

# Knowledge Discovery and Data Mining

**Computer Science 591Y**

Department of Computer Science  
University of Massachusetts Amherst

February 17, 2005



# Topics

---

- Example — Predicting water demand
- Evaluation functions
  - For classification
  - For regression, dependency finding, clustering
  - Assumptions and issues
- Homework #3

**Example**

---

# Predicting water demand

- In the 1990s, the **St. Louis County Water Company** supplied water to tens of thousands of St. Louis residents.
- They were a **regulated utility** whose rates were set by a government body, the **Missouri Public Service Commission**.
- In the early 1990s, the staff of the Commission devised **a new way to predict water demand**, and thus set rates.

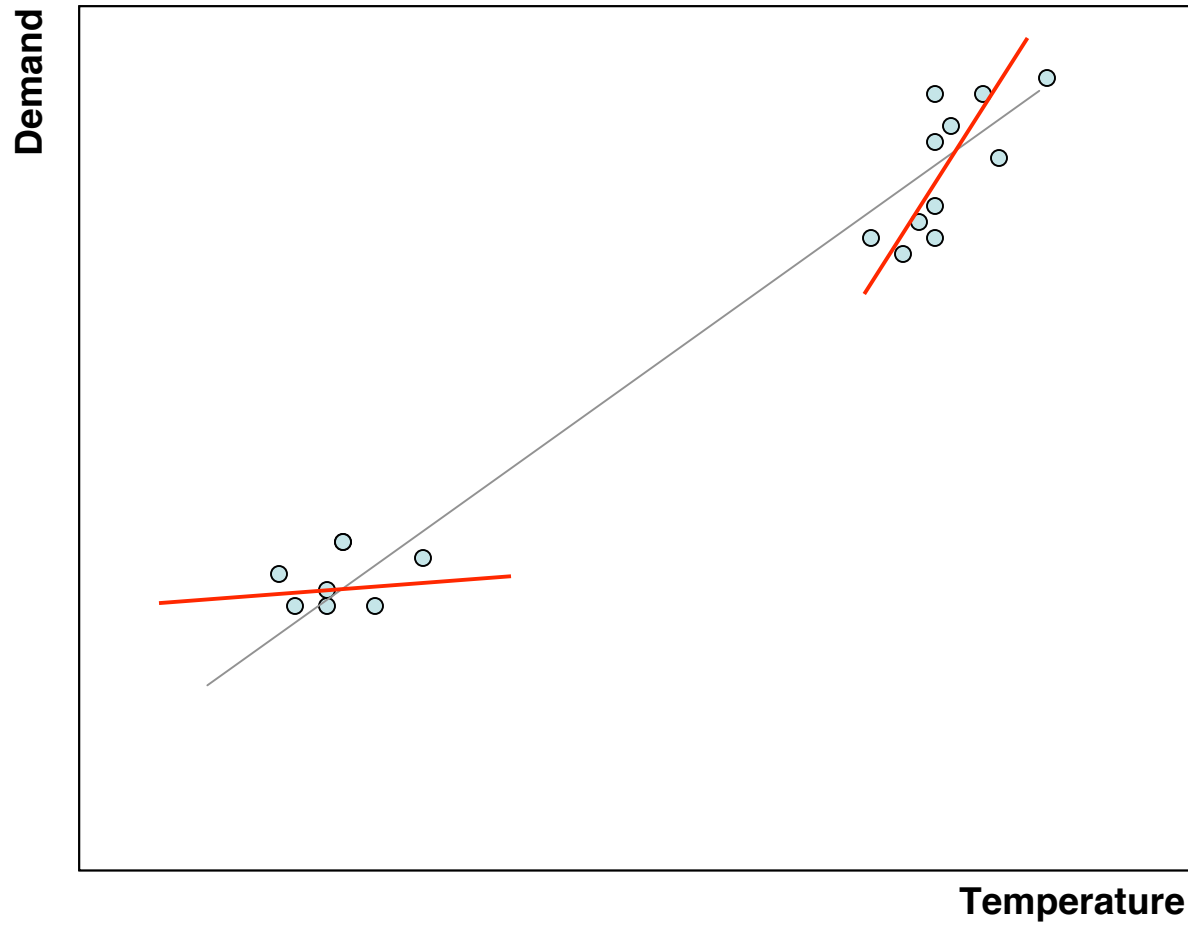
# The Task and a Problem

---

- Task description
  - Analyze the relationship between
    - Daily aggregate water demand and...
    - Several weather-related variables
  - Determine "average" weather for a year
  - Use the predictive model and "normal" weather to estimate "normal" demand
  - Set rates with "normal" demand so the company makes a reasonable profit in an average year
- Problem: Rates dropped substantially

# Unexpected findings

---



# Evaluation functions

---

# Evaluation functions in context

- Task description
- Data representation
- Knowledge representation
