Machine Learning B (2025)
Home Assignment 4

Carsten Jørgensen, student ID: skj730

Contents

1 The Airline question

Queston 1.1

For this question we are given the following assumptions:

- 100 tickets sold for a flight with 99 seats
- Each person has a 5% probability of not showing up (or 95% probability of showing up)
- People show up independently

Let X be the random variable representing the number of people who show up. Then X follows a binomial distribution with $n = 100$ (tickets sold) and $p = 0.95$ (probability of showing up). The event "more people show up than seats available" means $X > 99$. Since we sold exactly 100 tickets, the maximum possible value for X is 100. Therefore: $P(X > 99) = P(X = 100)$ and:

$$\Pr\big(X = 100\big) = \binom{100}{100} 0.95^{100} 0.05^{0} = e^{100 \ln(0.95)} \approx 0,0059205$$

So

$$Pr(\text{more passengers than seats}) \ = \ 0.95^{100} \ \leq \ 5.93 \times 10^{-3} \approx 0.6\%.$$

Queston 1.2.a

The question reads "Bound the probability of observing such sample and getting a flight overbooked." I interpret "such sample" and the event of selling 100 tickets for a flight that can only hold 99 passengers.

But under bullet (a) it also says "In the first approach we consider two events: the first is that in the sample of 10000 passengers, where each passenger shows up with probability p, we observe 95% of show-ups. The second event is that in the sample of 100 passengers, where each passenger shows up with probability p, everybody shows up. Note that these two events are independent. Bound the probability that they happen simultaneously assuming that p is known."

We shall provide answers to both questions. We are given the following information

- The airline observed that 5% of passengers don't show up (based on 10,000 reservations)
- They sell 100 tickets for a flight with 99 seats
- We need to find the probability of all 100 passengers showing up

Let p be the true probability of a passenger showing up. We consider two events:

- Event A: In a sample of 10,000 passengers, 95% show up (9,500 people)

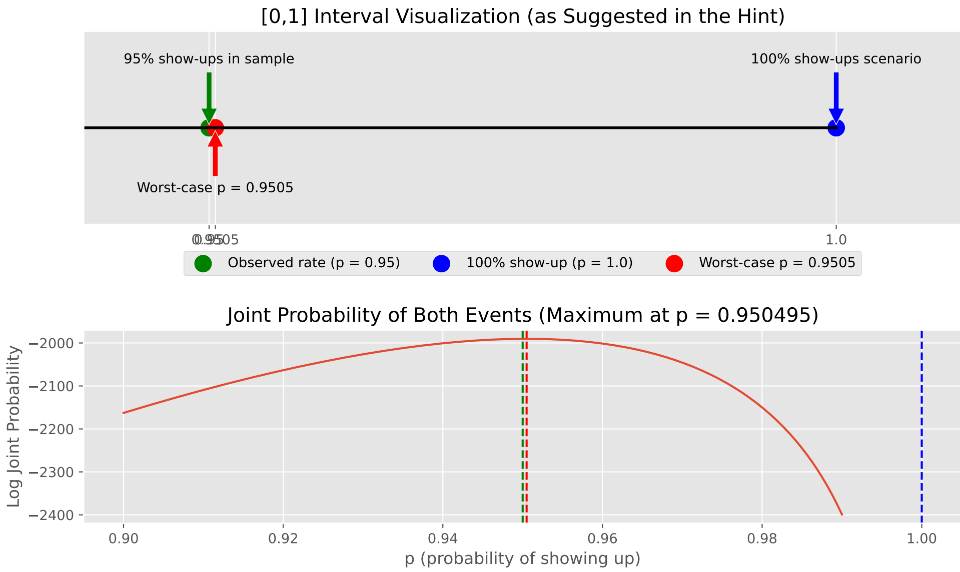

Figure 1: Visualization of hint

- Event B: In a sample of 100 passengers, all 100 show up (overbooking)

Probability of the two events when p is fixed

$$P_p(A) = \binom{10000}{9500} p^{9500}(1-p)^{500}$$
$$P_p(B) = p^{100}$$

so

$$P_p(A \cap B) = \binom{10000}{9500} p^{9600}(1-p)^{500}.$$

We compute the worst–case value with respect to the unknown p by defining:

$$f(p) = \ln P_p(A \cap B) = \ln\left(\binom{10000}{9500}\right) + 9600\ln(p) + 500\ln(1-p).$$

Differentiate f and set to zero gives us $9600/10100 \approx 0.950495$. Thus the joint probability attains its maximum at $p^* \approx 0.950495$.
The probability of overbooking with this worst-case p^* is:

$$P_{p^*}(B) = p^{*100} = 0.950495^{100} \approx 0.0062371$$

which lead to the conclusion that the probability of all 100 passengers showing up (causing overbooking) is bounded by 0.623%.

