HW3

Kevin Chang

October 13, 2025

1

Let Ω be a set of N bits, each equal to either 0 or 1. For a fixed integer $k < N$, sample a sequence $y_1, \ldots, y_{k+1}$ from Ω uniformly without replacement. Define the sample average of the first k elements as

$$\hat{m}_k = \frac{1}{k} \sum_{i=1}^{k} y_i.$$

and define the average of all of the bits as

$$m_N = \frac{1}{N} \sum_{y \in \Omega}^{N} y.$$

Show that

$$\mathbb{E}[(\hat{m}_k - y_{k+1})^2] = \frac{N}{N-1} \times \frac{k+1}{k} \times m_N(1 - m_N)$$

Proof. For $i \neq j$ under simple random sampling without replacement:

$$\mathbb{E}[y_i] = m_N, \qquad \mathrm{Var}(y_i) = m_N(1 - m_N),$$

- $\mathbb{E}[y_i y_j] = \mathbb{E}[\mathbb{E}[y_i y_j | y_i]] = \mathbb{E}[y_i \frac{M_n N - y_i}{N-1}] = \frac{N m_N^2 - m_N}{N-1}$.
- $\mathrm{Cov}(y_i, y_j) = \mathbb{E}[y_i y_j] - M_n^2 = -\frac{m_N(1-m_N)}{N-1}$.
- $\mathrm{Var}(\hat{m}_k) = \frac{1}{k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} \mathrm{Cov}(y_i, y_j) = \frac{1}{k^2} M_n(1 - M_n)(k - \frac{k(k-1)}{N-1}) = \frac{m_N(1-m_N)}{k} \cdot \frac{N-k}{N-1}$.
- $\mathrm{Cov}(\hat{m}_k, y_{k+1}) = \frac{1}{k} \sum_{i=1}^{k} \mathrm{Cov}(y_i, y_{k+1}) = -\frac{m_N(1-m_N)}{N-1}$.

Since $\mathbb{E}[\hat{m}_k] = \mathbb{E}[y_{k+1}] = m_N$, we have

$$\mathbb{E}\big[(\hat{m}_k - y_{k+1})^2\big] = \mathrm{Var}(\hat{m}_k - y_{k+1}) = \mathrm{Var}(\hat{m}_k) + \mathrm{Var}(y_{k+1}) - 2\,\mathrm{Cov}(\hat{m}_k, y_{k+1}).$$

Substituting the above expressions yields

$$\mathbb{E}\big[(\hat{m}_k - y_{k+1})^2\big] = m_N(1 - m_N) \left(\frac{N-k}{k(N-1)} + 1 + \frac{2}{N-1} \right) = m_N(1 - m_N) \cdot \frac{N(k+1)}{k(N-1)}.$$

□

2

Let X be a continuous random variable distributed over the closed interval $[0, 1]$. For label $Y = 0$, X is uniform:

$$p_{X|Y}(X|Y=0) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

For label $Y = 1$, the conditional pdf of X is as follows:

$$p_{X|Y}(X|Y=1) = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

- Find the decision rule that minimizes the probability of error.

$$\hat{Y}(x) = \mathbb{1}\left\{ \frac{p(x \mid Y=1)}{p(x \mid Y=0)} \;\geq\; \frac{p_0\left(loss(1,0) - loss(0,0)\right)}{p_1\left(loss(0,1) - loss(1,1)\right)} \right\} = \mathbb{1}\left\{ 2x \geq \frac{P[Y=0]}{P[Y=1]} \right\}$$

- Find the closed form expression for the operating characteristic of the likelihood ratio test (LRT), i.e., TPR as a function of FPR for the LRT.
  - $\mathrm{FPR}(t) = \Pr[x \geq \frac{P[Y=0]}{2P[Y=1]} \mid Y=0] = 1 - \frac{P[Y=0]}{2P[Y=1]}$
  - $\mathrm{TPR}(t) = \Pr[x \geq \frac{P[Y=0]}{2P[Y=1]} \mid Y=1] = \int_{\frac{P[Y=0]}{2P[Y=1]}}^{1} 2x\,dx = 1 - \left(\frac{P[Y=0]}{2P[Y=1]}\right)^2$.
  - Closed form:

$$\boxed{\mathrm{TPR} = 2\,\mathrm{FPR} - \mathrm{FPR}^2, \qquad 0 \leq \mathrm{FPR} \leq 1.}$$

