Data 102 (Fall 2022) **Names:**
Discussion Worksheet 3
09/09/2022

---

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 $\mathbb{E}[(\delta(X) - \theta)^2]$.)

   Solution:

$$\begin{aligned}\mathbb{E}(\delta(X) - \theta)^2 &= \mathbb{E}(\delta(X) - \theta + \mathbb{E}\delta(X) - \mathbb{E}\delta(X))^2 \\ &= \mathbb{E}(\delta(X) - \mathbb{E}\delta(X))^2 - 2\mathbb{E}(\delta(X) - \mathbb{E}\hat{\delta}(x))\mathbb{E}(\theta - \mathbb{E}\delta(X)) + \mathbb{E}(\theta - \mathbb{E}\delta(X))^2 \\ &= \underbrace{\mathbb{E}(\delta(X) - \mathbb{E}\delta(X))^2}_{\text{Variance}} + \mathbb{E}(\underbrace{\theta - \mathbb{E}\delta(X)}_{\text{Bias}})^2.\end{aligned}$$

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

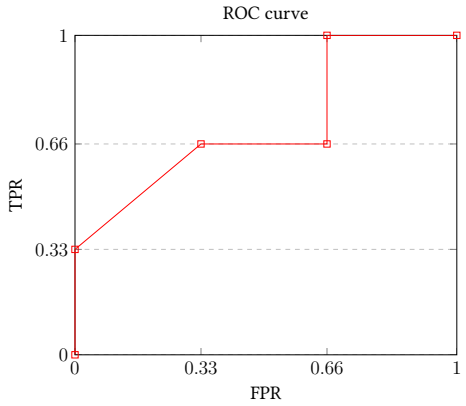
Y	$f(X_1, X_2)$	X_1	X_2
0	-1	-1	0.5
1	-0.5	-1	0.75
0	0	-1	1
1	1	0.2	-0.3
1	0.25	-0.25	0
0	0.25	-0.05	-0.3

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

   Solution: At a given decision threshold α, if $f(X_1, X_2) > \alpha$, then the decision is a positive classification, and if $f(X_1, X_2) \leq \alpha$, then the decision is a negative classification. Since the model $f(X_1, X_2)$ only takes five different values on this dataset, only five different decision thresholds lead to different true and false positive rates.

   - $\alpha < -1$: TPR $= 1$, FPR $= 1$
   - $-1 \leq \alpha < -0.5$: TPR $= 1$, FPR $= \frac{2}{3}$
   - $-0.5 \leq \alpha < 0$: TPR $= \frac{2}{3}$, FPR $= \frac{2}{3}$
   - $0 \leq \alpha < 0.25$: TPR $= \frac{2}{3}$, FPR $= \frac{1}{3}$
   - $0.25 \leq \alpha < 1$: TPR $= \frac{1}{3}$, FPR $= 0$
   - $1 \leq \alpha$: TPR $= 0$, FPR $= 0$

   Plotting each (FPR, TPR) combination as a point on a plot results in the following ROC curve:



(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 $\frac{2}{3}$?

Solution: No, this is not possible for this dataset. Every deterministic threshold corresponds to a point on the ROC curve drawn above, and every randomized threshold corresponds to a *convex combination* of points on the ROC curve. By inspecting the picture above, we see that $(\frac{1}{3}, \frac{2}{3})$ is slightly outside of the convex hull.

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 $\sum_{t=1}^{\infty} \gamma_t = 1$

1: Set $\alpha_1 = \gamma_1 \alpha$.
2: **for** $t = 1, 2, ...,$ **do**
3: $\quad$ p-value P_t arrives.
4: $\quad$ **if** $P_t \leq \alpha_t$ **then**
5: $\quad\quad$ Reject P_t.
6: $\quad$ 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?

Solution: At most $(\frac{1}{32} + \frac{1}{2})\alpha$.

(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?

Solution: At most $(\frac{1}{32} + \frac{1}{16})\alpha$.

