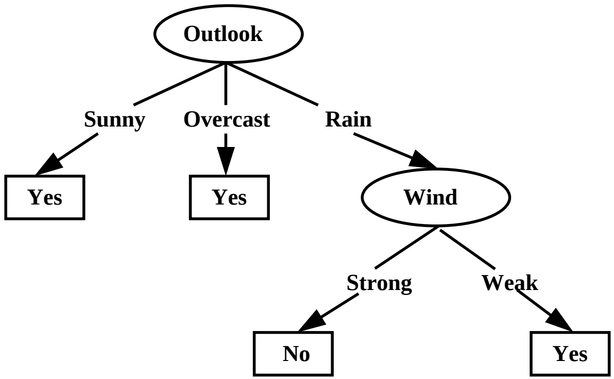
Bias Correction in Classification Tree Construction
ICML 2001

Alin Dobra   *Johannes Gehrke*
Department of Computer Science
Cornell University

December 15, 2001

Classification Tree Construction

Outlook	Temp.	Humidity	Wind	Play Tennis?
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No



Motivating Examples

Experiment 1:

- Two class labels equally probable: 1 and 2.
- Two predictor variables X_1 with 10 possible values and X_2 with 2 possible values. Both variables *uncorrelated* with the class label.
- Random datasets with $N = 100$ data-points.
- Split criteria: `gini` gain (Breiman 1984).

$$\Delta g \stackrel{\text{def}}{=} \sum_{i=1}^{n} P[X=x_i] \sum_{j=1}^{k} P[C=c_j|X=x_i]^2 - \sum_{j=1}^{k} P[C=c_j]^2$$

Result: X_1 is chosen 80 times more often than X_2.

Conclusion: `gini` gain *biased* towards predictor variables with more categories.

Motivating Examples(cont.)

Experiment 2:

- Same setup as before but X_2 slightly correlated:

$$P[C = 1|X_2 = x_{21}] = 0.51$$

- Experiments for $N = 100$ and $N = 1000$ training data-points

Result:

N	odds X_1 vs X_2
100	62:1
1000	32:1

Conclusion: `gini` gain chooses the *wrong* predictor variable to split on with high probability if only *weak correlations* are present.

Outline of the Talk

- Motivation
- Bias in split variable selection
  - Formal definition of the bias.
  - Experimental demonstration of the bias.
- Correction of the bias
  - General method for bias removal.
  - Correction of the bias of `gini` gain.
  - Explanation of the bias of `gini` gain.
  - Experimental evaluation.
- Summary and future work.

Bias in Split Selection

- How do we define this bias formally?

Null Hypothesis (H_0): Class labels are independent of predictor variables and come from pure coin flips with a multi-face coin with probabilities $p_1, \cdots, p_k$.

Notation:

- $\mathcal{D}$: random dataset distributed according to H_0.
- $s(\mathcal{D}, X)$: value of split criterion s when applied to dataset $\mathcal{D}$ and predictor variable X.

Definition:

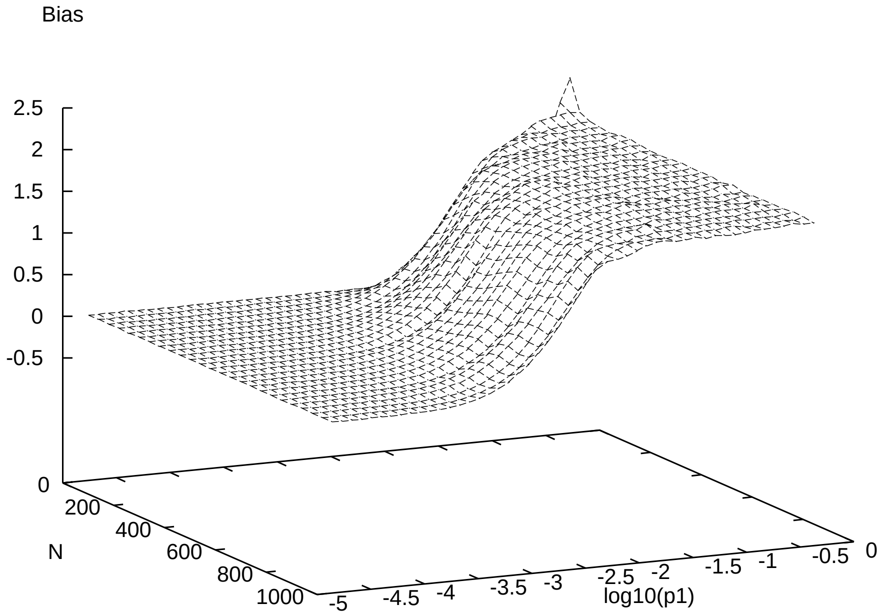
$$\text{Bias}(X_1, X_2) = \log_{10}\left(\frac{P[s(\mathcal{D}, X_1) > s(\mathcal{D}, X_2)]}{1 - P[s(\mathcal{D}, X_1) > s(\mathcal{D}, X_2)]}\right)$$