Figure 1 show a visualization of the hint given in the question. Now let compute a bound for the probability of both events occuring simultaneously. Using Stirling's formula, we compute $\ln(\binom{1000}{9500}) \approx 1981.1516704182736$, so

$$\begin{aligned}
\ln P_p^* &\approx 1981.1516704182736 + 9600 \ln(0.950495) + 500 \ln(0.049505) \\
&\approx 1981.15 - 487.41482358487605 - 1502.8408022038295 \\
&\approx -9.103955370432004
\end{aligned}$$

so $P_{p^*}(A \cap B) \approx \exp(-9.103955370432004) \approx 0.0001112$ which shows the the probability of observing the events simultaneously is bound by 0.01112%.

Question 1.2.b

For this questions we have to bound the probability of observing a sample of 10000 with 95% show ups AND a 99-seats flight with all 100 passengers showing up by following the below sampling protocol.

Draw $N = 10\,100$ i.i.d. Bernoulli(p) variables and split them

$$S = \{X_1, \dots, X_n\}, \qquad n = 10\,000,$$

$$S' = \{Y_1, \dots, Y_m\}, \qquad m = 100,$$

with $S \cup S' = \{0,1\}^N$ fixed and the split chosen at random.

Define the two empirical means

$$\bar{S} \;=\; \frac{1}{10\,000} \sum_{i=1}^{10\,000} X_i, \qquad \bar{F} \;=\; \frac{1}{100} \sum_{j=1}^{100} Y_j.$$

The event we wish to bound is

$$E \;=\; \bigl\{\bar{S} \geq 0.95 \;\wedge\; \bar{F} = 1\bigr\}.$$

Note we consider $\bar{S} \geq 0.95$ as I assume that as long as we are considering something worse that exactly 95% the upper bound is still valid.

Because the variables are independent, for every parameter $p \in [0,1]$

$$\mathbb{P}_p(E) \;=\; \mathbb{P}_p(\bar{S} \geq 0.95)\, \mathbb{P}_p(\bar{F} = 1). \tag{1}$$

Historical sample - $\bar{S}$: Hoeffding's inequality gives

$$\mathbb{P}_p(\bar{S} \geq 0.95) \;\leq\; \exp\Bigl(-2 \cdot 10\,000\,(0.95 - p)^2\Bigr). \tag{2}$$

Flight sample - $\bar{F}$: All 100 passengers show up iff every $Y_j = 1$, hence

$$\mathbb{P}_p(\bar{F} = 1) \;=\; p^{100}. \tag{3}$$

We find the uniform (worst–case) bound over p by combining (1)–(3):

$$\mathbb{P}_p(E) \;\leq\; g(p) := p^{100} \exp\Bigl(-2 \cdot 10\,000\,(0.95 - p)^2\Bigr)$$

We maximize $g(p)$ by solving

$$\frac{d}{dp}\log g(p) \;=\; \frac{100}{p} + 4 \cdot 10\,000\,(0.95 - p) = 0,$$

which yields

$$40\,000\,p^2 - 38\,000\,p - 100 = 0 \quad\Longrightarrow\quad p^\star \approx 0.95262433.$$

Evaluate g at $p^\star$:

$$\log g(p^\star) = 100\ln(0.95262433) - 2 \cdot 10\,000(0.95 - 0.95262433)^2 \approx -4.991207190565722,$$
$$g(p^\star) \approx e^{-4.991207190565722} \approx 0.006797453715245451 \approx 6.8 \times 10^{-3}.$$

Hence the worst (largest) value of g is attained at $p^\star$.

Therefore, for *every* unknown show-up probability $p \in [0, 1]$,

$$\mathbb{P}_p(E) = \mathbb{P}_p\Big(\bar{S} \geq 0.95 \;\wedge\; \bar{F} = 1\Big) \;\leq\; 6.8 \times 10^{-3}$$

i.e., the protocol's probability of simultaneously observing a 95% show-up rate in the size-10 000 sample *and* a 100% show-up rate on the 99-seat flight is at most about 0.68%.

