MAC 6962 – Tópicos Matemáticos para Computação Contemporânea
Notas de aula

Leonardo Nagami Coregliano

10 de agosto de 2026

Estas notas de aula são baseadas em um curso desenvolvido por Leonardo Nagami Coregliano na Universidade de São Paulo. A edição atual se baseia em uma iteração do curso dada no segundo semestre de 2026.

Sumário

Capítulo 1

Concentração de distribuições

1.1 Modelando um problema de aprendizagem

Considere o seguinte problema de classificação:

Problema 1.1.1. Dada uma imagem, determinar (algoritmicamente) se na imagem há um gato.

Como não sabemos muito bem como definir o que é um "gato", poderíamos ao invés disso ter uma fase de pré-processamento na qual fornecemos várias imagens exemplo rotuladas como "gato" ou "não-gato" e esperar que o algoritmo determine uma "regra" do que é um "gato". Como modelamos isso matematicamente?

Notação 1.1.2. Seja X o conjunto (finito) de todas as imagens possíveis. Seja $Y \stackrel{\text{def}}{=} \{0,1\}$, onde 1 significa "gato" e 0 significa "não-gato". Uma "regra" é uma função $X \to Y$. Imaginamos que há uma função que é a "verdade" de classificação $F\colon X \to Y$ (desconhecida por nós), e nosso algoritmo será uma função que retorna uma regra $H\colon X \to Y$ dados m exemplos rotulados $(x_1, y_1), \ldots, (x_m, y_m) \in X \times Y$ (onde m é livre). Isso significa que nosso algoritmo é uma função da forma

$$\mathcal{A}\colon \bigcup_{m\in\mathbb{N}} X^m \times Y^m \to Y^X$$

(onde Y^X é o espaço de todas as funções $X \to Y$).

Problema 1.1.3 (Aprendizagem total)**.** Dado um conjunto finito X e tomando $Y \stackrel{\text{def}}{=} \{0,1\}$, encontrar um algoritmo

$$\mathcal{A}\colon \bigcup_{m\in\mathbb{N}} X^m \times Y^m \to Y^X$$

tal que existe um $m_{X,\mathcal{A}}$ grande o sufficiente tal que para todo $F\colon X \to Y$, e todo $m \geq m_X$, existem exemplos $(x_1, \ldots, x_m) \in X$ tal que

$$\mathcal{A}((x_i, F(x_i))) = F.$$

Quão grande tem de ser m_X?

Lema 1.1.4 (Uma obviedade)**.** *O Problema 1.1.3 tem uma solução $\mathcal{A}$ com $m_{X,\mathcal{A}} = |X|$. Mais especificamente, o algoritmo*

$$\mathcal{A}((x_i, y_i)_{i=1}^m)(x) \stackrel{\text{def}}{=} \begin{cases} y_i, & \text{se } i \in [m] \text{ é o mínimo indice com } x_i = x, \\ 0, & \text{caso contrário,} \end{cases} \tag{1.1}$$

é uma solução com $m_{XX,\mathcal{A}} = |X|$.

Demonstração. Note que se $m \geq m_{X,\mathcal{A}}$, então podemos passar todos os pontos de X para $\mathcal{A}$, portanto $\mathcal{A}$ retornará F. ■

Exercício 1.1.5. Assumindo $|X| \geq 2$, encontre uma solução $\mathcal{A}$ do Problema 1.1.3 com $m_{X,\mathcal{A}} = |X| - 1$.

Discussão 1.1.6. Note porém que a formulação do Problema 1.1.3 não é justa: não é muito realista que nós podessamo escolher os exemplos $x_1, \ldots, x_m$. (Quais seriam as melhores imagens de "gatos" e "não-gatos" para um algoritmo que nós não entedemos completamente?)

Ademais, nem todas as imagens (elementos de X) serão encontradas no dia-a-dia e também elas não aparecerão com a mesma frequência.

É mais natural imaginarmos que existe uma probabilidade μ sobre X (também desconhecida) que diz quão frequentemente as imagens ocorrem e que os nossos exemplos também vêm dessa probabilidade.

Mais especificamente, tenhamos um algoritmo fixado $\mathcal{A}$ e obtenhamos uma quantidade m de exemplos suficientemente grande a partir de (μ, F), isto é, sejam $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ sorteados i.i.d. (independentemente, identicamente distribuidos) de acordo com μ, e sejam $\boldsymbol{y}_i \stackrel{\text{def}}{=} F(\boldsymbol{x}_i)$ e seja $H \stackrel{\text{def}}{=} \mathcal{A}((\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1}^m)$, quão grande tem de ser m para termos uma chance de que $F = H$ μ-quase sempre (i.e., $\mu(\{x \in X \mid F(x) = H(x)\}) = 1$)?

