Project Dilemmas

- How do I know when I'm done? How do I know what I've accomplished?
  - clearly define focus/goal from beginning
    - design a search method that handles plateaus
    - improve some ML method's robustness wrt noise
    - find the best ever solution to some really hard/interesting application
    - model what makes problems within a domain difficult
  - should be able to draft introduction and conclusion to your final paper almost from the start… at least the big picture

I did the proposal; now what?

- Hypothetical ToDo List
  - Outline final paper, add to outline as research progresses
  - Update literature review ***periodically*** (check citeseer)
  - Send email to author(s) of most related work requesting any recent papers or code (if appropriate)
  - Think about evaluation: formulate hypotheses, design experiments, propose analyses
  - Implement or obtain support code, e.g., simulators, test case generators (if appropriate)
  - Gather data sets
  - Code easy/greedy/strawman solution and see how it does
  - Implement base version of your solution
  - Run pilot experiment to confirm your intuitions, identify pitfalls early, fix parameter tunings, etc.
  - Revisit original goals, hypotheses
  - Run experiments
  - Analyze results
  - Finish writing paper
  - Ask someone else to read paper, then edit it based on suggestions.

Empirical Evaluation in AI

"Computer science is an empirical discipline. … Each new program that is built is an experiment. It poses a question to nature, and its behavior offers clues to an answer. Neither machines nor programs are black boxes; they are artifacts that have been designed, both hardware and software, and we can open them up and look inside. We can relate their structure to their behavior and draw many lessons from a single experiment."

from Newell and Simon's Turing Award Lecture in 1975

Purposes of Evaluation

- demonstrate that your ideas work, your intuitions were correct (or not)
- show how program/algorithm performance varies under different circumstances
- convince others of the value of what you've done (substantiation, justification)
- situate research in context of field

Evaluation Myths

- Because we build programs, we understand them.
- It is enough to build a program.
- Theory simply has to get more sophisticated.
- We can control our experiments because they are on a computer.
- Results on one dataset will generalize to other data sets.
- Synthetic problems are a fine proxy for actual problems.
- What isn't controlled doesn't matter.

Types of Evaluation

- Demo/Proof of Concept/Assessment
  - show that far fetched idea actually might work
- System Performance Evaluation
  - efficiency of data structures and methods
  - operational profile
- Comparison – Who's best
  - how does new algorithm compare to its predecessors; what does it add?
- Hypothesis testing: Manipulation or Observation
  - program is motivated by some hypothesis about problem, structure, function of algorithms…Show that the hypothesis holds (or doesn't…).

Three Basic Research Questions

- Description:

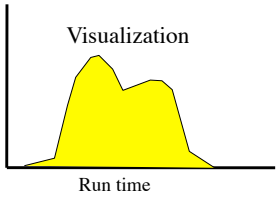
  "What will happen when…" → Try it and see
- Prediction:

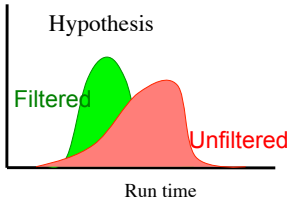
  "Does this model accurately predict what will happen when…" → Yes, No or Maybe
- Explanation:

  "Does this model accurately explain what will happen when…" → Yes, No or Maybe

Three Classes of Empirical Methods

- Visualization, summarization, exploration
- Experiment designs
- Statistical methods





Essential Components of Empirical Studies

- a subject (program, system)
- one or more tasks that the subject must perform
- an environment
- metrics of performance/behavior
- a protocol.

Hypothesis Testing

- Experiments are based on hypothesis:
  - My system has significantly better performance than state of the art.
  - Heuristics significantly improve performance.
  - Negative feedback makes little difference to performance.

Basic Terminology

- Independent variable
  - What is being actively manipulated or at least controlled
- Dependent variable
  - A phenomenon that can be measured and whose value is expected to *depend* on the values of the independent variables

Hypotheses relate values of independent variables to observations of dependent variables.