- Evaluation function
- Search technique
- Inference method

# Uses of evaluation functions

---

- Guides the search *inside of* algorithms
  - Selects which path to pursue in heuristic search
  - Decides when to stop searching
  - Identifies which of the set of discovered models or patterns should be output
- Evaluates the results *outside of* algorithms
  - Informs analysts about the absolute quality of a discovered model or pattern
  - Helps analysts compare the relative quality of different algorithms or outputs

# Contingency tables

---

		Actual		
		+	-	
Predicted	+	49	6	55
	-	11	34	45
		60	40	100

# Contingency tables

---

		Actual			
		0	0.5	1	
Predicted	0	45	4	0	49
	0.5	5	27	5	37
	1	0	1	25	26
		50	32	30	112

# Simple measures on tables

- True positive rate (TPR) =  $TP / (TP + FN)$
- False positive rate (FPR) =  $FP / (FP + TN)$
- Recall =  $TP / (TP + FN) = TPR$
- Precision =  $TP / (TP + FP)$
- Specificity =  $TN / (FP + TN)$
- Sensitivity = TPR

		Actual	
		+	-
Predicted	+	TP	FP
	-	FN	TN

# Expected values in cells

---

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

$$N = a + b + c + d$$

		Actual	
		+	-
Predicted	+	a	b
	-	c	d

$$E(a) = p(\text{predictedpos}) \cdot p(\text{actualpos} \mid \text{predictedpos}) \cdot N$$

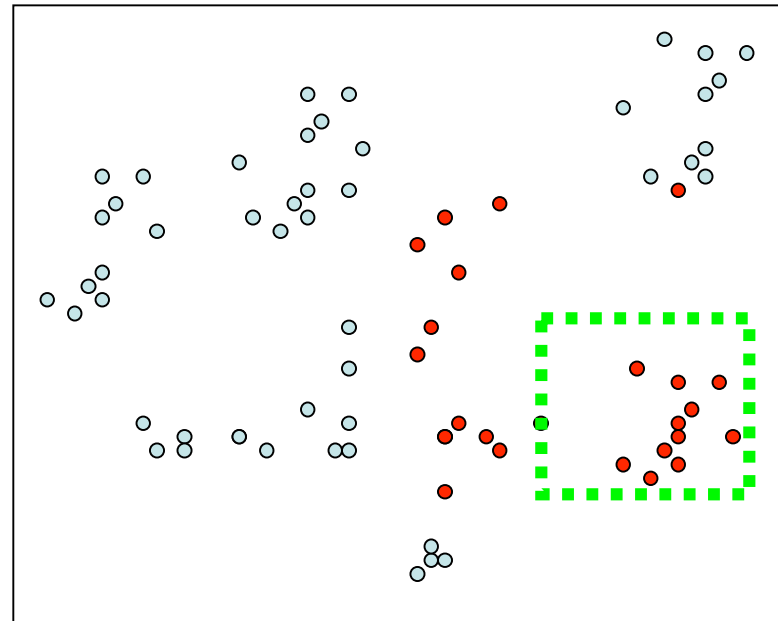
$$E(a) = p(\text{predictedpos}) \cdot p(\text{actualpos}) \cdot N$$

