470 | 51470 | 51470 | 33332 |
| 10 | 57030 | 57030 | 55961 | 55167 | 48391 | 45986 | 36625 |
| 20 | 57222 | 53548 | 49764 | 46039 | 42826 | 39530 | 29267 |
| 50 | 50680 | 45159 | 38959 | 36192 | 34564 | 30657 | 23313 |
| 100 | 43109 | 37236 | 31534 | 29255 | 28335 | 25027 | 19020 |
| 1000 | 19614 | 16546 | 13807 | 12952 | 12584 | 11367 | 9159 |

*Small values of $s$ are in favor of independence.*

# Approximating Distributions



# New Score

- Independence

$$P(D|IND).\sum_{(i,j)} P_X^e(i).P_Y^e(j)\log\left(\frac{P_X^e(i).P_Y^e(j)}{P^e(i,j)}\right)$$

- Dependence

$$P(D|DEP).\sum_{(i,j)} P^e(i,j)\log\left(\frac{P^e(i,j)}{P_X^e(i).P_Y^e(j)}\right)$$

# Results

| Size | BD2 | BD4 | BD2MOD |
|---:|---:|---:|---:|
| 3 | 2.32649 | 2.32649 | 2.32649 |
| 5 | 2.218363 | 2.218363 | 2.218363 |
| 10 | 2.091777 | 2.091873 | 2.091647 |
| 20 | 2.003681 | 2.003812 | 2.003637 |
| 50 | 1.937799 | 1.937997 | 1.937789 |
| 100 | 1.913013 | 1.913072 | 1.912986 |
| 1000 | 1.889333 | 1.889339 | 1.889333 |

# Classification

- Friedman, Geiger, Goldszmidt (1997) Specialized scores:

$$\prod_i P(Y_i | X_i, M)$$

- Acid, Campos, Castellano (2005) General scores provide good results.
- Friedman et al. use different procedure to estimate parameters!!

# The modified score for classification

- Independence

$$P(D|IND).\sum_{(i,j)} P_X^e(i).P_Y^e(j)\log\left(\frac{P_Y^e(j)}{P^e(j|X=i)}\right)$$