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	$1/36$	$4/9$

   Solution: We want to test the hypotheses: $\begin{cases} H_0 : \text{samples came from Norman} \\ H_1 : \text{samples came from Alice} \end{cases}$

   So $\delta = 1$ corresponds to picking Alice, and $\delta = 0$ corresponds to picking Norman. Using the intuition from Neyman-Pearson, we want to set $\delta = 1$ for outcomes that have a large likelihood ratio P_A/P_N. We compute these ratios in the table below:

	P_A	P_N	LR	δ
RR	$1/4$	$1/36$	9	1
RB	$1/6$	$1/36$	6	1
RG	$1/12$	$1/9$	$3/4$	0
BR	$1/6$	$1/36$	6	1
BB	$1/9$	$1/36$	4	1
BG	$1/18$	$1/9$	$1/2$	0
GR	$1/12$	$1/9$	$3/4$	0
GB	$1/18$	$1/9$	$1/2$	0
GG	$1/36$	$4/9$	$9/144$	0

   Pick the likelihood-ratio threshold to be $3/4$; i.e. reject the null if LR $> 3/4$. Then,

$$\begin{aligned} \mathbb{P}(\text{reject the null}|\text{null is true}) &= \mathbb{P}(\text{you observed RR, RB, BR, or BB}|\text{Norman sampled}) \\ &= \frac{4}{36} = \frac{1}{9}. \end{aligned}$$

The corresponding TPR is

$$\begin{aligned}\mathbb{P}(\text{reject the null}|\text{null is false}) &= \mathbb{P}(\text{you observed RR, RB, BR, or BB}|\text{Alice sampled}) \\ &= \frac{27}{36} = \frac{3}{4}.\end{aligned}$$

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 $\arg\max$.

Solution: Note that

$$\mathbb{E}[\ell(\theta, \delta(X))] = \mathbb{P}(\theta \neq \delta(X)) \tag{3}$$

To minimize this, for any particular value of $X = x$ find the value $a^* = \delta(x)$ that solves

$$\begin{aligned}a^* &= \arg\min_{a\in\mathbb{R}} \mathbb{P}(\theta \neq a|X = x) \\ &= \arg\min_{a}(1 - \mathbb{P}(\theta = a|X = x)) \\ &= \arg\max_{a} \mathbb{P}(\theta = a|X = x).\end{aligned}$$

(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

$$\begin{cases}\mathbb{P}(\theta = \theta_1|X = x_1) = 0.1 \\ \mathbb{P}(\theta = \theta_1|X = x_2) = 0.9 \\ \mathbb{P}(\theta = \theta_2|X = x_1) = 0.6 \\ \mathbb{P}(\theta = \theta_2|X = x_2) = 0.4\end{cases} \tag{4}$$

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

Solution: $\delta^*(x_2) = \theta_1$.

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

Solution: Following the pointwise minimization strategy, for any particular value of $X = x$ we find the value $a^* = \delta(x)$ that solves

$$a^* = \min_{a\in\mathbb{R}} \mathbb{E}_{\theta}[(1/2)(\theta - a)^2 \mid X = x].$$

To do this, we take the derivative with respect to a and set it to zero. Swapping the differentiation and expectation operators and applying the chain rule gives

$$f'(a) = \mathbb{E}_\theta[a - \theta \mid X = x].$$

Further setting the derivative to zero gives

$$f'(a) = 0 \implies a^* = \mathbb{E}[\theta | X = x].$$

That is, for any particular value of $X = x$, we should take $\delta^*(x) = \mathbb{E}[\theta \mid X = x]$. That means that the decision rule that minimizes the Bayes risk for the squared error loss is $\delta^*(X) = \mathbb{E}[\theta \mid X]$, the posterior mean.

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

$$c(n) = \begin{cases} 1 & \text{tests are independent or positively correlated (this is just B-H)} \\ \sum_{j=1}^{n} \frac{1}{j} & \text{tests are dependent or negatively correlated} \end{cases} \tag{5}$$

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

Feedback Form

On a scale of 1-5, where 1 = much too slow and 5 = much too fast, how was the pace of the discussion section?

1 2 3 4 5

Which problem(s) did you find most useful?

Which were least useful?

Any other feedback?