Optimal Streaming Feature Selection

Bob Stine and Dean Foster
Dept of Statistics, Wharton School
University of Pennsylvania

Wharton
Department of Statistics

Plan

- Streaming feature selection
  - What it is, why it's useful,
  - Variable selection problems introduced
- Alpha investing
  - Controls expected number false positives
  - Test a possibly infinite sequence of hypotheses
  - Flexibility presents many alternatives
- Optimal method
  - Universal spending rule
  - Compare performance to oracles

Streaming Feature Selection

- Context
  - L_0 variable selection
  - Eg: Picking variables for a regression model

    $$\hat{y} = b_0 + b_1 x_1 + \ldots + b_k x_k + \mathbf{?}$$
- Streaming selection
  - External source offers new feature z
  - Judge whether to add to model or not
- Novelties
  - Choice of z may depend on prior outcomes

    Construction of interactions, transformations
  - Don't require full X matrix at start of search

    Data base query, text mining, ...
  - Can be done very fast compared to stepwise

Over-Fitting?

- Control fitting process
  - Wide data set permits choice among many
  - Unbounded set of rules to make new features
- Cross-validation
  - Less data to fit model, estimate parameters
  - Time to repeat fitting process
- Hypothesis testing, p-value based
  - Requires computing 'honest' p-values
  - FDR requires full set of p-values
  - Alpha-investing tests sequence one-at-a-time...

Alpha-Investing

- Testing sequence of hypotheses $H_1, H_2, \ldots$
- Alpha wealth
  - Start with allowance $W_1 = \omega$ for Type I error
  - 'Invest' $0 \le \alpha_i \le W_i$ in test of H_i
  - Compute p-value pi of test of H_i
  - Test determines wealth available for testing H_{i+1}

$$W_{i+1} = W_i - \alpha_i + \omega\, I\{p_i \le \alpha_i\}$$

- Comments
  - Controls mFDR
  - Resembles alpha-spending rule
  - Can replicate Bonferroni, FDR, ...

Optimal Alpha Investing

- Question
  - How much of the current wealth W_i should one invest in the test of H_i?
- Answer
  - It depends! Easy to start thinking strategically
  - If expect a 'good variable', set α_i close to W_i to earn more for later tests
    … but …
    If variables are 'really good', can test with small $\alpha_i \approx 0$ and still reject and have more for later
- Wrong issue
  - Ought to be thinking about the choice of the next variable z rather than the amount

Problem

- Estimators
  - Test *n* hypotheses H_i: $\mu_i = 0$ vs $\mu_i \neq 0$
  - Tests define estimators 'testimator'
  - Goal: accumulate minimal risk over *n* tests

Risk versus counting 'correct rejections'

- Class of strategies
  - Monotone spending rules that set $\alpha_{i+1} \leq \alpha_i$
  - Represent spending rule by density $f(i)$

$$0 \leq f(i+1) \leq f(i), \ \Sigma f(i) = 1$$

  - Denote cumulative risk $R(\mu, f)$

Investing Rules

- Geometric
  - Invest fixed proportion ψ of current wealth in next test

    $g(i) = \psi(1-\psi)^i, \quad i=0,1,2,\ldots$
- Uniform
  - Distribute wealth equally over tests

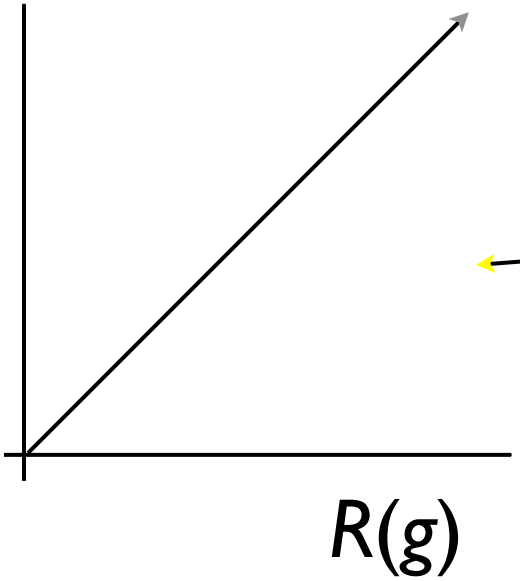
    $h(i) = 1/n$
- Universal
  - Version of Rissanen's universal prior for integers

    $u(i) = c/[(i+1)\log(i+2)] \quad i=0,1,2,\ldots$
  - Invests much initially $u(0) = 0.614$
    but much less later $u(30) =$

Comparing Investing Rules

- Question rephrased
  - Arbitrary sequence $\mu_1, \mu_2, \ldots \mu_n$
  - Two monotone density functions *g* and *h*
  - What is the cumulative risk implied by using investing rule *g* vs investing rule *h*?
- Graph feasible set

Coding
All rules work well somewhere!

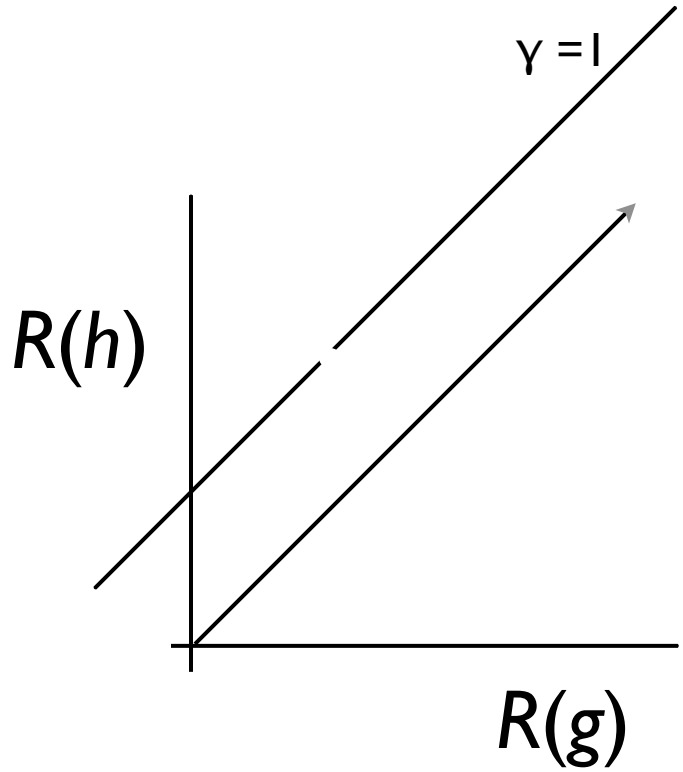


There exists sequence of means for which risks are $R(g), R(h)$

Finding Feasible Set

- Compute from Bellman equations
  - Reduce to one-dimensional maximization

    $\max_\mu R(\mu,f) - \gamma R(\mu,g)$ for various γ
  - messy, big recursion



References

- Need some to other than us???

- Streaming feature selection
  - Lin, Foster, and Ungar (2011), JASA
- Alpha-investing
  - Foster and Stine (2008), JRSS-B
- Honest p-values
  - Foster and Stine (2003), JASA
-

That's all for now...
Thanks for coming!