2 PAC learnability

Question 2a

Proof. Suppose $\mathcal{C}$ is efficiently PAC learnable using $\mathcal{H}$ in the standard model. This means that for any $\epsilon, \delta > 0$, there exists a polynomial-time algorithm $\mathcal{A}$ that, given access to labeled examples, outputs a hypothesis $h \in \mathcal{H}$ such that with probability at least $1 - \delta$:

$$\Pr_{x \sim D}[h(x) \neq c(x)] \leq \epsilon$$

We construct an algorithm $\mathcal{A}_{pn}$ for the positively-negatively PAC learning model:

1. Create a mixed distribution D' by sampling from EX_c^+ and EX_c^- with equal probability:
   - With probability $\frac{1}{2}$, draw x from $\mathcal{D}_c^+$ and return $(x, 1)$
   - With probability $\frac{1}{2}$, draw x from $\mathcal{D}_c^-$ and return $(x, 0)$
2. Run algorithm $\mathcal{A}$ on this mixed distribution with parameters $\frac{\epsilon}{2}$ and δ
3. Return the hypothesis h that $\mathcal{A}$ outputs

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

$$\Pr_{(x,y)\sim D'}[h(x) \neq y] \leq \frac{\epsilon}{2}$$

so we have:

$$\Pr_{(x,y)\sim D'}[h(x) \neq y] = \frac{1}{2} \cdot \Pr_{x\sim \mathcal{D}_c^+}[h(x) \neq 1] + \frac{1}{2} \cdot \Pr_{x\sim \mathcal{D}_c^-}[h(x) \neq 0] \leq \frac{\epsilon}{2}$$

Since both terms are non-negative, we must have:

$$\Pr_{x\sim \mathcal{D}_c^+}[h(x) \neq 1] \leq \epsilon \quad \text{and} \quad \Pr_{x\sim \mathcal{D}_c^-}[h(x) \neq 0] \leq \epsilon$$

Which is equivalent to:

$$\Pr_{x\sim \mathcal{D}_c^+}[h(x) = 0] \leq \epsilon \quad \text{and} \quad \Pr_{x\sim \mathcal{D}_c^-}[h(x) = 1] \leq \epsilon$$

Therefore, $\mathcal{C}$ is efficiently positively-negatively PAC learnable using $\mathcal{H}$ for the algorithm $\mathcal{A}_{pn}$. □

Question 2b

Proof. Suppose $\mathcal{C}$ is efficiently positively-negatively PAC learnable using $\mathcal{H}$. This means that for any $\epsilon, \delta > 0$, there exists a polynomial-time algorithm $\mathcal{A}$ that, given access to EX_c^+ and EX_c^-, outputs a hypothesis $h \in \mathcal{H} \cup \{h_0, h_1\}$ such that with probability at least $1-\delta$:

$$\Pr_{x\sim \mathcal{D}_c^+}[h(x) = 0] \leq \epsilon \quad \text{and} \quad \Pr_{x\sim \mathcal{D}_c^-}[h(x) = 1] \leq \epsilon$$

Here I assume that for this to hold we draw at least m- negative examples and at least m^+ positive examples from a polynomial in $1/\epsilon, 1/\delta$ from EX_c^- resp. $\mathrm{EX}^- + c$ such that .. holds.

Let $\mathcal{D}$ be some probability distribution over negative and positive examples and draw m from this distribution.
We do not know the numbers of negative and positive elements in the m sample. If we somehow could draw the m samples guaranteeing that the samples contains at least m- and at least m^+ samples then positively-negatively PAC-learning would imply standard PAC-learning:

$$\begin{aligned}
\Pr_{x\sim D}[h(x) \neq c(x)] &= \Pr_{x\sim D}[h(x) \neq c(x) \mid c(x) = 0] \Pr_{x\sim D}[c(x) = 0] \\
&\quad + \Pr_{x\sim D}[h(x) \neq c(x) \mid c(x) = 1] \Pr_{x\sim D}[c(x) = 1] \\
&\leq \epsilon\Big(\Pr_{x\sim D}[c(x) = 0] + \Pr_{x\sim D}[c(x) = 1]\Big) \\
&= \epsilon
\end{aligned}$$

since

$$\begin{aligned}
\Pr_{x\sim D}[h(x) \neq c(x) \mid c(x) = 0] &= \Pr_{x\sim D}[h(x) \neq 0 \mid c(x) = 0] \\
&= \Pr_{x\sim D}[h(x) = 1 \mid c(x) = 0] \\
&= \Pr_{x\sim \mathcal{D}_c^-}[h(x) = 1]
\end{aligned}$$

and similar for $\Pr_{x\sim D}[h(x) \neq c(x) \mid c(x) = 0]$.