- Suppose we require the TPR to be at least $(1+\epsilon)$FPR, where $\epsilon > 0$ is a fixed constant. Find TPRmax(ϵ), the maximal value of TPR that is achievable under this constraint.
  - Let $u = \mathrm{FPR} \in [0,1]$. On the LRT ROC we have

$$\mathrm{TPR}(u) = 2u - u^2.$$

  - Impose $\mathrm{TPR} \geq (1+\epsilon)\,\mathrm{FPR}$ with $\epsilon > 0$:

$$2u - u^2 \;\geq\; (1+\epsilon)u \iff u\big((1-\epsilon) - u\big) \;\geq\; 0.$$

  - Thus the feasible set is

$$\mathcal{U}_\epsilon = \begin{cases} [0,\, 1-\epsilon], & 0 \leq \epsilon \leq 1, \\ \{0\}, & \epsilon \geq 1. \end{cases}$$

    Since $\mathrm{TPR}(u) = 2u - u^2$ is increasing on $[0,1]$, the maximum over $\mathcal{U}_\epsilon$ occurs at

$$u^\star = \begin{cases} 1-\epsilon, & 0 \leq \epsilon \leq 1, \\ 0, & \epsilon \geq 1. \end{cases}$$

  - Hence

$$\boxed{\mathrm{TPR}_{\max}(\epsilon) = \begin{cases} 2(1-\epsilon) - (1-\epsilon)^2 \;=\; 1 - \epsilon^2, & 0 \leq \epsilon \leq 1, \\ 0, & \epsilon \geq 1. \end{cases}}$$

3

Suppose you have n expert forecasters. Each individual forecaster $i \in [n]$ has a known true positive rate and false positive rate (TPRi , FPRi). Define forecaster 0 to be the artificial forecaster that only predicts $y = 0$ and forecaster $n+1$ to be the artificial forecaster who only predicts $y = 1$. Consider the set of forecasts defined by the following procedure:

- 1. Pick two numbers i and j between 0 and $n+1$ (inclusive) and a scalar $p \in [0,1]$.
- 2. Generate a random number r from the uniform distribution on $[0,1]$.
- 3. If $r < p$, return the prediction of forecaster i. Otherwise, return the prediction of forecaster j.

Answer:

- For a given $t \in [0,1]$, we can tune the randomized forecaster to achieve TPR $= t$. The expected TPR is

$$\mathrm{TPR}(p; i, j) = p \cdot \mathrm{TPR}_i + (1-p) \cdot \mathrm{TPR}_j.$$

Thus, for any target t satisfying $\min\{\mathrm{TPR}_i, \mathrm{TPR}_j\} \le t \le \max\{\mathrm{TPR}_i, \mathrm{TPR}_j\}$, one may set

$$p = \frac{t - \mathrm{TPR}_j}{\mathrm{TPR}_i - \mathrm{TPR}_j},$$

which guarantees $\mathrm{TPR}(p; i, j) = t$. In particular, by mixing the trivial forecasters 0 (with $\mathrm{TPR}_0 = 0$) and $n+1$ (with $\mathrm{TPR}_{n+1} = 1$), we obtain

$$\mathrm{TPR}(p; 0, n+1) = p,$$

so any $t \in [0, 1]$ is attainable by choosing $p = t$.

- The expected FPR is given by

$$\mathrm{FPR}(p; i, j) = p \cdot \mathrm{FPR}_i + (1 - p) \cdot \mathrm{FPR}_j.$$

Hence, randomization between two forecasters yields every point on the line segment connecting their ROC coordinates $(\mathrm{FPR}_i, \mathrm{TPR}_i)$ and $(\mathrm{FPR}_j, \mathrm{TPR}_j)$. By extension, the entire family of randomized forecasters spans the *convex hull* of all ROC points. In particular, by mixing forecasters 0 and $n+1$, we can obtain every point on the diagonal TPR = FPR of the ROC plane.

4

In this problem, we consider an automated resume screening tool which is used by a company to sort candidates based on whether or not they are predicted to be invited for an on site interview after an initial phone screen. Let the random variable X denote the features of a candidate's application and Y denote whether a candidate is invited for an on site interview, where $Y = 1$ indicates that an individual was invited.

- Suppose that there are many qualified individuals looking for jobs and that paying recruiters to call applicants is expensive. As a result, it is comparatively half as costly for the company to miss a candidate who would have been invited on site than it is to spend time calling an individual who is not invited for an interview (i.e. for some $\alpha > 0, C_{10} = \alpha, C_{01} = \frac{\alpha}{2}$, and other costs are zero). Show that the company's optimal decision rule for resume screening has the form $s(x) = \mathbb{E}[Y|X = x] \ge t$, and find the value of t.

$$R(1 \mid x) = C_{10} \Pr(Y = 0 \mid X = x) = \alpha\,[1 - s(x)],$$
$$R(0 \mid x) = C_{01} \Pr(Y = 1 \mid X = x) = \frac{\alpha}{2}\, s(x),$$

where $s(x) = \Pr(Y = 1 \mid X = x) = \mathbb{E}[Y \mid X = x]$.
Choose $\hat{Y} = 1$ iff $R(1 \mid x) \le R(0 \mid x)$, i.e.,

$$\alpha\,[1 - s(x)] \;\le\; \frac{\alpha}{2}\, s(x) \iff 1 - s(x) \le \frac{s(x)}{2} \iff s(x) \ge \frac{2}{3}.$$