- Dependence

$$P(D|DEP).\sum_{(i,j)} P^e(i,j)\log\left(\frac{P^e(j|X=i)}{P_Y^e(j)}\right)$$

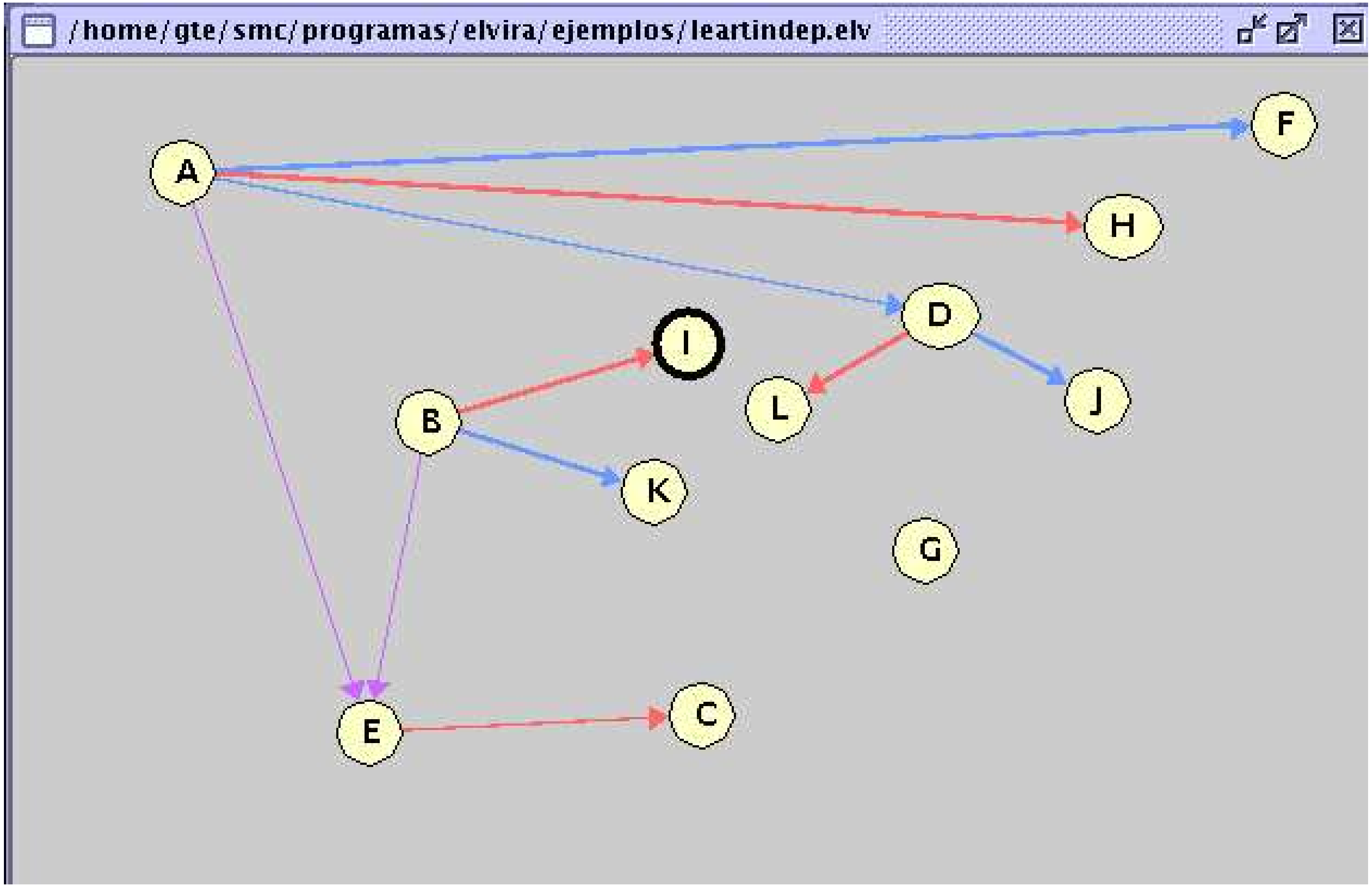
# Results

| Size | $CHI^{0.05}$ | BD2 | BD4 | BD2MOD |
|---:|---:|---:|---:|---:|
| 3 | 1.173371 | 1.152434 | 1.152434 | 1.152434 |
| 5 | 1.096266 | 1.090344 | 1.090344 | 1.090344 |
| 10 | 1.02057 | 1.014398 | 1.014494 | 1.014266 |
| 20 | 0.962427 | 0.96028 | 0.960411 | 0.960236 |
| 50 | 0.919406 | 0.918781 | 0.918979 | 0.918771 |
| 100 | 0.903223 | 0.903026 | 0.903085 | 0.903 |
| 1000 | 0.888041 | 0.888003 | 0.888009 | 0.888003 |

# Main Conclusions

- Bayesian scores are good if hypothesis can be assumed and we want to minimize the number of errors.
- It can be dangerous a blind application: it can produce with a very short sample a result of dependence when there is independence. Learn a networks with 12 independent binary variables from a sample of 4: a complete structure, with only 3 isolated variables.
- It can be convenient to use other procedures as statistical tests. They can provide better results (difficulty to extend to general networks, but useful in classification trees).
- We have given a new score, which can also be used for classification.
- We are working in a new score, with a behaviour similar to statistical tests for small samples and similar to the Bayesian score for large samples.

# Example




Network learned with a Bayesian score and a sample of 4 from 12 independent variables.

# A New Score Based on Imprecise Probability

- It is based on computing a lower and upper value for the score, assuming some variation in the parameters of the Dirichlet model.
- The $S$ parameter will not do too much for small samples.
- For $\alpha_i$ parameters we can not consider all the freedom in Walley IDM. The reason is that if we have observed $Z = z_i$ then the $\underline{P}(D)$ converges to 0 when $\alpha_i \to 0$.
- What we do is to allow an imprecise model in which each $\alpha_i$ verifies $S/(2k) \leq \alpha_i \leq S/(2k) + S/2$.

# A New Score

We can consider the following:

- A global sample size $S$.
- For each unconditional probability distribution about $Z$ all parameters $\alpha_i$ such that $\sum_i \alpha_i = S$ and $S/(2k) \leq \alpha_i \leq S/(2k) + S/2$.
- For each conditional probability $Z|X$ consider all the distributions for $Z$ in which $\sum_i \alpha_i = S_x$ and $S_x/(2k) \leq \alpha_i \leq S_x/(2k) + S_x/2$ where $S_x$ is the parameter associated to $x$ (its $\alpha_x$ if it is an unconditional probability). Assign to $z$ the parameter $S_z = \sum_z \alpha_i$.

# A New Score

- Compute the dominance interval for independence and dependence:

$$\left[\min_{\alpha} \frac{P(D|Dep)}{P(D|Indep)}, \max_{\alpha} \frac{P(D|Dep)}{P(D|Indep)}\right]$$

- Decide for Dependence if the lower limit is greater than 1.0
- Decide for Independence if the upper limit is lower than 1.0
- Non decide if 1.0 is in the interval
- Alternative rule (preferring independence in the case of non-dominance): decide for independence except if the lower limit is greater than 1.0.

# Approximate Method

- The interval can be difficult to compute.
- Consider a constant sample size $S$ for all unconditional and conditional probabilities and divide it between the number of parents configurations of the variable.
- In the case of two variables $X$ and $Y$, the probabilities for independence and dependence are $\sum_i \alpha_{i,j} = \alpha_j$:

$$BSI = \frac{\Gamma(S)}{\Gamma(N+S)} \left( \prod_i \frac{\Gamma(N_X(i)+\alpha_i)}{\Gamma(\alpha_i)} \right) \cdot \frac{\Gamma(S)}{\Gamma(N+S)} \left( \prod_j \frac{\Gamma(N_Y(j)+\alpha_j)}{\Gamma(\alpha_j)} \right)$$

- Dependence

$$BSD = \frac{\Gamma(S)}{\Gamma(N+S)} \left( \prod_i \frac{\Gamma(N_X(i)+\alpha_i)}{\Gamma(\alpha_i)} \right) \cdot \prod_i \frac{\Gamma(S/2)}{\Gamma(N_X(i)+S/2)} \left( \prod_j \frac{\Gamma(N(i,j)+\alpha_{i,j})}{\Gamma(\alpha_{i,j})} \right)$$

