Data 102 (Fall 2022)
Discussion Worksheet 3
09/09/2022

Names:

1. **Bias-variance.** Decompose the squared-error loss between a parameter θ and estimator δ into bias and variance terms. (Recall that the squared error is $\mathbb{E}[(\delta(X) - \theta)^2]$.)

2. **ROC Curves.** Consider the toy dataset in the table below; Y is the label, X_1, X_2 are features, and we use the prediction function $f(X_1, X_2)$.

Table 1: Example dataset

Y	$f(X_1, X_2)$	X_1	X_2
0	-1	-1	0.5
1	-0.5	-1	0.75
0	0	-1	1
1	1	0.2	-0.3
1	0.25	-0.25	0
0	0.25	-0.05	-0.3

   (a) Draw the ROC curve for the prediction function f with respect to the label Y.

   (b) Is it possible to choose a (possibly randomized) decision threshold for f, such that the expected true positive rate is $\frac{1}{3}$, and the expected false positive rate is $\frac{2}{3}$?

3. **LORD Procedure**:
Recall the LORD Procedure.

Algorithm 1 The LORD Procedure

Input FDR level α, non-increasing sequence $\{\gamma_t\}_{t=1}^{\infty}$ such that $\sum_{t=1}^{\infty} \gamma_t = 1$

Set $\alpha_1 = \gamma_1 \alpha$.
for $t = 1, 2, ...,$ **do**
  p-value P_t arrives.
  if $P_t \leq \alpha_t$ **then**
    Reject P_t.
  Update $\alpha_{t+1} = \gamma_{t+1} W_0 + \alpha \sum_{j=1}^{\infty} \gamma_{t+1-\tau_j} 1\{\tau_j < t\}$, where τ_j is the time of the j'th rejection.

   (a) You want to control the FDR with LORD at level α. Set $\gamma_t = 2^{-t}$. You are currently at time step $t = 5$, and the only rejection you've made so far was at time step $t = 4$. How small must the 5th p-value be in order for you to make a discovery at this time step?

(b) You again want to control the FDR with LORD at level α. Set $\gamma_t = 2^{-t}$. You are currently at time step $t = 5$, and the only rejection you've made so far was at time step $t = 1$. How small must the 5th p-value be in order for you to make a discovery at this time step?

4. Alice has a bag with 3 red marbles, 2 blue marbles, and 1 green marble. Norman has a bag with 1 red marbles, 1 blue marble, and 4 green marbles. You observe two samples with replacement from either Alice or Norman, and want to figure out which is which. You want to have the highest TPR while keeping the FPR at $\frac{1}{9}$. What decision rule do you pick? What is the corresponding TPR? Assume that **Norman is the null** and **Alice is the alternative**.

   To help you get started, the table below writes out the probabilities for all 9 outcomes:

	P_A	P_N
RR	1/4	1/36
RB	1/6	1/36
RG	1/12	1/9
BR	1/6	1/36
BB	1/9	1/36
BG	1/18	1/9
GR	1/12	1/9
GB	1/18	1/9
GG	1/36	4/9

5. **Bayes risk + Bayes-optimal classifier**

   Recall the Bayes risk for an arbitrary decision procedure $\delta(X)$ is

$$R(\delta) = \mathbb{E}_{\theta,X}[\ell(\theta, \delta(X))]. \tag{1}$$

   Ideally, we'd like to find the *best* decision procedure

$$\delta^* = \arg\min_{\delta} R(\delta). \tag{2}$$

   (a) Find δ^* for $\ell(\theta, \delta(X)) = \mathbf{1}[\theta \neq \delta(X)]$ (zero-one loss). δ^* should be a function of X, and involves at least one conditional probability and one $\arg\max$.

   (b) Suppose X can take on two possible values $\{x_1, x_2\}$ and θ can take on two possible values $\{\theta_1, \theta_2\}$. If

$$\begin{cases} \mathbb{P}(\theta = \theta_1 | X = x_1) = 0.1 \\ \mathbb{P}(\theta = \theta_1 | X = x_2) = 0.9 \\ \mathbb{P}(\theta = \theta_2 | X = x_1) = 0.6 \\ \mathbb{P}(\theta = \theta_2 | X = x_2) = 0.4 \end{cases} \tag{3}$$

   what is $\delta^*(x_2)$ under the 0-1 loss from part (a)?

   (c) (Optional) Find δ^* for $\ell(\theta, \delta(X)) = (1/2)(\theta - \delta(X))^2$ (squared-error loss)

6. **Benjamini-Yekutieli procedure** (Challenge Question)

   Suppose you are testing n hypotheses and want to control the FDR at level α. It turns out that Benjamini-Hochberg is only guaranteed to work when the hypotheses are *independent* or *positively correlated.* Construct an example with negatively correlated hypotheses where Benjamini-Hochberg fails.

   Remark: The Benjamini-Yekutieli procedure, a generalization of Benjamini-Hochberg, controls the FDR regardless of independence assumptions, and therefore is guaranteed to work in all cases. It is shown below. The only difference from Benjamini-Hochberg is the $c(n)$ function highlighted in red.

Algorithm 2 The Benjamini-Yekutieli Procedure

Input FDR level α, set of n p-values $P_1, \ldots, P_n$

1: Sort the p-values $P_1, \ldots, P_n$ in non-decreasing order $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(n)}$

2: Find $K = \max\{i \in \{1, \ldots, n\} : P_{(i)} \leq \frac{\alpha}{n \cdot c(n)} i\}$, where

$$c(n) = \begin{cases} 1 & \text{tests are independent or positively correlated (this is just B-H)} \\ \sum_{j=1}^{n} \frac{1}{j} & \text{tests are dependent or negatively correlated} \end{cases} \tag{4}$$

3: Reject the null hypotheses (declare discoveries) corresponding to $P_{(1)}, \ldots, P_{(K)}$
