Machine Learning B (2025)
Home Assignment 4

Bar Segal xsb740

17/05/2025

Contents

1 The Airline question

Part 1: Single flight (known no–show rate 0.05)

Setup. Sell $n = 100$ tickets for a plane with 99 seats. Each passenger independently shows up with probability $p = 0.95$ (and no-shows with probability 0.05).

Random variable. Let

$$X = \text{number of passengers who show up.}$$

Then

$$X \sim \text{Binomial}(n = 100,\ p = 0.95).$$

Overbooking event. An overbooking occurs when more passengers show up than there are seats:

$$\Pr[\text{overbooked}] = \Pr[X > 99].$$

Because $X \leq 100$, this is the same as

$$\Pr[X > 99] \ = \ \Pr[X = 100].$$

Exact probability. All 100 ticket-holders must appear, so

$$\Pr[X = 100] \ = \ (0.95)^{100} \ = \ e^{\,100\ln(0.95)} \ \approx \ 5.92 \times 10^{-3} \ \approx \ 0.59\%.$$

Conclusion. The airline will face an overbooked flight with probability

$$\boxed{(0.95)^{100} \ \approx \ 0.006 \text{ (about } 0.6\%)},$$

or roughly once in every 170 flights.

Remark: Since the event $X > 99$ reduces to the single point $X = 100$, we can compute the probability exactly with one binomial mass. Therefore there is no need to invoke tail inequalities such as Chernoff or Hoeffding bounds here; those tools are useful when the event involves a wider tail region or when an explicit calculation is infeasible.

Part 2: Historical sample of 10 000 + one flight of 100

We want a bound on the probability of observing

(a) a **sample of 10 000** passengers with exactly 95 % show-ups, and

(b) the **next flight of 100** passengers all showing up,

when the true (unknown) show-up probability is some $p \in [0, 1]$.

Two directions to bound this probability are:

(a) Treat the two samples as independent Let

$$E_1 = \{X_1 = 9\,500\} \quad \text{with } X_1 \sim \mathrm{Bin}(10\,000, p), \qquad E_2 = \{X_2 = 100\} \quad \text{with } X_2 \sim \mathrm{Bin}(100, p).$$

For a fixed p,

$$\Pr(E_1 \wedge E_2 \,|\, p) \;=\; \binom{10\,000}{9\,500} p^{9\,600}\,(1-p)^{500}.$$

Hoeffding on E_1. $\Pr(E_1 \,|\, p) \le \exp\!\big[-2{\cdot}10\,000\,(|p-0.95|)^2\big]$. Hence

$$\Pr(E_1 \wedge E_2 \,|\, p) \;\le\; p^{100}\,\exp\!\big[-2{\cdot}10\,000\,(|p-0.95|)^2\big].$$

Maximising the RHS over $p \in [0,1]$ gives the worst case $p^\star \approx 0.953$ and

$$\boxed{\Pr(E_1 \wedge E_2) \le 6.8 \times 10^{-3}}.$$

(This matches the "≈ 0.0068" hinted at in the exercise.)

(b) "Split-after-sampling" (VC-style) argument

1. Draw once a sequence of 10 100 independent Bernoulli(p) variables. Condition on the event that exactly 9 600 are "show" and 500 are "no-show".
2. Randomly allocate 100 of the 10 100 passengers to the next flight, the remaining 10 000 form the historical sample.

Let $C = \{\text{exactly } 9\,600 \text{ shows out of } 10\,100\}$. For a fixed p,

$$\Pr(C \,|\, p) \;=\; \binom{10\,100}{9\,600} p^{9\,600}(1-p)^{500}.$$

Given the composition (9 600 show, 500 no-show), the probability that all 100 chosen for the flight come from the 9 600 shows is hypergeometric:

$$\Pr(E_2 \,|\, C) \;=\; \frac{\binom{9\,600}{100}}{\binom{10\,100}{100}} \;\approx\; 6.08 \times 10^{-3}.$$

Hence, for any p,

$$\Pr(E_1 \wedge E_2 \,|\, p) \;=\; \Pr(C \,|\, p)\,\frac{\binom{9\,600}{100}}{\binom{10\,100}{100}}.$$

The factor $\Pr(C \,|\, p)$ is maximised at $p^\star = 9\,600/10\,100 \approx 0.9505$, giving $\Pr(C \,|\, p^\star) \approx 1.83 \times 10^{-2}$. Therefore

$$\Pr(E_1 \wedge E_2) \;\le\; 1.83 \times 10^{-2} \times 6.08 \times 10^{-3} \;\approx\; 1.1 \times 10^{-4}.$$

$$\boxed{\Pr(E_1 \wedge E_2) \le 1.1 \times 10^{-4}}.$$

(If the composition probability $\Pr(C)$ is omitted, then we recover the looser bound 6.1×10^{-3} stated earlier.)

