Online Decision-Making, Neyman-Pearson, and Binary Classification

Data 102 Spring 2022

Lecture 4

Weekly Overview

- Last week: multiple hypothesis testing

- **Today:**
  - Making decisions with feedback (online decision-making)
  - Hypothesis testing with a known alternative (Neyman-Pearson)
  - Connection to binary classification

- Next time: connecting decision-making and frequentist/Bayesian views

Recap: controlling error rates

- Goal: adjust p-value threshold to account for multiple tests
- No adjustment
  - Does not control any error rate: could have lots of false positives
- Bonferroni
  - Controls FWER: P(at least one FP)
  - How it works: to get FWER $\leq \alpha$, use p-value threshold α/m for all tests
- Benjamini-Hochberg
  - Controls FDR: how many of our discoveries are wrong, on average (column-wise error rate)
  - How it works: (1) sort p-values (indexed by k), (2) draw the line $y = k * \alpha/m$, (3) find the largest p-value that's under the line, and (4) use that p-value as the threshold

Recap: Multiple Hypothesis Testing

Algorithm	**What it controls**	**How it works**
"Naive thresholding"	FPR per-test	For FPR α, use p-value threshold α
Bonferroni correction	Family-wise error rate: FWER	For FWER α across m tests, use p-value threshold α/m
Benjamini-Hochberg (B-H)	False discovery rate: FDR	For FDR α across m tests: 1. Sort the p-values, index by k (starting at k=1) 2. Find the largest p-value that's below $k*\alpha/m$ (in other words, the last p-value below the line $y=k*\alpha/m$) 3. Use that p-value as the p-value threshold for all tests

Understanding FWER vs FDR

- FWER: P(any test is a FP)
- FDR: E[FDP] = expected proportion of discoveries that are wrong

- When might we prefer one or the other?
  - Example: missile detection
  - Example: website improvements

- Can you come up with one example where we should use FWER, and one where we should use FDR?

Online Decision-Making

- Sometimes, we don't get to see all the data up front
  - Early decisions may influence later ones
  - Example: expensive sequence of medical tests
  - Example: A/B tests for website optimization
- In **online** decision-making, we see a p-value and must make a decision right then
  - We can't go back and change the decision later
- Bonferroni can be used in online settings *if* we know the total number of tests: why?
- B-H can't be used in online settings: why?

LORD

Algorithm 1 The LORD Procedure

input: FDR level α, non-increasing sequence $\{\gamma_t\}_{t=1}^{\infty}$ such that $\sum_{t=1}^{\infty} \gamma_t = 1$,
initial wealth $W_0 \leq \alpha$

Set $\alpha_1 = \gamma_1 W_0$

for $t = 1, 2, \ldots$ **do**

p-value P_t arrives

if $P_t \leq \alpha_t$, reject P_t

$\alpha_{t+1} = \gamma_{t+1} W_0 + \gamma_{t+1-\tau_1}(\alpha - W_0)\mathbf{1}\{\tau_1 < t\} + \alpha \sum_{j=1}^{\infty} \gamma_{t+1-\tau_j} \mathbf{1}\{\tau_j < t\}$,

where τ_j is time of j-th rejection $\tau_j = \min\{k : \sum_{l=1}^{k} \mathbf{1}\{P_l \leq \alpha_l\} = j\}$

end

LORD

- At time i, we have alpha-wealth α_i : how optimistic we are about making a discovery at time i
    - This is our p-value threshold

- Every time we make a discovery, we gain “wealth” that decays over time

- Current wealth is the sum of decayed wealth from all previous discoveries

Neyman-Pearson: if we know the alternative

Multiple Hypothesis Testing: Four Approaches

Algorithm	**What it controls**	**How it works**
“Naive thresholding”	FPR per-test	For FPR α, use p-value threshold α
Bonferroni correction	Family-wise error rate: FWER	For FWER α across m tests, use p-value threshold α/m
Benjamini-Hochberg (B-H)	False discovery rate: FDR	For FDR α across m tests: 1. Sort the p-values, index by k (starting at k=1) 2. Find the largest p-value that’s below k*α/m (in other words, the last p-value below the line y=k*α/m) 3. Use that p-value as the p-value threshold for all tests
LORD	FDR, online	● Use p-value threshold α_t for test t ● Each time you make a discovery, gain wealth (that decays over time)

Hypothesis testing vs binary classification

- Both are trying to make binary decisions
  - Hypothesis testing: do our data support the null or alternative?
  - Binary classification: class 0 or class 1?
- In hypothesis testing, we work with p-values: $P(\text{data} \mid R = 0)$
  - For one test, we pick a threshold to control FPR (Neyman-Pearson: also maximize TPR)
  - For multiple tests, we'll define group-wise error rates and try to control those
  - We observe a particular sequence of p-values, but the sequence could have been different
- In binary classification, we (often) work with arbitrary numbers we can threshold to get a decision
  - Often 0-1 and can be interpreted as probabilities (e.g., logistic regression)
  - Pick threshold by using an ROC curve (TPR, FPR) or precision-recall curve
  - Notebook demo