$$E(a) = ((a + b)/N) \cdot ((a + c)/N) \cdot N$$

# Movement in the table

---

		Actual	
		+	-
Predicted	+	a	b
	-	c	d



# Chi-square statistic

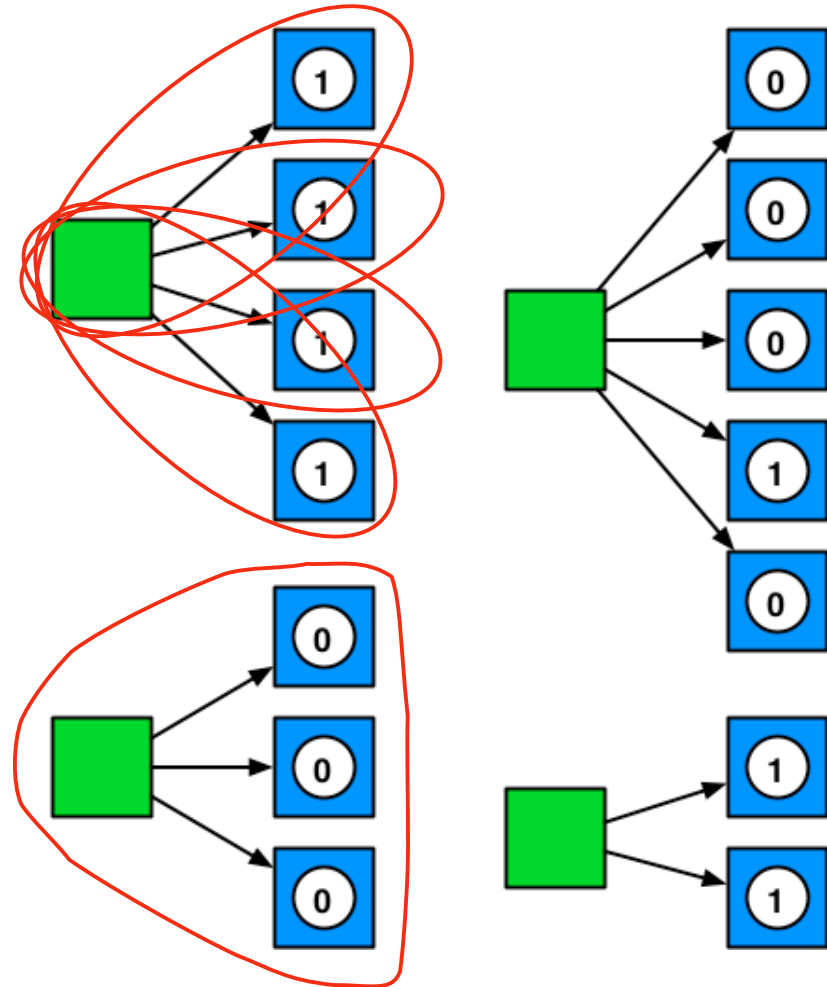
---

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

	Polio	No Polio	Polio	No Polio
Vaccine	57	200,688	99.6	201,129.4
Placebo	142	201,087	99.4	200,645.6
	Observed		Expected	