2 Equivalence of the Two PAC Variants

(a) Standard $\Longrightarrow$ Pos/Neg

Assume a concept class $\mathcal{C}$ is efficiently PAC learnable by the hypothesis class $\mathcal{H}$. That is, there exists a polynomial-time algorithm

$$\mathcal{A}_{\mathrm{PAC}}\big(\mathrm{EX}(c; D), \varepsilon, \delta\big)$$

which, for every distribution D over the instance space and every target concept $c \in \mathcal{C}$, outputs $h \in \mathcal{H} \cup \{h_0, h_1\}$ satisfying

$$\Pr_{x \sim D}\big[\, h(x) \neq c(x) \,\big] \;\leq\; \varepsilon \quad \text{with probability} \quad \geq 1 - \delta.$$

We show that $\mathcal{C}$ is also efficiently positive–negative (PN) PAC learnable using the same hypothesis class.

Notation. For any target c, let D_c^+ (resp. D_c^-) be the distribution of instances conditioned on $c(x) = 1$ (resp. $c(x) = 0$). The oracles

$$\mathrm{EX}_c^+ \;:\; x \sim D_c^+, \qquad \mathrm{EX}_c^- \;:\; x \sim D_c^-$$

supply positive and negative examples without labels (labels are implicit).

Algorithm $\mathcal{B}$ for PN-PAC learning

1. **Setting parameters:** Let
$$m \;=\; m_{\mathrm{PAC}}\big(\varepsilon/2,\; \delta\big),$$
where m_{PAC} is the sample bound required by $\mathcal{A}_{\mathrm{PAC}}$. (Polynomial in $1/\varepsilon$ and $\log(1/\delta)$.)

2. **Creating a synthetic mixed sample:** For each $i = 1, \ldots, m$:
   (a) Flip an unbiased coin.
   (b) If heads, draw $x_i \sim \mathrm{EX}_c^+$ and set the label $y_i \leftarrow 1$.
   (c) If tails, draw $x_i \sim \mathrm{EX}_c^-$ and set the label $y_i \leftarrow 0$.

   Denote the resulting labelled multiset by $S = \{(x_i, y_i)\}_{i=1}^m$. Note that every (x_i, y_i) is distributed exactly as a draw from the mixture distribution $D_{\mathrm{mix}} := \frac{1}{2} D_c^+ + \frac{1}{2} D_c^-$.

3. **Running the standard PAC learner:** Invoke
$$h \;\leftarrow\; \mathcal{A}_{\mathrm{PAC}}\big(S,\; \varepsilon/2,\; \delta\big)$$
and return h.

Correctness analysis:

By the guarantee of $\mathcal{A}_{\mathrm{PAC}}$, with probability at least $1-\delta$,

$$\Pr_{x\sim D_{\mathrm{mix}}}[\,h(x)\neq c(x)\,]\;\leq\;\tfrac{\varepsilon}{2}.$$

Write

$$p^{+}\;=\;\Pr_{x\sim D_c^{+}}[h(x)=0],\qquad p^{-}\;=\;\Pr_{x\sim D_c^{-}}[h(x)=1].$$

Because D_{mix} selects a positive and a negative example each with probability $1/2$,

$$\Pr_{D_{\mathrm{mix}}}[h(x)\neq c(x)]=\tfrac{1}{2}p^{+}+\tfrac{1}{2}p^{-}\leq\tfrac{\varepsilon}{2}.$$

Multiplying by 2 gives $p^{+}+p^{-}\leq\varepsilon$. Since both terms are non-negative, each must be $\leq\varepsilon$:

$$\Pr_{x\sim D_c^{+}}[h(x)=0]\;\leq\;\varepsilon,\quad\Pr_{x\sim D_c^{-}}[h(x)=1]\;\leq\;\varepsilon.$$

Thus h meets the PN-PAC accuracy requirement. The probability of success is at least $1-\delta$, the same as for $\mathcal{A}_{\mathrm{PAC}}$.

Efficiency:

The algorithm draws exactly $m=m_{\mathrm{PAC}}(\varepsilon/2,\delta)$ examples and performs only a coin flip and an oracle call per example, plus one invocation of $\mathcal{A}_{\mathrm{PAC}}$. Because m and the running time of $\mathcal{A}_{\mathrm{PAC}}$ are polynomial in $1/\varepsilon$ and $\log(1/\delta)$, $\mathcal{B}$ is an efficient PN-PAC learner.