Experimental Setup

- Synthetic datasets generated according to H_0.
- Two predictor variables: X_1 with domain size $n_1 = 10$ and X_2 with domain size $n_2 = 2$.
- Number of data-points N between 10 and 1000.
- p_1 between 0 and 1/2.
- 100000 Monte Carlo trials to estimate $P[s(\mathcal{D}, X_1) > s(\mathcal{D}, X_2)]$.
- Exactly the same instances used for all split criteria.
- Fair coin used to break ties.

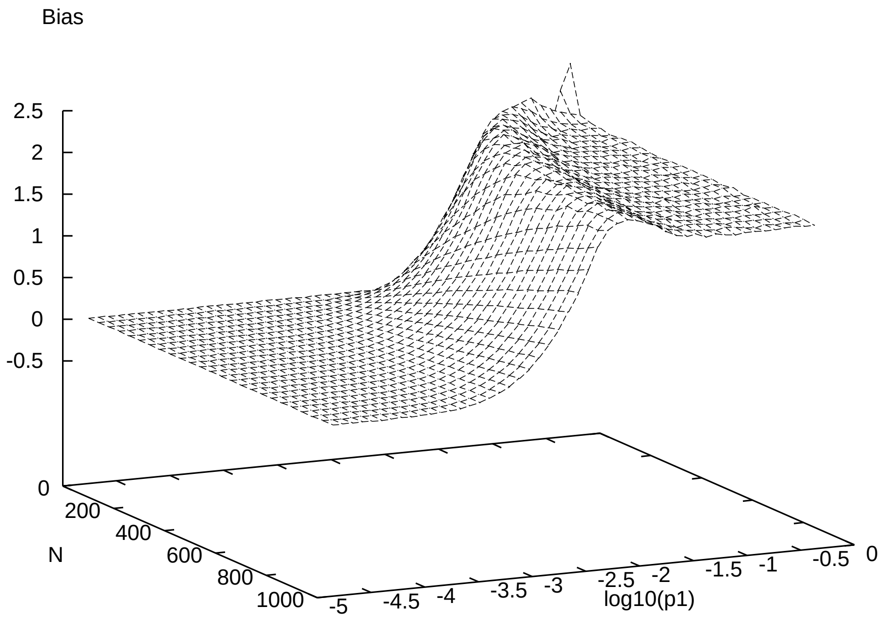
Experimental Bias of Gini Gain

$$\Delta g \stackrel{\text{def}}{=} \sum_{i=1}^{n} P[X=x_i] \sum_{j=1}^{k} P[C=c_j|X=x_i]^2 - \sum_{j=1}^{k} P[C=c_j]^2$$