Terminology

- Data scales: categorical, ordinal/ranked, interval (relative), ratio (absolute)
- Factorial experiment: sample every combination of possible factors
- Nonparametric tests: make no assumptions about the distribution of data

Key Problem

- Correctly attributing the cause of a change (or lack thereof) in the dependent variable.
- Extraneous variables
  - any variable other than the independent variables that effects the dependent variable

→ *Experimental control*

- Manipulation experiment: Manipulate independent variable(s) and nothing else, then measure differences in dependent variable(s)

Handling Extraneous Variables

- Strategies
  - Construct sequence of experiments or add more independent variables (if possible)
  - Treat the extraneous variables as sources of variance and assume (hope!) that they exert roughly the same influence across the dependent variables

Spurious Variables

- Sometimes a bias creeps into how sample is selected or produced…
  - Consider running code on machines in the lab which unbeknownst to you are part of a memory upgrade. *Coincidentally*, algorithm A is being run on the machines with less and algorithm B on the machines with more.

Spurious Variable! (aka confounded variable or sampling bias)

Control Strategies

- Incorporate "control" or baseline conditions
- Use random sampling to control for noise variables (and avoid spurious variables)
- If too much noise, then high variance and if necessary, run new experiment with new independent variables.

Summary of Control

- Experiments test whether one factor influences another, as indicated by variables.
- Levels of variables define conditions.
- Control for extraneous variables by systematically varying conditions.
- Control for noise variables by random sampling.
- If noise variables are systematically associated with independent and dependent variables, we have sampling bias.

Guidelines for Experiment Design

1. Experiment Procedure: include independent & dependent variables, protocol, sampling strategy, number of trials, intervals of observation collection
2. Example of a data table: how will the variables be expected to combine?
3. Example of your analysis: what tests will you run on the data once you have it?
4. Discussion of possible outcomes and how they relate back to original hypothesis

Canonical AI Comparison Experiment Protocol

1. For each algorithm A being compared,
   a. For each parameter setting P of A,
      1) If ML: Train A(P) on a training data set D'.
      2) For each testing data set D,
         a) Run A on D collecting metrics M
         b) Compare actual results to expected results if some expectation
         c) Compute performance metrics
   b. Compare performance on settings P for algorithm A
2. Compare performance across set A on best P for each A using statistical tests for significance

Canonical AI Comparison Experiment Protocol (cont.)

Independent variables:

- algorithm set A
- parameter settings P
- data set(s) D and D' (testing and training)

Dependent variables:

- metrics M

Key Questions

- To what methods should you compare?
- What parameters should be manipulated?
- What data sets should be used?
- If ML: How should you divide into training and testing?
- What metrics should be used?
- On what platform (hardware/software) will your experiment run?
- What statistical tests should be run?

Choosing Methods/Algorithms

- Use strawman to show problem difficulty
- Use state of the art methods to show improvement
- Use similar methods to show influence of specific changes/additions
- Use code supplied by author(s) whenever possible to remove claims of poor programming or other bias (e.g., bad parameter choice)

Choosing Parameters

- Follow recommendations of author(s), either their defaults or what they have used in their comparisons
- Sample the parameter space in pilot experiments to determine best settings and assess variance
- Make the parameter be an independent variable

Choosing Data Sets and Metrics

- Follow common practices (S.O.P) in the area, e.g.
  - use benchmarks to expedite comparison
  - use well established metrics (e.g., # evaluations, % to global optima, % to best known solution for optimization problems, accuracy, computation time).
- Construct data sets that show performance scale-up and boundaries (e.g., a mix of easy and hard).
- Add new problems or metrics as justified by algorithm motivation or by interest (have a reason!)
- If not possible to use all benchmarks, then pick some with specific goal in mind and randomly select from others.

Choosing Platform

- Beware of comparing your results against previously published ones!
- Avoid comparing CPU times as too many exogenous factors contribute. Substitute other metrics for computation, e.g., # evaluations.
- Try to run all experiments on same platform (realize that systems people upgrade software regularly which might change your platform performance mid-experiment).
- Understand the platform specific aspects of your and other's code.
- Nice your processes if distributing on many machines.

Statistics 101: Data Definitions

- *Measurement* is a variable value associated with an individual.
- *Sample* is a collection of measurements.
- A sample should be representative of the *population* from which it is drawn.
- *Distribution* refers to either 1) the frequency of different measurements for a variable or 2) the shape that characterizes the frequencies.