Conclusion. Efficient standard PAC learnability of $\mathcal{C}$ implies efficient positive–negative PAC learnability by the same hypothesis class $\mathcal{H}$.

(b) Pos/Neg $\Longrightarrow$ Standard

Assume a concept class $\mathcal{C}$ is efficiently positive–negative (PN) PAC learnable with respect to the same hypothesis class $\mathcal{H}$. That is, there exists a polynomial–time algorithm

$$\mathcal{A}_{\mathrm{PN}}\big(\mathrm{EX}_c^{+},\mathrm{EX}_c^{-},\varepsilon,\delta\big)$$

which, given independent samples from the conditional distributions D_c^{+} (positives) and D_c^{-} (negatives), outputs $h\in\mathcal{H}\cup\{h_0,h_1\}$ satisfying

$$\Pr_{x\sim D_c^{+}}\big[h(x)=0\big]\leq\varepsilon\quad\text{and}\quad\Pr_{x\sim D_c^{-}}\big[h(x)=1\big]\leq\varepsilon$$

with probability at least $1-\delta$. We show that $\mathcal{C}$ is also efficiently PAC learnable in the standard model.

Notation. Let D be the (unknown) distribution from which the standard example oracle

$$\mathrm{EX}(c; D) \ : \ (x, c(x)) \sim D$$

draws labelled examples. Write $p \ = \ \Pr_{x\sim D}[\, c(x) \ = \ 1\,]$ for the positive class prior. By definition $D = pD_c^+ + (1-p)D_c^-$.

Algorithm $\mathcal{A}$ for the standard PAC model

1. **Parameter preparation.** Let

$$m \ = \ m_{\mathrm{PN}}\big(\varepsilon/2,\ \delta/3\big), \qquad N \ = \ \Big\lceil \frac{8m}{\varepsilon}\,\ln\frac{6}{\delta}\Big\rceil,$$

   where m_{PN} is the (polynomial) sample bound required by $\mathcal{A}_{\mathrm{PN}}$.

2. **Draw N mixed examples.** Query $\mathrm{EX}(c; D)$ N times. Split the sample into

$$S^+ \ = \ \big\{x_i : c(x_i) = 1\big\}, \qquad S^- \ = \ \big\{x_i : c(x_i) = 0\big\}.$$

3. **Handle extreme imbalance.**

   - If $|S^+| < m$, return h_0 (the constant 0 hypothesis).
   - Else if $|S^-| < m$, return h_1 (the constant 1 hypothesis).

4. **Run the PN learner.** Feed the first m positives and the first m negatives to $\mathcal{A}_{\mathrm{PN}}\big(\cdot; \varepsilon/2, \delta/3\big)$ and return the hypothesis h it produces.

Correctness analysis

Case 1: Both $|S^+|, |S^-| \geq m$. The examples supplied to $\mathcal{A}_{\mathrm{PN}}$ are i.i.d. from D_c^+ and D_c^-, so with probability $\geq 1 - \delta/3$,

$$\Pr_{x\sim D_c^+}\big[h(x) = 0\big] \leq \tfrac{\varepsilon}{2}, \qquad \Pr_{x\sim D_c^-}\big[h(x) = 1\big] \leq \tfrac{\varepsilon}{2}.$$

Consequently,

$$\Pr_{x\sim D}[h(x) \neq c(x)] \ = \ p\Pr_{D_c^+}[h(x) = 0] + (1-p)\Pr_{D_c^-}[h(x) = 1] \ \leq \ \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2} \ = \ \varepsilon.$$

Case 2: $|S^+| < m$ (symmetric for S^-). A multiplicative Chernoff bound applied to the N Bernoulli labels shows that, except with probability $\leq \delta/3$, $|S^+| < m$ implies the true prior $p < \varepsilon/4$. In that event the constant classifier h_0 makes error $\Pr_{x\sim D}[h_0(x) \neq c(x)] = p < \varepsilon$.

Union bound. Failure in Case 1 occurs with probability $\leq \delta/3$, and failure in either constant branch contributes at most $\delta/3$. Therefore the overall failure probability is $\leq \delta$, so $\mathcal{A}$ is a valid PAC learner.

Efficiency

Both N and m are polynomial in $1/\varepsilon$ and $\log(1/\delta)$. Sampling, counting, and running $\mathcal{A}_{\mathrm{PN}}$ (thus step 4) are all polynomial-time operations. Hence the entire algorithm $\mathcal{A}$ is efficient.