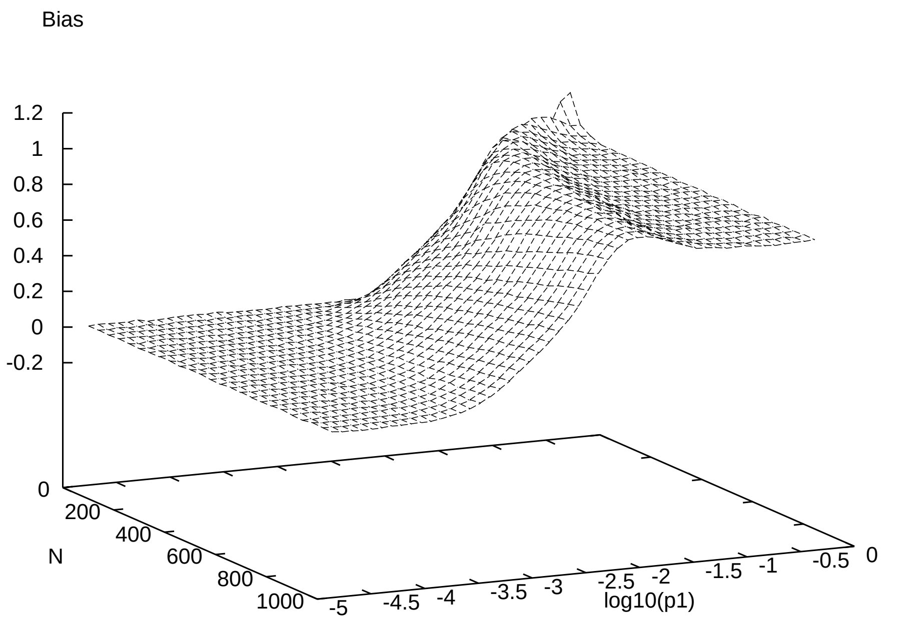
Experimental Bias of Information Gain

$$IG \stackrel{\text{def}}{=} \sum_{j=1}^{k} \Phi(P[C=c_j]) + \sum_{i=1}^{n} \Phi(P[X=x_i]) - \sum_{j=1}^{k}\sum_{i=1}^{n} \Phi(P[C=c_j \wedge X=x_i]),\ \Phi(p) = -p\log p$$



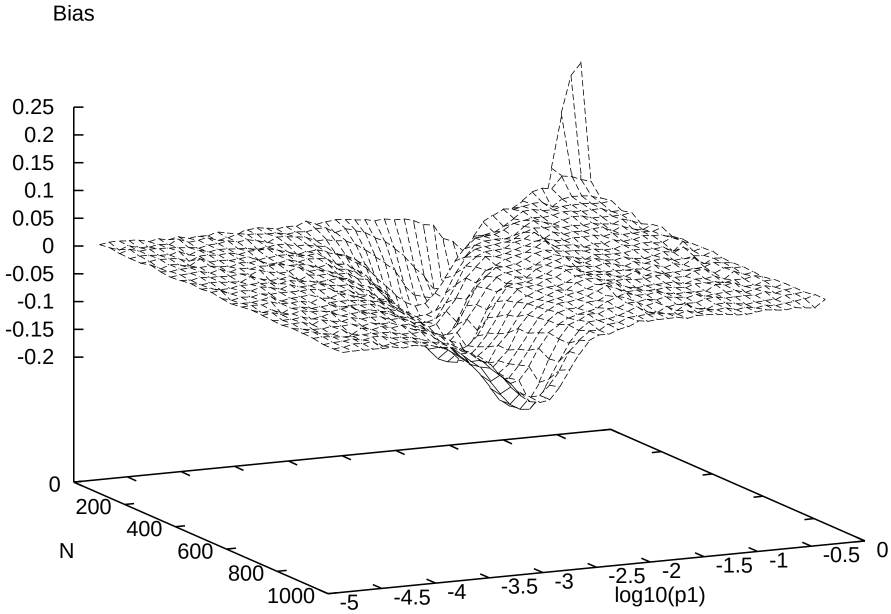
Experimental Bias of Gain Ratio

$$GR \stackrel{\mathrm{def}}{=} \frac{IG}{\sum_{i=1}^{n} \Phi(P[X=x_i])}$$



Experimental Bias of χ^2-test

$$\chi^2 \stackrel{\text{def}}{=} \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{(A_{ij} - E(A_{ij}))^2}{E(A_{ij})}, \quad E(A_{ij}) = \frac{N_i S_j}{N}$$



Outline of the Talk

- Motivation
- Bias in split variable selection
  - Formal definition of the bias.
  - Experimental demonstration of the bias.
- Correction of the bias
  - General method for bias removal.
  - Correction of the bias of `gini` gain.
  - Explanation of the bias of `gini` gain.
  - Experimental evaluation.
- Summary and future work.