Scales of Data

- *Categorical* or *nominal*: measurement assigns category labels
- *Ordinal*: values are ranks without any magnitude information
- *Interval*: distances between values are meaningful
- *Ratio*: Like interval plus there is a fixed reference point (e.g., absolute 0)

Distributions

Univariate distribution

Planner	Success?	Time
FF	T	.5
UCPOP	F	1800
SGPlan	T	1.5
FF	F	30
UCPOP	F	1800
SGPlan	F	500
…	…	…

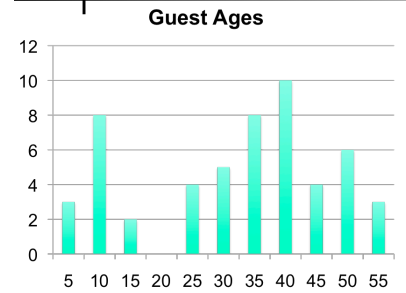
Joint distribution

Partitions divide distributions into sub-parts according to a variable value.

Statistical Methods

- Based on distributions of
  - individuals across categories (Categorical)
  - ranks (Ordinal)
  - real numbers (Interval, Ratio)
- Can transform more informative into less (e.g., ranks into categories) or smooth noisy data depending on hypothesis

Visualizing One Variable



Gaps suggest unequal influences of another factor.

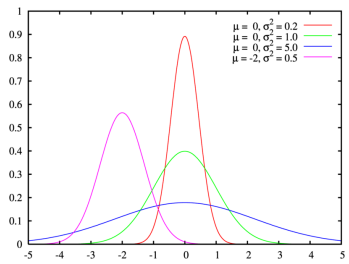
Histogram

Terms:
- bin
- frequency
- mode

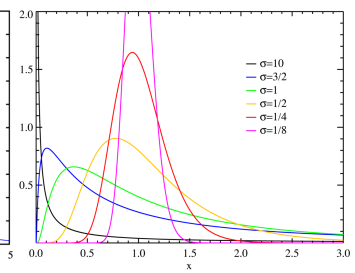
Univariate Summary Statistics

Statistic	Time
Size	63
Mean	25.56
Median	23.92
Mode	9.49
Skew	1.79
Minimum	9.39
Maximum	84.43
Range	75.04
Standard deviation	12.27
Variance	150.61

Common Distributions



Normal or Gaussian



Log-normal: one tailed

Joint Distributions of Categorical Data

Hypothesis:

Younger people prefer lighter chocolate.

	White	Milk	Dark	TOTAL
Undergrad				
Grad				
Faculty				
TOTAL				

Contingency Table

Statistical Analyses

- Issues
  - Parametric? Tests rely on assumptions about population parameters (e.g., normally distributed data)
  - Type of data (nominal, ordinal, ratio or interval)?
  - Hypothesis testing or modeling?
  - Multiple comparisons?

Null Hypothesis

- Basis of statistical hypothesis testing
- Reverse of what you are hypothesizing – that chance is responsible for an effect observed in data
- In running tests, we are trying to reject the null hypothesis: H_0

Type 1 and Type 2 Errors

1. Error of rejecting the null hypothesis when it actually is true aka "the level of significance" or α
2. Error of accepting the null hypothesis when it is false (missing a real difference) aka "the power of the test"

Statistical Test for Categorical: Contingency Table

Alg	Solved	Failed	Total
A	11	17	28
B	17	15	32
C	14	14	28
TOTAL	42	46	88

Observed Frequencies

Alg	Solved	Failed	Total
A	13.36	14.64	28
B	15.27	16.73	32
C	13.36	14.64	28
TOTAL	42	46	88

Expected Frequencies

- Determine how closely an observed distribution matches an expected one – Goodness-of-Fit
- Test of independence: calculate expected frequencies from totals
- Chi squared statistic

$$\chi^2 = \sum_{i,j} \frac{\left(f_{ij} - \hat{f}_{ij}\right)^2}{\hat{f}_{ij}}$$

- Compare statistic to distribution to determine p *http://www.graphpad.com/quickcalcs/PValue1.cfm*