# Approximate Method

- The interval can be difficult to compute.
- Consider a constant sample size $S$ for all unconditional and conditional probabilities and divide it between the number of parents configurations of the variable.
- In the case of two variables $X$ and $Y$, the probabilities for independence and dependence are $\sum_i \alpha_{i,j} = \alpha_j$:

$$BSI = \frac{\Gamma(S)}{\Gamma(N+S)} \left( \prod_i \frac{\Gamma(N_X(i)+\alpha_i)}{\Gamma(\alpha_i)} \right) \cdot \frac{\Gamma(S)}{\Gamma(N+S)} \left( \prod_j \frac{\Gamma(N_Y(j)+\alpha_j)}{\Gamma(\alpha_j)} \right)$$

- Dependence

$$BSD = \frac{\Gamma(S)}{\Gamma(N+S)} \left( \prod_i \frac{\Gamma(N_X(i)+\alpha_i)}{\Gamma(\alpha_i)} \right) \cdot \prod_i \frac{\Gamma(S/2)}{\Gamma(N_X(i)+S/2)} \left( \prod_j \frac{\Gamma(N(i,j)+\alpha_{i,j})}{\Gamma(\alpha_{i,j})} \right)$$

# Approximate Method

- The interval can be difficult to compute.
- Consider a constant sample size $S$ for all unconditional and conditional probabilities and divide it between the number of parents configurations of the variable.
- In the case of two variables $X$ and $Y$, the probabilities for independence and dependence are $\sum_i \alpha_{i,j} = \alpha_j$:

$$BSI = \frac{\Gamma(S)}{\Gamma(N+S)} \left( \prod_j \frac{\Gamma(N_Y(j) + \alpha_j)}{\Gamma(\alpha_j)} \right)$$

- Dependence

$$BSD = \prod_i \frac{\Gamma(S/2)}{\Gamma(N_X(i) + S/2)} \left( \prod_j \frac{\Gamma(N(i,j) + \alpha_{i,j})}{\Gamma(\alpha_{i,j})} \right)$$

# Approximation

Independence:

$$BSI = \frac{\Gamma(S)}{\Gamma(N+S)} \left( \prod_j \frac{\Gamma(N_Y(j)+\alpha_j)}{\Gamma(\alpha_j)} \right)$$

Dependence

$$BSD = \prod_i \frac{\Gamma(S/2)}{\Gamma(N_X(i)+S/2)} \left( \prod_j \frac{\Gamma(N(i,j)+\alpha_{i,j})}{\Gamma(\alpha_{i,j})} \right)$$

**Rule:** (to compute the lower value of the interval) Assign $\alpha_{i,j}$ to the lowest value $S/8$ if $N(i,j)$ is maximum (in $j$ for each $i$).

Make $\alpha_{i,j} = S/4$ if $N(i,0) = N(i,1)$

Compute: $\alpha_j = \sum_i \alpha_{i,j}$

# Experiments

Simulated in the same conditions than BDE with $S = 2$ and this score considered for $S = 2$.

| Indep. | $S = 2$ | PRIOR | NEW | Dep. | $S = 2$ | PRIOR | NEW |
|---|---|---|---|---|---|---|---|
| 3 | 33397 | 0 | 0 | 3 | 50006 | 100000 | 100000 |
| 5 | 26910 | 1148 | 1148 | 5 | 51470 | 92202 | 92202 |
| 10 | 20864 | 1234 | 5694 | 10 | 48391 | 80284 | 71045 |
| 20 | 15858 | 1344 | 5490 | 20 | 42826 | 67633 | 57973 |
| 50 | 10467 | 1090 | 4813 | 50 | 34564 | 51458 | 42721 |
| 100 | 7448 | 964 | 4024 | 100 | 28335 | 40192 | 33210 |
| 1000 | 2337 | 351 | 1563 | 1000 | 12584 | 16428 | 13636 |

PRIOR are the results modifying the 'a priori' distribution for Dependence-independence.

We have few errors from independence to dependence for small samples.

For large samples the results are similar to the Bayesian procedure.

# Classification Trees

- They are an important tool for the supervised classification problem.
- Introduced by Quinlan (ID3,C4.5).
- We will apply the upper entropy score to its construction.

# The Supervised Classification Problem

We assume a set of variables or attributes $\mathbf{X} = (X_1, \ldots, X_n)$.
Each variable $X_i$ will take values on a finite set $U_{X_i}$.
We have a class variable $C$, with values in $U_C$.
We have a database of values for these variables:

| $X_1$ | $X_2$ | $\ldots$ | $X_n$ | $C$ |
|---|---|---|---|---|
| $x_1^1$ | $x_2^1$ | $\ldots$ | $x_n^1$ | $c_1$ |
| $x_1^2$ | $x_2^2$ | $\ldots$ | $x_n^2$ | $c_2$ |
| $x_1^3$ | $x_2^3$ | $\ldots$ | $x_n^3$ | $c_3$ |
| $x_1^4$ | $x_2^4$ | $\ldots$ | $x_n^4$ | $c_4$ |

We want to induce a model $M$ such that if $\mathbf{x}$ is a value of $\mathbf{X}$.