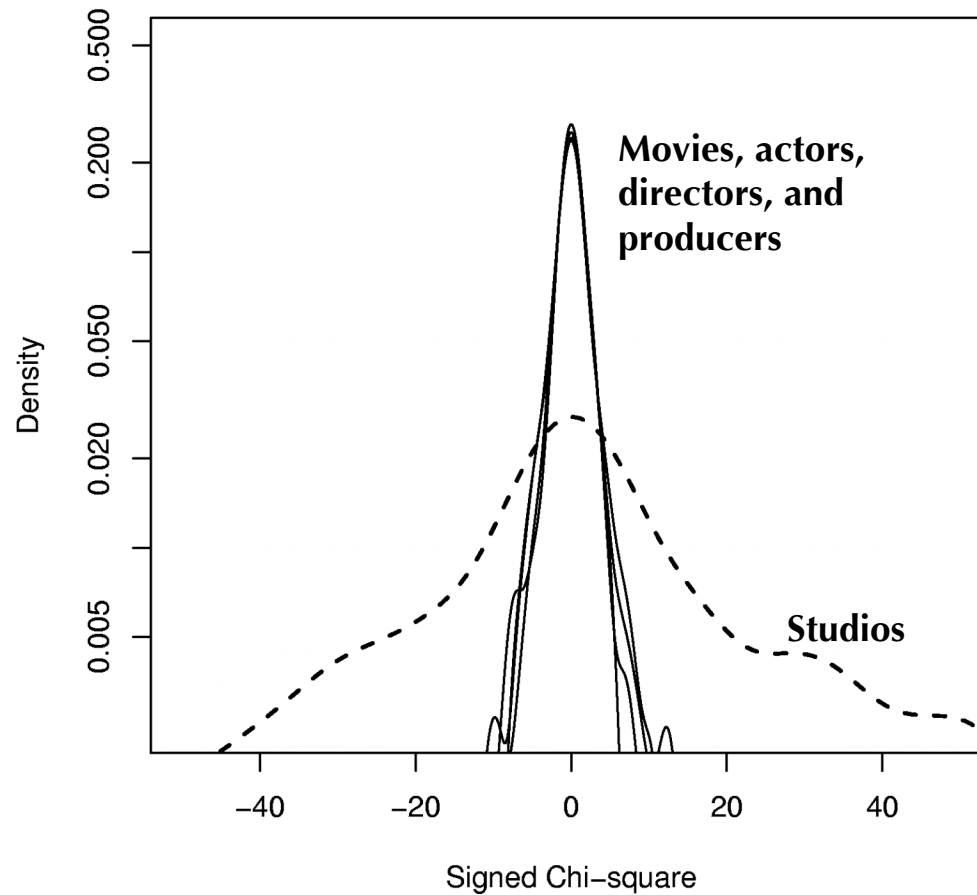
# Assuming independence

- The confidence of any statistical association varies with sample size (N)
- Consider evaluating the association between characteristics of **groups** and their **members**
- What is the "effective" sample size?
  - $N = |\text{members}|$
  - $N = |\text{groups}|$
  - $|\text{members}| \geq N \geq |\text{groups}|$