Interval or Ratio Single Var: T-test

Prob	Alg A	Alg B	Δ
1	0.01	0.04	-0.03
2	0.13	0.01	0.12
3	8.72	10.11	-1.39
4	0.05	0.04	0.01
5	36.0	60.0	-24.00
6	3.5	8.22	-4.72
AVG	8.07	13.07	-5.00
Sd	14.10	23.43	9.49

- Determine whether the means of two populations on some outcome differ, e.g., two levels of a categorical independent variable
- Use t-statistic to compare two samples

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{SD_1^2}{\eta_1} + \frac{SD_2^2}{\eta_2}\right)}}$$

- Calculate p from t distribution
- http://www.graphpad.com/quickcalcs/ttest1.cfm?Format=SD

Wilcoxon Rank Sums

Alg A	Alg B	Sorted
0.01	0.04	0.01,0.02
0.13	0.02	0.04,0.05
8.72	10.11	0.06,0.09
0.06	0.05	0.12,0.13
36.0	60.0	2.11,3.5
3.5	8.22	3.89,5.24
2.11	3.89	8.22,8.72
15.3	19.75	9.33,10.11
5.24	9.33	15.3,19.75
0.09	0.12	36.0,60.0

- Non-parametric alternative to t-test
- Order values by rank and compute ranks of a group
- Convert to Z and use Normal distribution
- http://www.fon.hum.uva.nl/Service/Statistics/Wilcoxon_Test.html

$$w_A = 1 + 5 + 6 + 8 + 9 + 10 + 12 + 14 + 17 + 19 = 101$$

Interval or Ratio Multiple Var: ANOVA

Alg A	Alg B	Alg C
3.57	9.92	3.53
2.70	8.12	6.58
0.94	0.53	14.61
4.00	3.49	10.64
2.80	5.52	8.84
1.35	4.29	4.82

- Compare means of >2 groups (k): ANalysis Of VAriance
- Assumption: = SDs for all groups
- Between Mean Squares

$$BMS = \frac{\sum_k \eta_k \left(\bar{x}_k - \bar{x}\right)^2}{|k| - 1}$$

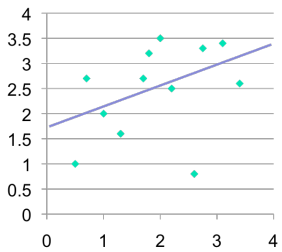
- Within Sum of Squares

$$WMS = \left(\frac{\sum_k (\eta_k - 1) SD_k}{N - |k|}\right)$$

- F

$$F = \left(\frac{BMS}{WMS}\right)$$

Model: Linear Regression



- Function that relates the sum of weighted independent variables to a dependent variable

$$y = a_0 + a_1 x_1$$

- Quality of fit

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)}{TotalVariance}$$

Data Analysis for the Four Studies

- Assessment, Exploratory
  - Visualizations of distributions, histograms, contingency tables, scatterplots, time series
  - Descriptive statistics
  - Modeling, e.g., regression
- Manipulation, Observation
  - Tests of effects on means, e.g., t-test, ANOVA
  - Tests of interaction effects on means, e.g., ANOVA
  - Tests of effects on proportions, e.g., chi-square
  - Tests of predictive power, e.g., R^2
  - Tests of distribution assumptions

Experiment Pitfalls

- Poorly specified hypothesis
- Reliance on flaky and/or too few users/data sets
- Bias in the user base/data set
- Inappropriate comparison methods
- Varying too many variables/parameters of experiment simultaneously
- Biased evaluation metric (confounding) or data selection
- Too many or too few statistical tests

Strategies for Avoiding Pitfalls

- Fully specify experiment before running it
- Run pilot experiment
- Put serious thought into “right” hypothesis
- Think about what you hope to say when experiment concludes… will experiment support your saying it?

Evaluation Plan

- Hypotheses
- Type of experiment: comparison, performance assessment, modelling
- Independent variables (what you are testing)
- Dependent variables (what you are measuring) and why
- Exogenous variables (what you think might matter, but don't have time to test and so are setting some way), e.g., length of each trial, platform, code parameters, data set?
- Estimated resources required: how many trials, how much time/trial, how many machines, other
- Analyses: statistics? Which ones? How many?
- What do you expect to say from experiment?