Empirische Methoden für Künstliche Intelligenz

Introduction

Sebastian Koralewski, Simon Stelter

Institute for Artificial Intelligence
Universität Bremen

9th April, 2018

Outline

General Info

- Lecturers: Sebastian, Simon (PhD students at IAI)
- Language: German and English
- Credits: 4 ECTS (2 SWS)
- Course type: block seminar
- Course number: 03-BE-710.98e
- Location: TAB Building, Room 0.31 EG

Course Content

Based on the book
"Empirical Methods for AI", Paul Cohen, MIT Press, 1995.

- What are empirical methods?
- Why use them?
- Experiment design
- Data analysis

What does “empirical” mean?

- Relying on observations, data, experiments
- Empirical work should complement theoretical work
  - Theories often have holes
  - Theories are suggested by observations
  - Theories are tested by observations
  - Conversely, theories direct our empirical attention

Why We Need Empirical Methods?

Cohen, 1990 Survey of 150 AAAI Papers

- Roughly 60% of the papers gave no evidence that the work they described had been tried on more than a single example problem.
- Roughly 80% of the papers made no attempt to explain performance, to tell us why it was good or bad and under which conditions it might be better or worse.
- Only 16% of the papers offered anything that might be interpreted as a question or a hypothesis.
- Theory papers generally had no applications or empirical work to support them, empirical papers were demonstrations, not experiments, and had no underlying theoretical support.
- **The essential synergy between theory and empirical work was missing**

Theory, not Theorems

Theory based science need not be all theorems.
Otherwise science would be mathematics.

- Computer programs are formal objects
  $\rightarrow$ so let us reason about them entirely formally?
- Two reasons why we cannot or will not:
  - theorems are hard
  - some questions are empirical in nature
    e.g. even though our problem is intractable in general, are there instances met in practice that are easy to solve?

Empirical Computer Science / AI

- Treat computer programs as natural objects
  like fundamental particles, chemicals, living organisms
- Build (approximate) theories about them:
  - construct hypotheses
  - test with empirical experiments
  - refine hypotheses and modelling assumptions

Empirical Computer Science / AI [2]

Advantage over Other Sciences

- Cost: no need for expensive super-colliders
- Control: unlike the real world, we often have complete command of the experiment
- Reproducibility: in theory, computers are entirely deterministic
- Ethics: no ethics panels needed before you run experiments

Experimental Life Cycle

1. Exploration
2. Hypothesis construction
3. Experiment
4. Data analysis
5. Drawing of conclusions
6. Go back to step 1.

Evaluation begins with claims

- The most important, most immediate and most neglected part of evaluation plans
- What you measure depends on what you want to know, on what you claim
- Claims:
  - X is bigger/faster/stronger than Y
  - X varies linearly with Y in the range we care about
  - X and Y agree on most test items
  - Frequency distributions (how often events happen) are equal
  - Classification variables such as smoking history and heart disease history are unrelated
  - It doesn't matter who uses the system (no effects of subjects)
  - My algorithm scales better than yours (e.g., a relationship between size and runtime depends on the algorithm)
- **Non-claim**: I built it and it runs fine on some test data

Always run pilot experiments

- A pilot experiment is designed less to test the hypothesis than to test the experimental apparatus to see whether it can test the hypothesis.
- Use pilot experiments to adjust independent and dependent measures, see whether the protocol works, provide preliminary data to try out your statistical analysis, in short, test the experiment design.

Explain the variance

- The job of empirical science is to explain why things vary, to identify the factors that cause things to be different
- High variance usually means a causal factor has a sizeable effect and is being ignored
- High variance is an opportunity to learn something, not a pest to be bludgeoned with data

The logic of hypothesis testing

- Example: toss a coin ten times, observe eight heads. Is the coin fair (i.e., what is its long run behavior?) and what is your residual uncertainty?
- You say, "If the coin were fair (*null hypothesis*), then eight or more heads is pretty unlikely, so I think the coin is not fair."
- Like proof by contradiction: Assert the opposite (the coin is fair) show that the sample result ($\geq$ 8 heads) has low probability p, reject the assertion, with residual uncertainty related to p.
- Estimate p with a sampling distribution.