Conclusion. If $\mathcal{C}$ is efficiently PN-PAC learnable by $\mathcal{H}$, then $\mathcal{C}$ is efficiently PAC learnable in the standard model by $\mathcal{H}$ as well.

3 Growth Function

Let the growth function of a hypothesis set $\mathcal{H}$ be defined by

$$m_{\mathcal{H}}(n) \;=\; \max_{S\subseteq\mathcal{X},\,|S|=n} \big|\{\, h|_S : h\in\mathcal{H}\}\big|,$$

i.e. the maximum number of distinct dichotomies that $\mathcal{H}$ can realise on any n points.

Finite $\mathcal{H}$: the immediate upper bound

3.1 Finite $\mathcal{H}$: the immediate upper bound

Assume $|\mathcal{H}| = M < \infty$. Fix an arbitrary set $S \subseteq \mathcal{X}$ of n points. Every $h \in \mathcal{H}$ induces exactly one labelling of S, so

$$\big|\{h|_S : h\in\mathcal{H}\}\big| \;\leq\; M.$$

Conversely, the total number of all possible binary labellings of S equals 2^n, which can never be exceeded. Taking the maximum over all S preserves the inequality on each side; hence

$$\boxed{m_{\mathcal{H}}(n) \;\leq\; \min\{\, M,\, 2^n\}}.$$

Tightness of the bound. The two terms inside the min can both be achieved: if $M \leq 2^n$ we may choose $\mathcal{H}$ with exactly M mutually different hypotheses that realise M distinct labellings on some n–point sample, whereas if $2^n \leq M$ we can take $\mathcal{H}$ to contain all 2^n binary labellings of a fixed n–point set. Therefore the bound is tight in both parameter regimes

3.2 Exactly two hypotheses imply $m_{\mathcal{H}}(n) = 2$

Now let $\mathcal{H} = \{h_1, h_2\}$ with $h_1 \neq h_2$.

Upper bound Part 1 with $M = 2$ immediately yields $m_{\mathcal{H}}(n) \leq 2$.

Lower bound. Since $h_1 \neq h_2$, there exists $x^* \in \mathcal{X}$ such that $h_1(x^*) \neq h_2(x^*)$. Take any set S of n points that contains x^*. Then $h_1|_S \neq h_2|_S$, so $\big|\{h_1|_S, h_2|_S\}\big| = 2$, forcing $m_{\mathcal{H}}(n) \geq 2$.

Combining the bounds we obtain

$$\boxed{m_{\mathcal{H}}(n) = 2 \quad \text{for every } n \geq 1 \text{ whenever } |\mathcal{H}| = 2}.$$

(For completeness, on the empty sample we have $m_{\mathcal{H}}(0) = 1$.)

3.3 Sub-multiplicative property: $m_{\mathcal{H}}(2n) \leq m_{\mathcal{H}}(n)^2$

For every hypothesis class $\mathcal{H}$ and every $n \in \mathbb{N}$,

$$m_{\mathcal{H}}(2n) \;\leq\; \big(m_{\mathcal{H}}(n)\big)^2.$$

Fix any sample $T = \{x_1, \ldots, x_{2n}\}$ of size $2n$. Partition it into two blocks of equal size:

$$S_1 = \{x_1, \ldots, x_n\}, \qquad S_2 = \{x_{n+1}, \ldots, x_{2n}\}.$$

For a hypothesis $h \in \mathcal{H}$ denote its restrictions by $h|_{S_1} \in \{0,1\}^n$ and $h|_{S_2} \in \{0,1\}^n$. The dichotomy that h realizes on the full sample T is completely described by the ordered pair $\big(h|_{S_1},\, h|_{S_2}\big)$.

—

Counting ordered pairs. By definition of the growth function,

$$\#\{\, h|_{S_1} \mid h \in \mathcal{H} \,\} \;\leq\; m_{\mathcal{H}}(n), \qquad \#\{\, h|_{S_2} \mid h \in \mathcal{H} \,\} \;\leq\; m_{\mathcal{H}}(n).$$

Hence the number of distinct ordered pairs—and therefore the number of distinct dichotomies on T—is at most $m_{\mathcal{H}}(n) \times m_{\mathcal{H}}(n) = m_{\mathcal{H}}(n)^2$.

—

Maximising over all samples. Because the bound holds for every size-$2n$ sample T, it holds for the one that maximises the count, giving $m_{\mathcal{H}}(2n) \leq m_{\mathcal{H}}(n)^2$.

Thank you.