- Now suppose that unemployment has gone down, and there are no longer many qualified candidates looking for jobs. As a result, it is instead twice as costly to miss good candidates than it is to call ones who are not invited for an interview (i.e. for some $\beta > 0, C_{10} = \beta, C_{01} = 2\beta$, and other costs are zero). How does the optimal decision rule change?

$$R(1 \mid x) = C_{10} \Pr(Y = 0 \mid X = x) = \beta\,[1 - s(x)],$$
$$R(0 \mid x) = C_{01} \Pr(Y = 1 \mid X = x) = 2\beta\, s(x),$$

where $s(x) = \Pr(Y = 1 \mid X = x) = \mathbb{E}[Y \mid X = x]$.
Choose $\hat{Y} = 1$ iff $R(1 \mid x) \le R(0 \mid x)$, i.e.,

$$\beta\,[1 - s(x)] \;\le\; 2\beta\, s(x) \iff s(x) \ge \frac{1}{3}.$$

Suppose now that some score function $\hat{s}(x)$ has been estimated from historical data, and a threshold rule is applied to assign individuals the screening predictions $\hat{Y} = 1$ for those who will be considered more closely by recruiters and $\hat{Y} = 0$ for those who will not. In the United States, it is illegal to discriminate against job applicants on the basis of religion, and your job is to evaluate this tool with that in mind. Below is a table which shows the predictions and outcomes for applicants split by membership in a minority religious group, with $A = 1$ indicating that an individual is a member of this group and $A = 0$ indicating that they are not. We have data from 500 candidates in the religious group and 5, 000 candidates not in the religious group.

A = 1			
	$Y = 0$	$Y = 1$	Total
$\hat{Y} = 0$	360	40	400
$\hat{Y} = 1$	40	60	100
Total	400	100	500

A = 0			
	$Y = 0$	$Y = 1$	Total
$\hat{Y} = 0$	4050	450	4500
$\hat{Y} = 1$	200	300	500
Total	4250	750	5000

- With membership in the religious group as the sensitive attribute, does this classifier satisfy independence? Does it satisfy sufficiency? Justify your answer.
  - Independence requires $\hat{Y} \perp A$, i.e.,

    $$\Pr(\hat{Y} = 1 \mid A = 1) = \Pr(\hat{Y} = 1 \mid A = 0).$$

    From the tables:

    $$\Pr(\hat{Y} = 1 \mid A = 1) = \tfrac{100}{500} = 0.20, \qquad \Pr(\hat{Y} = 1 \mid A = 0) = \tfrac{500}{5000} = 0.10.$$

    These differ, so the classifier *does not* satisfy independence.
  - Sufficiency requires $Y \perp A \mid \hat{Y}$. For a binary predictor, this is equivalent to equal PPV and equal $\Pr(Y = 1 \mid \hat{Y} = 0)$ across groups:

    $$\Pr(Y = 1 \mid \hat{Y} = 1, A = 1) = \tfrac{60}{100} = 0.60, \quad \Pr(Y = 1 \mid \hat{Y} = 1, A = 0) = \tfrac{300}{500} = 0.60;$$

    $$\Pr(Y = 1 \mid \hat{Y} = 0, A = 1) = \tfrac{40}{400} = 0.10, \quad \Pr(Y = 1 \mid \hat{Y} = 0, A = 0) = \tfrac{450}{4500} = 0.10.$$

    These match, hence the classifier *does* satisfy sufficiency.
- For the criteria that the classifier doesn't satisfy, propose a group-dependent change to the threshold that results in a classifier that does satisfy the criteria. You do not need to specify exact quantities, rather comparisons with the current threshold. You should not propose a trivial threshold that results in 0% or 100% acceptance rates.

  Independence requires $\Pr(\hat{Y} = 1 \mid A = 1) = \Pr(\hat{Y} = 1 \mid A = 0)$. From the tables, current acceptance rates are 0.20 (group $A = 1$) and 0.10 (group $A = 0$), so the classifier violates independence.

  A group-dependent thresholding scheme can enforce independence by equalizing acceptance rates. **Raise the threshold for $A = 1$ and lower it for $A = 0$** so that both groups' acceptance rates meet at an intermediate value (e.g., $\approx 15\%$). This reduces acceptance for $A = 1$ and increases it for $A = 0$.
- Compare and contrast the value of the intervention you suggested in part 4 for the following two circumstances:
  - You learn that the historical data comes from a hiring manager who is a member of the religious group and has been heard telling fellow members that they have an "in" regardless of their qualifications.