$$\mathbf{x} \longrightarrow \boxed{M} \longrightarrow c \in U_C$$

# Notation

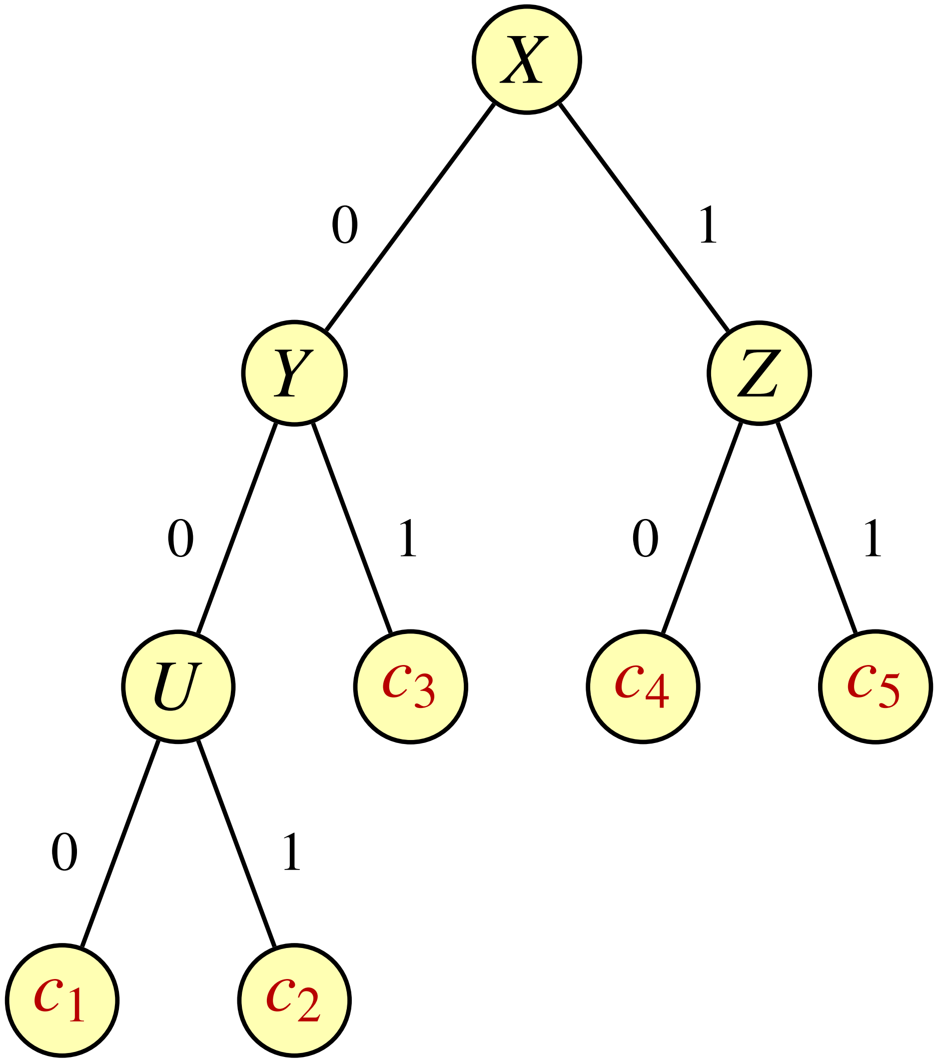
- If $\mathbf{Y} \subseteq \mathbf{X}$ is a subset of the variables, an assignation $\sigma \equiv [\mathbf{Y} = \mathbf{y}]$ is called a Configuration.
- A configuration is complete if assigns values to all the variables in $\mathbf{X}$.
- If $\sigma$ is a configuration and $Z$ is a variable and $z \in U_Z$, then $\sigma(Z = z)$ is the result of adding the value $Z = z$ to $\sigma$.
- If $D$ is a database and $\sigma$ is a configuration, then $D[\sigma]$ is the subset of the database verifying $\sigma$, and it is called the restriction of the database to $\sigma$.

# Difficulties

- Noisy Data.-
- Exponential Number of Configurations.- The training database does not cover all the possible configurations for all the variables. The model should generalize to cases that do not coincide in all the values of all the variables.