I cannot figure out how to construct the m samples such that there at enough negative and positive samples. But intuitively is ought to be possible by selecting m large enough - at least in the case where $\mathcal{D}$ is not "to biased" toward either negative or positive samples. □

3 Growth Function

Question 3.1

Let $\mathcal{H}$ be a finite hypothesis set with $|\mathcal{H}| = M$ hypotheses. Prove that $m_{\mathcal{H}}(n) \leq \min\{M, 2^n\}$.

We can think of the bound by M as a cardinality bound and 2^n as a combinatorial bound.

Proof. $m_{\mathcal{H}}(n) \leq M$:

The growth function $m_{\mathcal{H}}(n)$ represents the maximum number of different ways n points can be labeled by hypotheses in $\mathcal{H}$. Since $\mathcal{H}$ contains exactly M different hypotheses, there can be at most M different labelings produced by these hypotheses on any set of n points. Therefore, $m_{\mathcal{H}}(n) \leq M$.

$m_{\mathcal{H}}(n) \leq 2^n$:

For any n points, the total number of possible dichotomies is 2^n, hence we have $m_{\mathcal{H}}(n) \leq 2^n$.

Since $m_{\mathcal{H}}(n)$ is bounded by both M and 2^n, we have $m_{\mathcal{H}}(n) \leq \min\{M, 2^n\}$. □

Question 3.2

Let $\mathcal{H}$ be a hypothesis space with 2 hypotheses (i.e., $|\mathcal{H}| = 2$). Prove $m_{\mathcal{H}}(n) = 2$.

Proof. From Question 3.1, we know that $m_{\mathcal{H}}(n) \leq \min\{M, 2^n\} = \min\{2, 2^n\} = 2$ for all $n \geq 1$.

Since $\mathcal{H}$ contains 2 distinct hypotheses that we will denote h_1 and h_2, they must disagree on the label of at least one input point (otherwise they would be the same hypothesis). Call this point $x^* \in \mathcal{X}$, so $h_1(x^*) \neq h_2(x^*)$.

Choose any sample S of size n that contains x^*. Then h_1 and h_2 label S differently, so $m_{\mathcal{H}}(n) \geq 2$ for all $n \geq 1$.

Combining both results, we have $m_{\mathcal{H}}(n) = 2$ for all $n \geq 1$. $\square$

Question 3.3

Prove that $m_{\mathcal{H}}(2n) \leq m_{\mathcal{H}}(n)^2$.

Proof. Take any set $S = \{x_1, \ldots, x_{2n}\}$ with $2n$ elements and partition it into two disjoint sets of equal size:

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

such that $S = S_1 \cup S_2$ and $S_1 \cap S_2 = \varnothing$.
For every hypothesis $h \in \mathcal{H}$ denote by d_h^S, $d_h^{S_1}$, $d_h^{S_2}$ the labelings it induces on S, S_1, S_2, respectively. Set

$$D(S) = \{d_h^S : h \in \mathcal{H}\}, \quad D(S_1) = \{d_h^{S_1} : h \in \mathcal{H}\}, \quad D(S_2) = \{d_h^{S_2} : h \in \mathcal{H}\}.$$

These are the sets of dichotomies. By definition of the growth function,

$$|D(S)| \leq m_{\mathcal{H}}(2n), \quad |D(S_1)| \leq m_{\mathcal{H}}(n), \quad |D(S_2)| \leq m_{\mathcal{H}}(n).$$

Define the Restriction map Φ as:

$$\Phi : D(S) \longrightarrow D(S_1) \times D(S_2), \qquad \Phi(d) = \big(d|_{S_1},\, d|_{S_2}\big).$$

where $d|_{S_i}$ is the restriction of the dichotomi to S_i. The mapping Φ is an injective map. Assume $\Phi(d) = \Phi(d')$ for two dichotomies $d, d' \in D(S)$. Then d and d' agree on both S_1 and S_2; since $S = S_1 \cup S_2$ is a disjoint union, they must also agree on all of S. Hence $d = d'$ and Φ is injective.

An injective map cannot increase cardinality, so

$$|D(S)| \;\leq\; |D(S_1) \times D(S_2)| \;=\; |D(S_1)|\,|D(S_2)| \;\leq\; m_{\mathcal{H}}(n)\, m_{\mathcal{H}}(n) \;=\; m_{\mathcal{H}}(n)^2.$$

The above bound holds for every sample S of size $2n$; taking the maximum over all such S yields the requested result

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

$\square$