Checklist for Experiment Design

1. Consider the experimental procedure
   - making it explicit helps to identify spurious effects and sampling biases
2. Consider a sample data table
   - identifies what results need to be collected
   - clarifies dependent and independent variables
   - shows whether data pertain to hypothesis
3. Consider an example of the data analysis
   - helps you to avoid collecting too little or too much data
   - especially important when looking for interactions

Guidelines for Experiment Design

- Consider possible results and their interpretation
  - may show that experiment cannot support/refute hypotheses under test
  - unforeseen outcomes may suggest new hypotheses
- What was the question again?
  - easy to get carried away designing an experiment and lose the BIG picture
- Run a pilot experiment to calibrate parameters

Bernard Moret's guidelines

Hallmarks of a good experimental paper

- clearly defined goals
- large scale tests: both in number and size of instances
- mixture of problems: real-world, random, standard benchmarks, ...
- statistical analysis of results
- reproducibility: publicly available instances, code, data files, ...

Bernard Moret's guidelines

Pitfalls for experimental papers

- simpler experiment would have given same result
- result predictable by (back of the envelope) calculation
- bad experimental setup: e.g. insufficient sample size, no consideration of scaling, ...
- poor presentation of data: e.g. lack of statistics, discarding of outliers, ...

Bernard Moret's guidelines

Ideal experimental procedure

- define clear set of objectives: which questions are you asking?
- design experiments to meet these objectives
- collect data: do not change experiments until all data is collected to prevent drift/bias
- analyse data: consider new experiments in light of these results

David Johnson's guidelines

- Provide explanations and back them up with experiment
  - adds to credibility of experimental results
  - improves our understanding of algorithms
  - leading to better theory and algorithms
  - can "weed" out bugs in your implementation!
- Ensure comparability (and give the full picture)
  - make it easy for those who come after to reproduce your results
  - provide meaningful summaries: give sample sizes, report standard deviations, plot graphs
  - do not hide anomalous results
  - report running times even if this is not the main focus

Outline

Topic 1: Basic Issues in Experiment Design

- Chapter 3 from the book
- Content:
  - manipulation and observation experiments
  - control, extraneous and noise variables: does x really cause y?
  - ceiling and floor effects: the result is 95%, is that good or bad?
  - sampling bias: more accidents with motorbikes than with cars, is it really the vehicle's fault?

  $\rightarrow$ very simple content and very well explained
- Pages: 37

Topic 2: Hypothesis Testing and Estimation

- Chapter 4 from a book
- Content:
  - hypothesis testing definition, null hypothesis
  - obtaining sampling distributions
  - z test, t test
  - how big should samples be?

  → relatively simple content, very well explained, essential for empirical evaluation
- Pages: 42

Topic 3: Computer-Intensive Statistical Methods

- Chapter 5 from a book
- Content:
  - Monte-Carlo sampling
  - bootstrap sampling
  - randomization tests, randomization version of the t test
  - extracting general knowledge about the population of samples

  $\rightarrow$ math-intensive, well explained, very often used
- Pages: 37

Topic 4: Performance Assessment

- Chapter 6 from a book
- Content:
  - analysis of variance
  - pairwise comparisons: comparing multiple systems
  - cross-validation
  - which performance measures to choose
  - tactics on performance assessment

  → very well explained, very often used, standard procedures in machine learning
- Pages: 51

Topic 5: Trade-offs in Automated Negotiations

- Journal paper, Artificial Intelligence, 2002
- *Using similarity criteria to make issue trade-offs in automated negotiations*, **Faratin, Peyman, Carles Sierra, and Nicholas R. Jennings**, artificial Intelligence 142.2 (2002): 205-237.
- Content:
  - from the area of game theory
  - automated agents negotiating trade-offs
  - making offers to provide services and making trade-offs over price
  - evaluation in a range of negotiation scenarios
  - independent and dependent variables explicitly specified
  - 4 hypothesis to be proven

  $\rightarrow$ many formal definitions, very good example of empirical evaluation
- Pages: 28 (33)

Topic 6: Monitoring Strategies for Embedded Agents

- Journal paper, Adaptive Behavior, 1996
- *Monitoring strategies for embedded agents: Experiments and analysis*, **Atkin and Cohen**, Adaptive Behavior 4.2 (1996): 125-172.
- Content:
  - from the area of robotics
  - embedded agents monitoring their environment
  - different monitoring strategies, pros and cons of each
  - series of experiments with different monitoring strategies: 5 different genetic algorithm methods, about 10 tasks and environments, thousands of trials
  - unexpected results during hypothesis proving and explanations thereof
  - comparing agents with humans