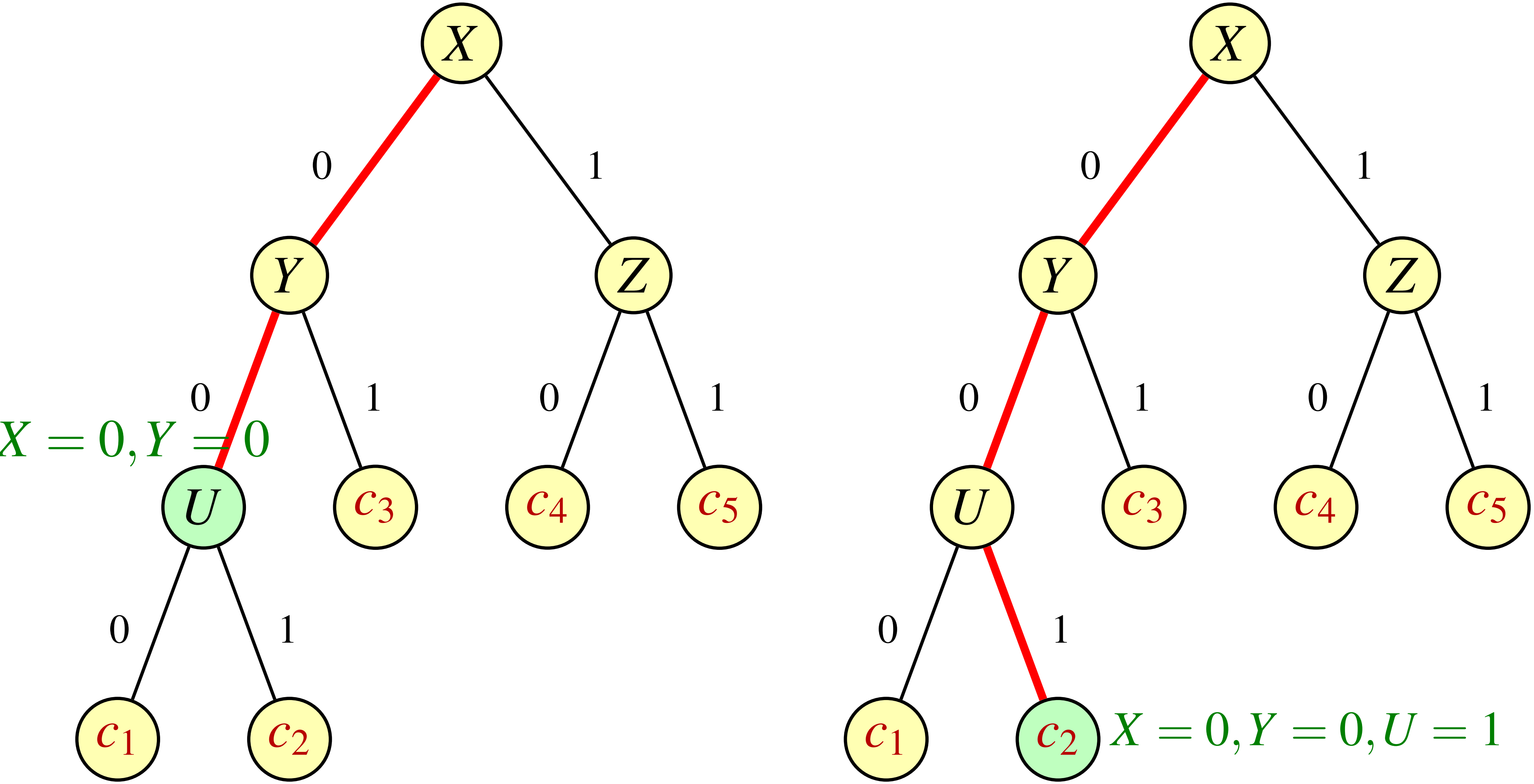
# Classification Trees

A classification tree is a classification model, given by a tree in which each inner nodes represents a variable $Y$ from $\mathbf{X}$.
It has so many children as possible values of $Y$.
Leaf nodes contain a value $c$ of class variable $C$.



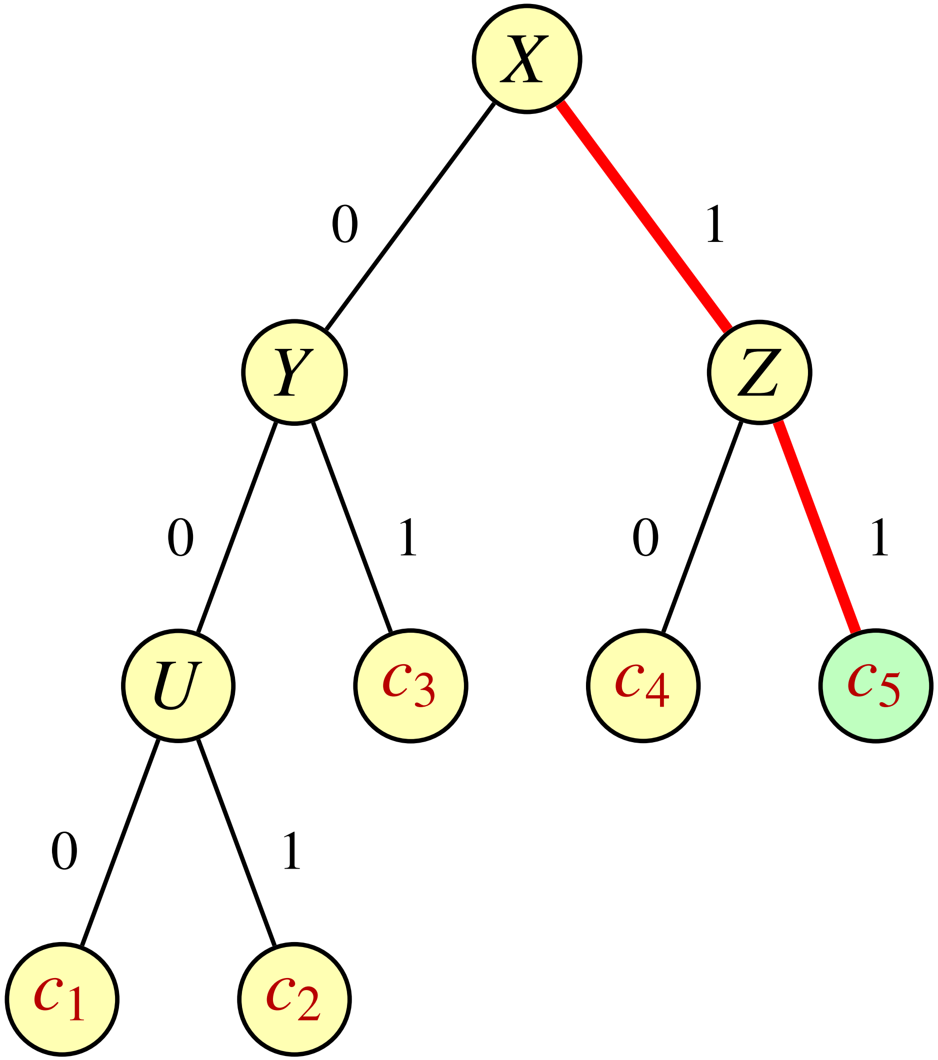
# Nodes and Configurations I

A node defines a configuration:

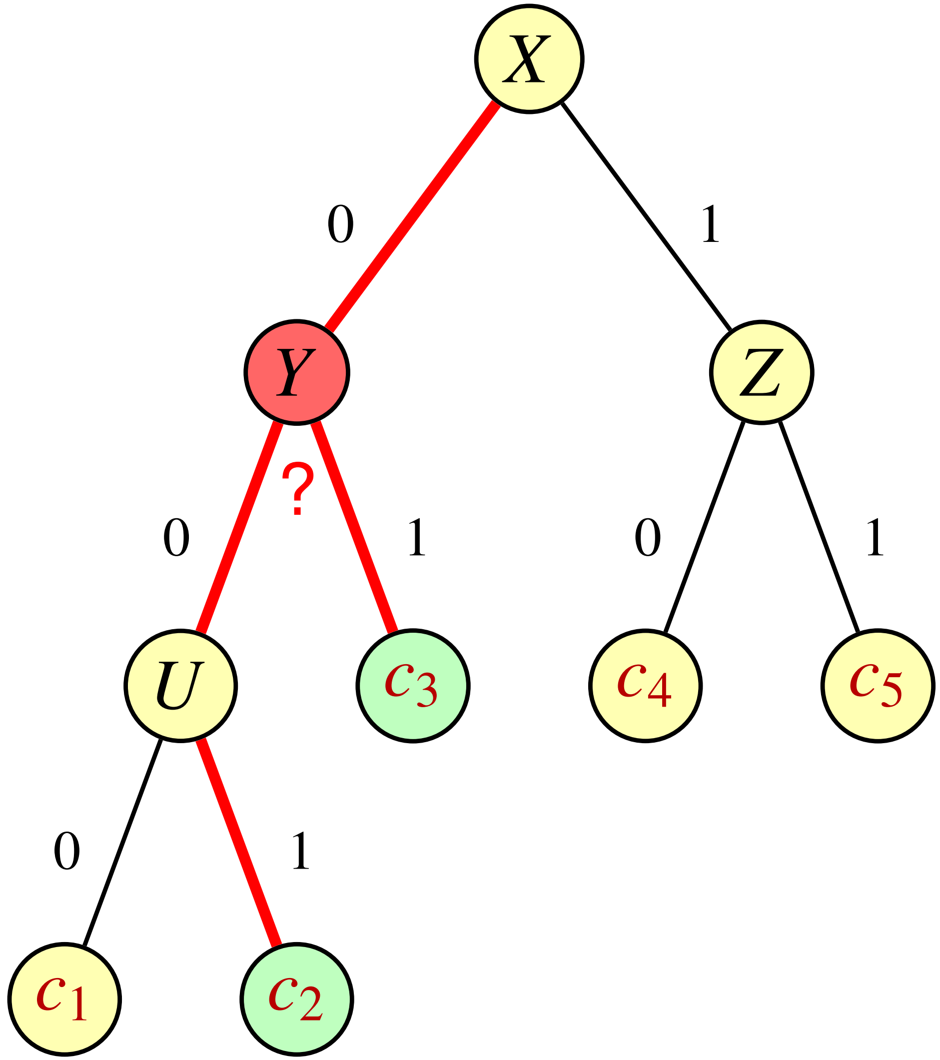


# Nodes and Configurations II

Each complete configuration defines a leaf: $X = 1, Y = 1, Z = 1, U = 0$




Incomplete configurations can be ambiguous: $X = 0, U = 1$



# Building Trees

There are several basic decisions:

- Decide if a node is an inner node or a leaf (model complexity).
- Decide which variable to assign to an inner node.
- Decide which value of $C$ to assign to a leaf node.

# Classic Probabilistic ID3

- Given a configuration, $\sigma$, and a variable $X$, then the frequencies of $X$ in $D[\sigma]$ will be denoted as $N_X^\sigma$. $N_X^\sigma(x)$ will be the number of cases in which $X = x$ in database $D[\sigma]$.
- $\hat{P}_X^\sigma$ will the relative frequencies computed from $N_X^\sigma$:

$$\hat{P}_X^\sigma(y) = \frac{N_X^\sigma(y)}{\sum_x N_X^\sigma(x)}$$

- If $P$ is a probability distribution, its entropy is denoted by $H(P)$.

# Classic Probabilistic ID3

- The decisions are taken node by node, starting by the root node and following by its children (recursively).
- We always decide to branch a node with configuration $\sigma$ except if all the cases in $D[\sigma]$ have the same value for class variable $C$.
- To decide the variable to branch, we consider its configuration $\sigma$ and then, the *information gain* for each variable:

$$I_\sigma(X) = H(\hat{P}_C^\sigma) - \sum_{x \in \Omega_X} \hat{P}_X^\sigma(x).H(\hat{P}_C^{\sigma(X=x)})$$

  We branch by the variable with more information gain.
- In each leaf with configuration $\sigma$, we choose the value of $C$ with highest frequency in $D[\sigma]$.