General Method for Bias Removal

Definition: The **p-value** of some observation x is the probability that randomly (according to H_0) a value at least as big as x is observed. Small p-value is proof against H_0.

Lemma: The p-value of any split criteria is unbiased if no value of the split criteria has significant probability according to H_0, irrespective of the tie-breaking strategy used.

Computation of the p-value of a Split Criteria

- Exact computation. Very expensive; efficient only for $n = 2$ and $k = 2$ (Martin 1997).
- Bootstrapping (Monte Carlo simulation) (Frank & Witten, 1998). Expensive for small p-values.
- Asymptotic approximations. E.g., χ^2-distribution approximation of the distributions of χ^2-test (Kass, 1980). Inaccurate for border conditions like small entries in the contingency table.
- Tight approximations. Approximate the distribution of the criterion with a parametric distribution such that the approximation is accurate everywhere.

Tight Approximation of the Gini Gain P-value

Can show theoretically:

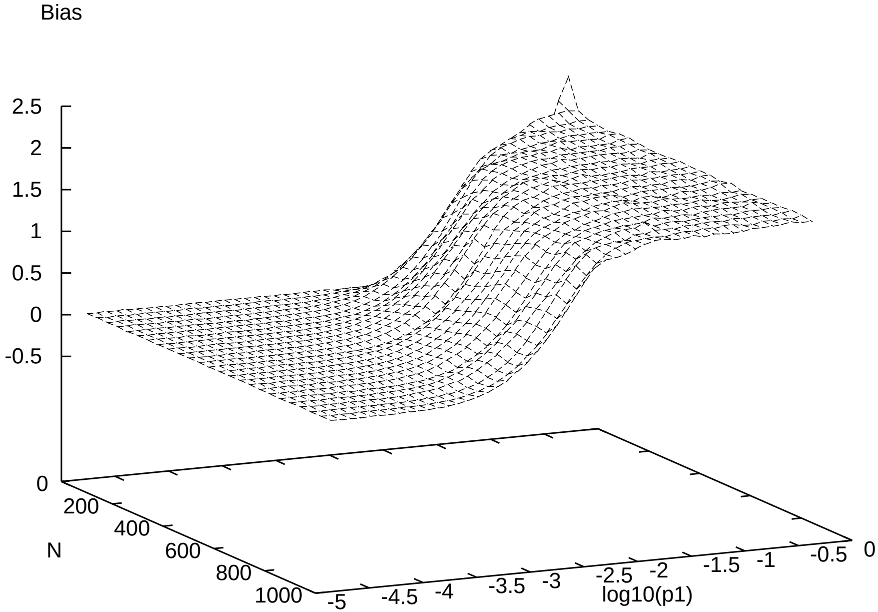
$$E(\Delta g) = \frac{n-1}{N}\left(1 - \sum_{j=1}^{k} p_j^2\right), \quad \text{Var}\,(\Delta g) \approx \frac{n-1}{N^2} f(p_j)$$

Experimental observation: Distribution of `gini` gain well approximated by a Gamma distribution (with cumulative distribution function Q).