    Answer: If past decisions (the labels Y) systematically favored $A = 1$ then the observed calibration (sufficiency) is with respect to *biased* labels. In this case, enforcing independence by *raising the threshold for $A = 1$ and/or lowering it for $A = 0$* can partially counteract historical favoritism and improve equity of opportunity in current screening.

– You learn that there is a well regarded religious university nearby that sends the resumes of highly qualified students to the company. Historically, these candidates have highly relevant skill sets and make up a majority of applications from the religious group.

Answer: If a strong nearby religious university yields a genuinely higher base rate of qualified applicants for $A = 1$, then the acceptance-rate gap may reflect real prevalence differences. Forcing independence by *raising* $A = 1$*'s threshold or lowering* $A = 0$*'s* could reduce utility: either excluding many truly qualified $A = 1$ candidates (higher false negatives) or including more unqualified $A = 0$ candidates (higher false positives). In this scenario, preserving *sufficiency* (calibration within groups) is typically more appropriate, while monitoring group-wise error rates (separation) may also be informative.

5

Apply the concepts from the course lectures to your final project.

(i) Provide your project abstract. This may have changed from last week, and that's ok.

(ii) What are the features X and labels Y ? What could they represent? How are they represented?

(iii) Describe the sources of these data. Are they simulated, generated, or retrospective? How costly are they to acquire?

(iv) What are the relevant prediction problems involved with your project? Why are these prediction targets of interest to yourself, to the research community, and to the broader public?

(v) Why do you believe that the labels can be predicted from the features?

(vi) What metrics might you use to evaluate the quality of your predictions? Do you expect there to be tradeoffs between these associated metrics?

1. We aim to use the BlueSky air traffic simulator to study and predict air traffic management (ATM) dynamics under high-density conditions. BlueSky is an open-source ATM simulator capable of running scenarios with hundreds of aircraft simultaneously, making it a powerful tool for generating realistic performance and navigation datasets. Our project will investigate predictive modeling tasks such as estimating potential conflicts between aircraft, predicting delays, and evaluating the effects of different traffic management policies. The ultimate objective is to contribute methods for safer and more efficient ATM systems by leveraging simulated data to explore machine learning–based approaches. We are especially interested in methods to achieve high-density advanced air mobility (AAM) using compositional, structured air traffic management policies.

2. **Features X and Labels Y.**

   Features (X): Aircraft state vectors (position, altitude, speed, heading), flight plans, separation distances, traffic density measures, and environmental conditions (e.g., wind speed if included). These are represented numerically as time series or structured arrays. Current and historical demand for drone delivery or air taxi services are possible features.

   Labels (Y): Potential conflict events (binary: conflict/no conflict), delay times (continuous), or safety risk scores (categorical/ordinal). These labels represent the prediction targets and are derived from simulation outputs or post-processed conditions. Future demand of airspace from drone delivery or air taxi services.

3. **Data Sources.** Data Sources. The data are primarily simulated using BlueSky, which allows for controlled and repeatable generation of ATM scenarios. Because the simulator is open-source and computationally efficient, generating data is relatively low-cost compared to collecting real-world air traffic data, which is proprietary, regulated, and expensive to acquire. For traditional aviation, there are many flight records and routes that can be analyzed. However, we anticipate the need to generate data to model the more complex and crowded air space conditions that are imagined in AAM scenarios.

4. **Prediction Problems.** Relevant prediction problems include:

   Conflict detection: predicting whether two aircraft will lose safe separation within a time horizon.

   Delay prediction: estimating arrival delays under varying traffic conditions.

   Policy evaluation: predicting safety or efficiency outcomes under different scheduling or routing policies.

   These prediction problems are of interest to:

   - *Ourselves:* to demonstrate the applicability of machine learning methods to ATM.
   - *Research community:* to benchmark new algorithms on open, reproducible simulation data.
   - *Broader public:* to promote safer, more efficient air travel and better management of crowded airspaces.

5. **Predictability of Labels from Features.** Aircraft states, traffic density, and flight plans are strong predictors of conflicts and delays because ATM dynamics are governed by physical constraints and well-established separation standards. The simulator provides both micro-level (individual aircraft trajectories) and macro-level (airspace-level congestion) features that directly drive the occurrence of conflicts and delays, making prediction feasible and meaningful. Past demand and routes can inform future demand (e.g., rush hour in road networks).

6. **Evaluation Metrics.** Given that our primary prediction task involves classification (e.g., conflict detection), we will evaluate model performance using metrics such as accuracy, precision, recall, and F1-score. The exact nature of the evaluation will depend on the final prediction task that we converge to.