# Problems

- It has a tendency to overfit the data and has poor behaviour for new cases.
  C4.5 Implements several pruning methods (errors in an additional set of cases, statistical tests, etc..)
- It has a tendency to chose variables with more possible cases.
  C4.5 Considers the relative information gain:

$$RI_\sigma(X) = \frac{I_\sigma(X)}{H(\hat{P}_X^\sigma)}$$

# Example

Imagine that in a node the content of the database for the remaining variables is:

| $X$ | $Y$ | $Z$ | $C$ |
|---|---|---|---|
| 1 | 1 | 1 | $a$ |
| 1 | 2 | 1 | $a$ |
| 0 | 0 | 1 | $a$ |
| 0 | 0 | 0 | $a$ |
| 0 | 0 | 1 | $b$ |
| 0 | 0 | 0 | $b$ |
| 0 | 0 | 0 | $b$ |
| 0 | 0 | 0 | $b$ |
| 0 | 0 | 0 | $b$ |

In a probabilistic approach $X$ is indifferent to $Y$ and preferred to $Z$.

# The Role of Imprecise Probability

When in a node with configuration $\sigma$ we estimate the probabilities of class variable $\hat{P}_C^\sigma$ we use maximum likelihood estimation.

If the sample size is $1$ then $H(\hat{P}_C^\sigma) = 0.0$

Estimating the probabilities of variable $C$ in $D[\sigma]$, we use the Imprecise Dirichlet Model (Walley, 1996) with parameter $S = 1$.

We get an interval for each case $c \in \Omega_C$:

$$\left[ \frac{N_C^\sigma(c)}{\sum_{c' \in U_C} N_C^\sigma(c') + 1}, \frac{N_C^\sigma(c) + 1}{\sum_{c' \in U_C} N_C^\sigma(c') + 1} \right]$$

Let us call this credal set $\mathcal{M}^\sigma$.

# Example

If $C$ has two values $\{a, b\}$, and the absolute frequencies of $C$ in $D[\sigma]$ are $(0, 1)$. The associated probability intervals are:

- For $C = a$: $[0, 0.5]$
- For $C = b$: $[0.5, 1]$

In the probabilistic case we considered a probability with zero entropy, and now we have the maximum entropy probability.
If the absolute frequencies of $C$ are $(2, 7)$. The associated probability intervals are:

- For $C = a$: $[0.2, 0.3]$
- For $C = b$: $[0.7, 0.8]$

# Measuring Uncertainty

Imagine that we have a credal set $\mathcal{M}$ (convex set of probability distributions), we have considered that the uncertainty has two components:

- Entropy $\overline{H}(\mathcal{M}) = \max\{H(P) \mid P \in \mathcal{M}\}$

# Credal Decisions Trees

We take decisions node by node. En each node with configuration $\sigma$, instead of

$$I_{sigma}(X) = H(\hat{P}_C^{\sigma}) - \sum_{x \in U_X} \hat{P}_X^{\sigma}(x).H(\hat{P}_C^{\sigma(X=x)})$$

we compute the imprecise information (Imprecise upper entropy criterion):

$$IMP_{sigma}(X) = \overline{H}(\mathcal{M}^{\sigma}) - \sum_{x \in U_X} \hat{P}_X^{\sigma}(x).\overline{H}(\mathcal{M}^{\sigma(X=x)})$$

Information can be negative!

# Branching?

We have used two criteria two decide whether to branch a node or to make it a leaf node:

- **Simple.-** We branch if there is a variable with positive information.
- **Double.-** We branch if there is a a single or a couple of variables, such that the information is positive after adding them.

$$IM_{sigma}(X,Y) = \overline{H}(\mathcal{M}^{\sigma}) - \sum_{(x,y)\in U_X \times U_Y} \hat{P}_X^{\sigma}(x,y).\overline{H}(\mathcal{M}^{\sigma(X=x,Y=y)})$$

# Selecting a Variable

- We chose the variable or couple of variables with maximum information. If we have selected a couple, then we single out the variable with maximum information of them.

# Decision in the leaves

- In general, to classify a variable in a leaf we use the **dominance criterion**.
- A value $c \in U_C$ is dominated if $\forall P \in \mathcal{M}$ we have that there is a value $c_0 \in U_C$ with $P(c_0) > P(c)$.
- In general we select the non-dominated cases.
- For this particular type of credal sets, $c$ is non-dominated if and only if there in no $c'$ such that $\underline{P}(c') > \overline{P}(c)$. Credal Classification introduced by Zaffalon.
- In some cases, for comparison with C4.5 we assign the value with highest frequency (Frequency Criterion).