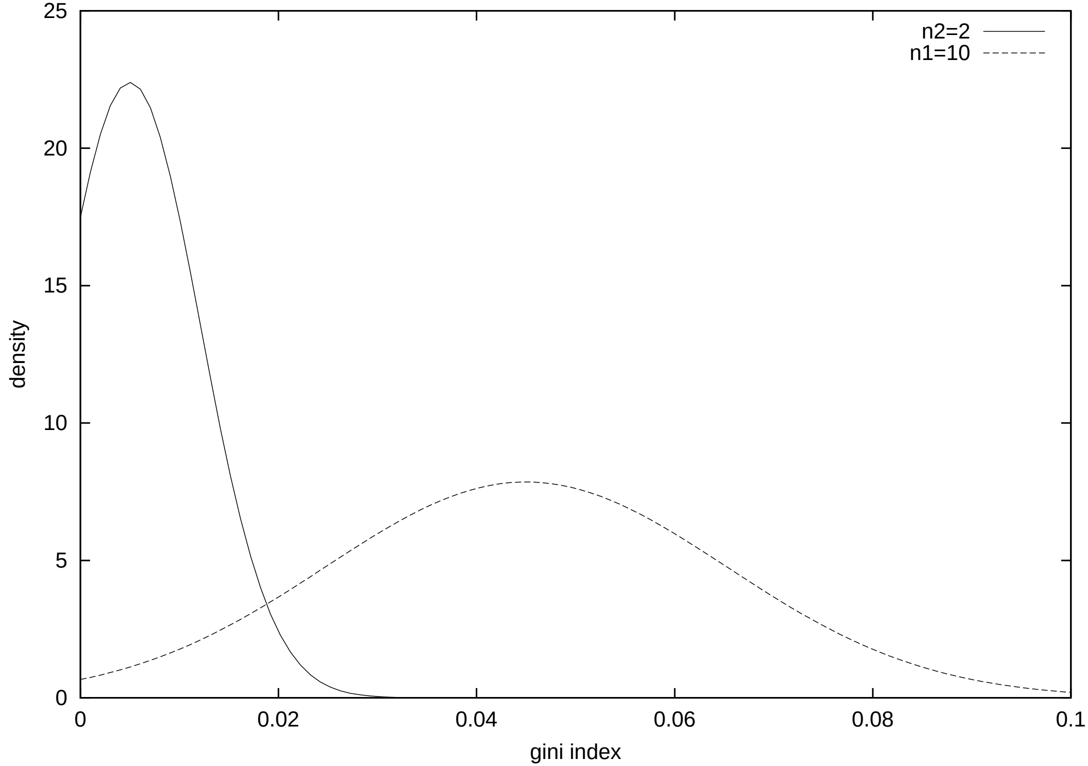
New split criteria:

$$\text{p-value}(\Delta g_e) = 1 - Q\left(\frac{E(\Delta g)^2}{\text{Var}\,(\Delta g)}, \frac{\Delta g_e \text{Var}\,(\Delta g)}{E(\Delta g)}\right)$$

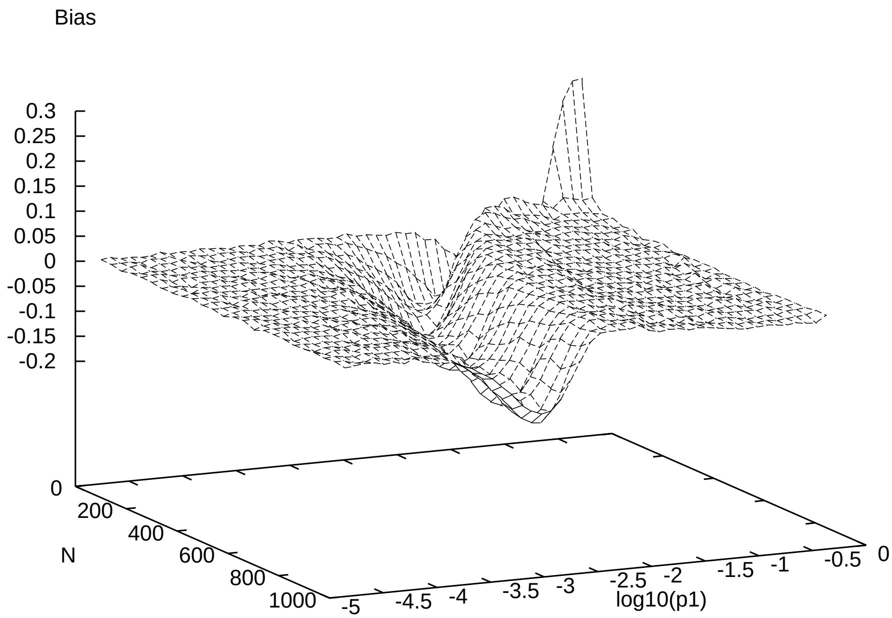
Experimental Bias of Gini Gain



Explanation of the bias of `gini` gain



Experimental Evaluation of the p-value of `gini` index



Summary and Future Work

- Defined bias as log-odds and showed experiments for different split criteria
- General method for bias removal: use p-value of existing criteria
- Corrected the bias of `gini` gain and showed experiments that the correction is successful
- Future work:
  - Analyze experimentally and theoretically the correction for the correlated case.
  - Find correction for CART type binary splits.