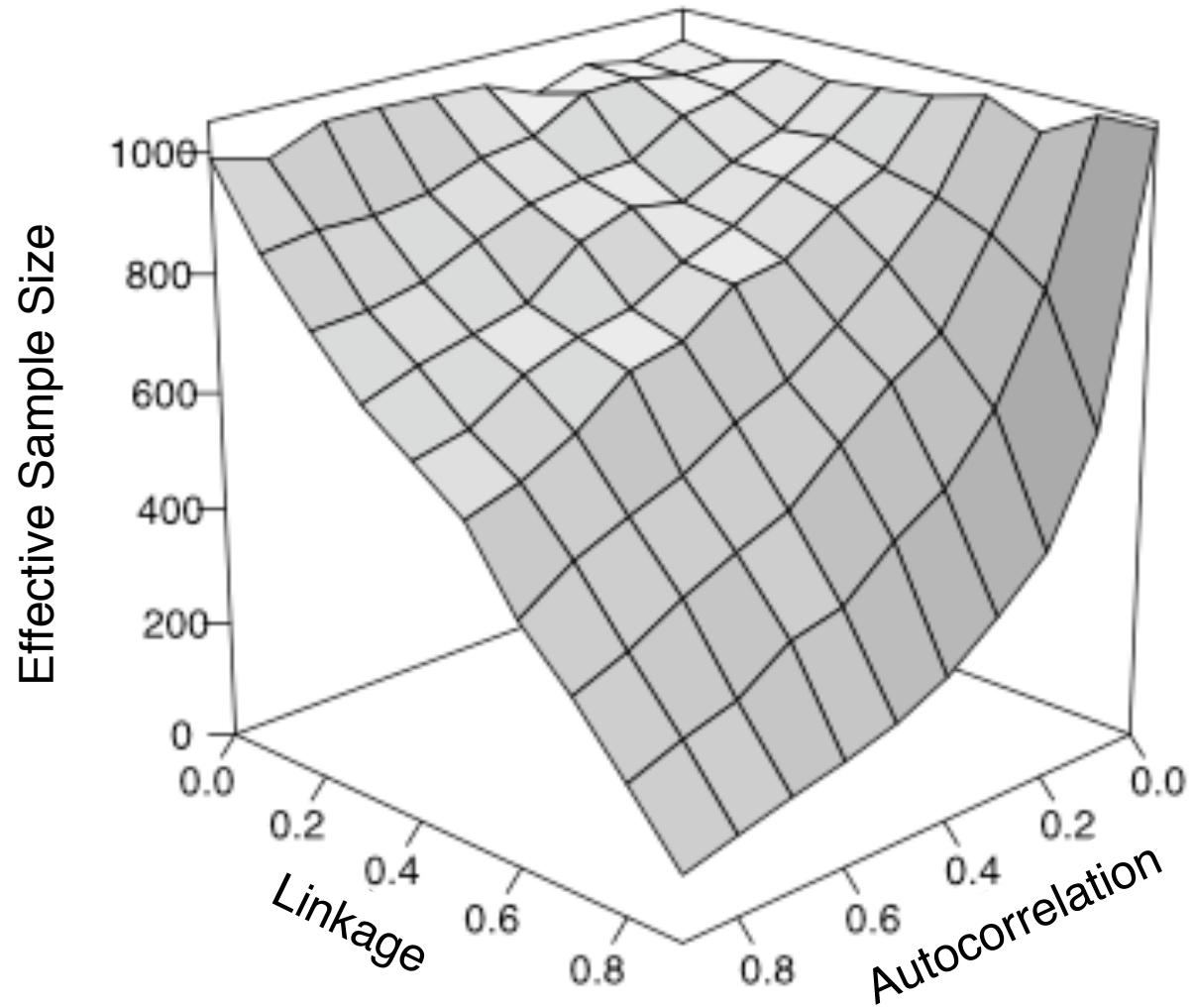
# Differing variance of feature scores

---



# Effective sample size

---



## Other issues

---

- Why not just use accuracy?
- Are the different "roles" of evaluation functions compatible?
- How do the issues with evaluation functions for classification generalize to other tasks?
  - Regression
  - Dependency finding
  - Clustering

# Homework

---

# Homework #3

(Due February 24)

- Identify, obtain, and prepare a data set where  $D \geq 500$ .  
 $D = NV - M$ , where  $N$  is the number of instances (rows),  $V$  is the number of variables (columns),  $M$  is the total number of missing data points,  $N \geq 40$ , and  $V \geq 4$ . Instance identifiers (e.g., name, unique code number, etc) do not count toward  $V$ .
- For example, if you analyzed the House races in 2004 by looking at the party of the winner, the margin of victory, the party split in the district, whether there was an opponent, and campaign spending, then  $N=435$ ,  $V=5$ , and  $M$  might be as large as 100 (because of unavailable data), then  $D \geq 2075$ .

## Homework #3 (continued)

---

- Write a one-page description of the data, suitable for distribution to the rest of the class. Submit as a PDF file.
- Download and install either the *Weka* toolkit or the *R* package (see the web page).
- Make a preliminary analysis of your dataset by constructing a classification tree, linear regression equation, or Bayesian network, or by conducting a cluster analysis.
- Write a one-page description of what the analysis indicates about your data. Make reference to specific portions of your model or patterns, and provide your qualitative judgment of the success of your first effort at analysis.