# Experiments

| UCI Repository | N. Tr | N. Ts | N. variables | N. classes |
|---|---|---|---|---|
| Breast Cancer | 184 | 93 | 9 | 2 |
| Breast | 457 | 226 | 10 | 2 |
| Heart | 180 | 90 | 13 | 2 |
| Hepatitis | 59 | 21 | 19 | 2 |
| Cleveland nominal | 202 | 99 | 7 | 5 |
| Cleveland | 200 | 97 | 13 | 5 |
| Pima | 512 | 256 | 8 | 2 |
| Vote1 | 300 | 135 | 15 | 2 |
| Australian | 460 | 230 | 14 | 2 |
| Monks1 | 124 | 432 | 6 | 2 |
| Soybean-small | 31 | 16 | 21 | 4 |

# Results Classic Methods

| Data set | NB(Tr) | NB(Ts) | C4.5(Tr) | C4.5(Ts) |
|---|---|---|---|---|
| Breast Cancer | 78.2 | 74.2 | 81.5 | 75.3 |
| Breast | 97.8 | 97.3 | 97.6 | 95.1 |
| Cleveland nominal | 63.9 | 57.6 | 69.3 | 51.5 |
| Cleveland | 78.0 | 50.5 | 73.5 | 54.6 |
| Pima | 76.4 | 74.6 | 79.9 | 75.0 |
| Heart | 87.8 | 82.2 | 83.3 | 75.6 |
| Hepatitis | 96.2 | 81.5 | 96.2 | 85.2 |
| Australian | 87.6 | 86.1 | 89.3 | 83.0 |
| Vote1 | 87.6 | 88.9 | 94.5 | 88.3 |
| Soybean-small | 100 | 93.8 | 100 | 100 |

# Results TU2 (single)

| Data set | Training | UC(Tr) | Test | UC(Ts) |
|---|---|---|---|---|
| Breast Cancer | 89.0 | 16.3 | 93.5 | 17.2 |
| Breast | 99.1 | 2.6 | 98.6 | 2.6 |
| Cleveland nominal | 73.6 | 21.2 | 74.4 | 13.1 |
| Cleveland | 82.6 | 34.0 | 80.3 | 31.9 |
| Pima | 86.6 | 15.6 | 86.2 | 15.2 |
| Heart | 93.9 | 8.8 | 93.8 | 10.0 |
| Hepatitis | 96.4 | 5.0 | 94.7 | 9.5 |
| Australian | 95.3 | 6.5 | 94.4 | 6.5 |
| Vote1 | 98.2 | 5.3 | 98.4 | 4.4 |
| Soybean-small | 100.0 | 0.0 | 100.0 | 0.0 |

# Global Comparison-Frequency Criterion

| Data set | TU2(Ts) | NB(Ts) | C4.5(Ts) |
|---|---|---|---|
| Breast Cancer | 90.3 | 74.2 | 75.3 |
| Breast | 97.8 | 97.3 | 95.1 |
| Cleveland nominal | 75.8 | 57.6 | 51.5 |
| Cleveland | 80.4 | 50.5 | 54.6 |
| Pima | 80.9 | 74.6 | 75.0 |
| Heart | 92.2 | 82.2 | 75.6 |
| Hepatitis | 95.2 | 81.5 | 85.2 |
| Australian | 93.5 | 86.1 | 83.0 |
| Vote1 | 97.8 | 88.9 | 88.3 |
| Soybean-small | 100 | 93.8 | 100 |

# Double Method

| Database | TU2(Ts) | NB(Ts) | C4.5(Ts) |
|---|---|---|---|
| Breast Cancer | 91.4 | 74.2 | 75.3 |
| Breast | 98.7 | 97.3 | 95.1 |
| Cleveland nominal | 74.7 | 57.6 | 51.5 |
| Cleveland | 80.4 | 50.5 | 54.6 |
| Pima | 82.4 | 74.6 | 75.0 |
| Heart | 94.4 | 82.2 | 75.6 |
| Hepatitis | 95.2 | 81.5 | 85.2 |
| Australian | 91.7 | 86.1 | 83.0 |
| Vote1 | 98.5 | 88.9 | 88.3 |
| Soybean-small | 100 | 93.8 | 100 |

| | Simple method | | Double method | |
|---|---|---|---|---|
| Function | Tr | Ts | Tr | Ts |
| TU1 | 81.5 | 80.6 | 94.4 | 91.7 |

# Conclusions

- Imprecise Dirichlet model and total uncertainty provide a good criterion for decision trees complexity (branching decisions).
- It is also more efficient as it is not based on complete expansion and posterior pruning.
- Maximum entropy is better than Maximum entropy + non specificity
- Information of credal sets is a good criterion to select variables
- Credal classification seems appropriate
- The double method is more complex, but the results improve

# Future Work

- Apply this methodology to build Bayesian networks
- Missing values
- To improve the criterion:
  An imprecise model for the marginal + Imprecise model for conditionals is not equivalent to Imprecise model in the joint probability
  There is no symmetry.

# Bibliography (non complete)


- J. Abellán, S. Moral, Building Classification Trees Using the Total Uncertainty Criterion. *International Journal of Intelligent Systems*, Vol. 18 (2003) 1215–1225.
  *A preliminary version of the method to build classification trees.*
- J. Abellán, S. Moral, Maximum of Entropy for Credal Sets. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 11 (2003) 587–597.
  *Properties of upper entropy.*
- S. Moral. An Empirical Comparison of Score Measures for Independence. En: *Proceedings of the Tenth International Conference IPMU 2004*, Vol. 2, Perugia, Italia (2004) 1307–1314.
  *The comparison of measures for independence, including the one based on upper entropy.*
- J. Abellán, S. Moral, Maximum of Entropy in Credal Classification. *Proceedings Isipta '03* (2003) 1–15. An extended version is submitted to: *International Journal of Approximate Reasoning*. *The procedure to build classification trees based on maximum entropy.*