  $\rightarrow$ nicely explained evaluation
- Pages: 33 (48)

Topic 7: Understanding Planner Behavior

- Journal paper, Artificial Intelligence, 1995
- *Understanding Planner Behavior*, **Howe and Cohen**, Artificial Intelligence 76.1-2 (1995): 125-166.
- Content:
  - from the area of planning
  - analyzing planner behavior to find reasons for failures and success
  - identifying interesting patterns of behavior
  - Phoenix fire-fighting world planner
  - 4 sets of execution logs gathered from about 400 runs of the planner
  - dependency of behavior on number of logs, noise in logs etc.
  - assessing information gain from applying the approach and effort to acquire information

  $\rightarrow$ extensive evaluation of the features of the developed system, very well written, easy to understand
- Pages: 43

Topic 8: Message Understanding Systems

- Journal paper, Computational linguistics, 1993
- *Evaluating message understanding systems: an analysis of the third message understanding conference (MUC-3)*, **Chinchor, Lewis and Hirschman**, Computational linguistics 19.3 (1993): 409-449.
- Content:
  - from the area of linguistics and natural language processing
  - survey of 15 systems extracting and summarizing information from newspaper articles
  - 3 types of systems: pattern-matching, syntax-driven, semantics-driven
  - 3 measures of effectiveness: recall, precision and fallout rates
  - making sure differences in effectiveness did not result from chance
  - computer-intensive method application: randomization tests

  $\rightarrow$ survey type of paper, details on particular systems not so important
- Pages: 41

Topic 9: Knowledge Extraction

- From the perspective of knowledge representation in Artificial Intelligence.
- Question: How do we actually gain information from text using Machine Learning techniques?
- 1) *Empirical Methods in Information Extraction*, **Clarie Cardie**, AI Magazine Volume 18 Number 4 (1997).
- 2) *Classifying the Semantic Relations in Noun Compounds via a Domain-Specific Lexical Hierarchy*, **Rosario and Hearst**, Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP-01).
- Summary: In this article, the author shows ways to extract information from corpus-based machine learning models.
- Pages: 15 + 9

Topic 10: Human-Machine-Interaction

- HMI is an important aspect of social robotics....
- Question: How do we actually evaluate spoken dialogues using empirical methods?
- 1) *PARADISE: A framework for evaluating spoken dialogue agents*, **Walker et al.**, Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics.
- 2) *The PARADISE Evoluation Framework: Issues and Findings*, **Hajdinjak et al.**, Computational Linguistics, 32(2), 263-272.
- Summary: The authors present PARADISE framework which enables the calculation of performance over subdialogues and whole dialoguies, specifies the relative contribution of various factors to performance and makes it possible to compare agents performing different tasks by normalizing for task complexity.
- Pages: 10 + 10

General Guidelines on the Paper

- should contain the following sections: Abstract, Introduction, *one or more Details*, Conclusion
- doesn't have to be very technical on the idea, has to be very detailed on evaluation (not applicable to book chapters)
- parts that are difficult to understand can be mentioned briefly, you can choose what to talk about
- important parts missing from the paper will have to be addressed in the presentation

Outline

Choosing the Topic

- Choose a topic from given list or suggest your own (has to be approved)
- Due: 22.04, Sunday, 23:59 Berlin time
- Write an email (`seba@cs.uni-bremen.de` or `stelter@cs.uni-bremen.de`) with your chosen topic
- First-come, first-served policy

Paper (Ausarbeitung)

- A scientific paper using Latex template
- Pages: 10 - 15
- Language: English
- Due: 10.06, Sunday, 23:59 Berlin time
- Read the paper carefully before 31.05!
- For questions: write an email
- Evaluation will be ready by 25.06 and will contain hints on what to add to the presentation

Presentation

- PDF slides using Latex template
- Duration: 20 min presentation, 10 min questions
- Language: slides in English, talk in German / English
- Date: 23.07 - 27.07, Will be decided through doodle
- **Make sure you are available in that week!**

Q & A

Thanks for your attention!

Course Content section slides inspired by the corresponding workshop presentation of Paul Cohen, Ian P. Gent and Toby Walsh.