Note que agora nós podemos ser azarados e obter exemplos ruins, e.g., com probabilidade baixa mas talvez não zero, poderíamos ter $\boldsymbol{x}_1 = \cdots = \boldsymbol{x}_m$, o que dificultaria deduzirmos o comportamento da função em outros pontos. Então é mais natural requerermos que estaremos corretos com alta probabilidade, i.e., probabilidade tão próxima de 1 quanto se queira.

Problema 1.1.7 (Aprendizagem provavelmente correta (PC))**.** Dado um conjunto X e tomando $Y \stackrel{\text{def}}{=} \{0, 1\}$, encontrar um algoritmo

$$\mathcal{A}\colon \bigcup_{m \in \mathbb{N}} X^m \times Y^m \to Y^X$$

tal que existe uma função $m_{X,\mathcal{A}}^{\text{PC}}\colon (0,1) \to \mathbb{N}$ tal que para toda probabilidade μ sobre X, todo $F\colon X \to Y$, todo $\delta \in (0,1)$ e todo $m \geq m_{X,\mathcal{A}}^{\text{PC}}(\delta)$, temos

$$\mathbb{P}_{\boldsymbol{x} \sim \mu^m}\left[\mathcal{A}\Big((\boldsymbol{x}_i, F(\boldsymbol{x}_i))_{i=1}^m\Big) = F \ \mu\text{-quase sempre}\right] \geq 1 - \delta.$$

Estudemos primeiramente uma variante do Problema 1.1.7 mais fácil, em que a probabilidade μ é uniforme:

Problema 1.1.8 (Aprendizagem provavelmente correta sob medida uniforme)**.** Dado um conjunto finito X e tomando $Y \stackrel{\text{def}}{=} \{0,1\}$, encontrar um algoritmo

$$\mathcal{A}\colon \bigcup_{m\in\mathbb{N}} X^m \times Y^m \to Y^X$$

tal que existe uma função $m^{\text{PCunif}}_{X,\mathcal{A}}\colon (0,1) \to \mathbb{N}$ tal que para a probabilidade μ uniforme sobre X, todo $F\colon X \to Y$, todo $\delta \in (0,1)$ e todo $m \geq m^{\text{PCunif}}_{X,\mathcal{A}}(\delta)$, temos

$$\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\Big[\mathcal{A}\big((\boldsymbol{x}_i, F(\boldsymbol{x}_i))_{i=1}^m\big) = F\Big] \geq 1-\delta.$$

Ideia: poderíamos tentar aplicar o Lema 1.1.4. Se virmos todos os pontos do espaço, o algoritmo do lema resolverá o problema. Isso motiva estudarmos o seguinte problema: queremos completar um álbum de figurinhas, se as figurinhas são distribuídas uniformemente, quantas figurinhas temos que comprar para completar o álbum?

Problema 1.1.9 (Coletor de Cupons[1] (Coupon Collector))**.** Seja X um conjunto finito e ponha $n \stackrel{\text{def}}{=} |X|$, seja μ a probabilidade uniforme em X e sejam $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ i.i.d. distribuídos de acordo com μ. Seja

$$\boldsymbol{T} \stackrel{\text{def}}{=} \min\{i \in \mathbb{N} \mid \forall x \in X, \exists j \leq i, x = \boldsymbol{x}_j\}$$

o tempo que leva até que todos os elementos de X aparecam na sequência aleatória $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$.

Quanto é $\mathbb{E}[\boldsymbol{T}]$? Podemos dar uma boa estimativa da forma $\mathbb{P}[\boldsymbol{T} \geq a] \leq b$?

Definição 1.1.10 (Valor esperado/esperança)**.** Dado um espaço de probabilidade enumerável (Ω, μ) é uma variável aleatória $\boldsymbol{T}\colon \Omega \to \mathbb{R}$ sobre Ω, o valor esperado de $\boldsymbol{T}$ é:

$$\mathbb{E}[\boldsymbol{T}(\omega)] = \sum_{\omega\in\Omega} \mu(\omega)\cdot\boldsymbol{T}(\omega).$$

Note que:

- (Linearidade da esperança) $\mathbb{E}[\boldsymbol{T_1} + a\cdot\boldsymbol{T_2}] = \mathbb{E}[\boldsymbol{T}_1] + a\cdot\mathbb{E}[\boldsymbol{T}_2]$.
- $|\mathbb{E}[\boldsymbol{T}]| \leq \mathbb{E}[|\boldsymbol{T}|]$.
- Se $\boldsymbol{T}_1, \ldots, \boldsymbol{T}_n$ são independentes, i.e., se

  $$\mathbb{P}[\boldsymbol{T}_1 = t_1, \ldots, \boldsymbol{T}_n = t_n] = \prod_{i=1}^n \mathbb{P}[\boldsymbol{T}_i = t_i],$$

  então $\mathbb{E}[\prod_{i=1}^n \boldsymbol{T}_i] = \prod_{i=1}^n \mathbb{E}[\boldsymbol{T}_i]$.
- (Probabilidade total) Para uma partição $(E_i)_{i\in I}$ de Ω em eventos, temos

  $$\mathbb{E}[\boldsymbol{T}] = \sum_{i\in I} \mathbb{E}[\boldsymbol{T} \mid E_i]\cdot\mathbb{P}[E_i].$$

[1] O nome mais conhecido do problema é esse, no qual estamos coletando cupons em vez de figurinhas.

Proposição 1.1.11. *No Problema 1.1.9, temos*[2]

$$\mathbb{E}[\boldsymbol{T}] = n \cdot H_n = n \cdot \sum_{i=1}^{n} \frac{1}{i} \leq n \cdot (\ln(n) + 1).$$

Demonstração. Para todo $i \in [n]$, seja $\boldsymbol{T}_k$ o tempo que leva para coletarmos nossa k-ésima figurinha única após termos $k-1$ figurinhas únicas, de forma que

$$\boldsymbol{T} = \sum_{k=1}^{n} \boldsymbol{T}_k.$$

Formalmente:

$$\boldsymbol{T}_k \stackrel{\text{def}}{=} \min\{i \in \mathbb{N} \mid |\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i\}| = k\} - \min\{i \in \mathbb{N} \mid |\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i\}| = k-1\}.$$

As variáveis $\boldsymbol{T}_k$ são independentes, mas não precisamos provar isso pois linearidade da esperança não requer independência:

$$\mathbb{E}[\boldsymbol{T}] = \sum_{k=1}^{n} \mathbb{E}[\boldsymbol{T}_k].$$

Fixe $k \in [n]$, qual é a distribuição de $\boldsymbol{T}_k$? Para $\boldsymbol{T}_k$, nós já temos $k-1$ figurinhas únicas, então cada vez que abrimos um novo pacote de figurinhas, a chance de ser uma figurinha nova é $p_k \stackrel{\text{def}}{=} (n-k+1)/n$ e cada novo pacote é independente do anterior. Isso significa que $\boldsymbol{T}_k$ tem *distribuição geometrica* $\text{Geom}(p_k)$:

$$\mathbb{P}[\boldsymbol{T}_k = t] = (1-p_k)^{t-1} \cdot p_k \qquad (t \in \mathbb{N}_+),$$

cuja esperança é fácil de se calcular:

$$\begin{aligned} \mathbb{E}[\boldsymbol{T}_k] &= \sum_{t \in \mathbb{N}_+} (1-p_k)^{t-1} \cdot p_k \cdot t = p_k \cdot \frac{d}{dx}\left(- \sum_{t \in \mathbb{N}_+} (1-x)^t \right)\Bigg|_{x=p_k} \\ &= p_k \cdot \frac{d}{dx}\left(-\frac{(1-x)}{1-(1-x)} \right)\Bigg|_{x=p_k} = p_k \cdot \frac{1}{x^2}\Big|_{x=p_k} = \frac{1}{p_k} = \frac{n}{n-k+1}. \end{aligned} \tag{1.2}$$

Portanto:

$$\mathbb{E}[\boldsymbol{T}] = \sum_{k=1}^{n} \mathbb{E}[\boldsymbol{T}_k] = \sum_{k=1}^{n} \frac{n}{n-k+1} = n \cdot H_n.$$

Finalmente, note que

$$H_n = \sum_{i=1}^{n} \frac{1}{i} \leq 1 + \int_1^n \frac{dx}{x} = \ln(n) + 1. \qquad \blacksquare$$

[2]O número H_n abaixo e chamado n-ésimo número harmônico.

Mas também gostaríamos de estimar a probabilidade de que $\boldsymbol{T}$ é muito maior do que o valor esperado. A ferramenta mais simples para se transformar uma cota no valor esperado numa cota de probabilidade é a desigualdade de Markov:

Proposição 1.1.12 (Markov)**.** *Se $\boldsymbol{X}$ é uma variável aleatória não-negativa ($\boldsymbol{X} \geq 0$) e $a > 0$, então*

$$\mathbb{P}[\boldsymbol{X} \geq a] \leq \frac{\mathbb{E}[\boldsymbol{X}]}{a}, \qquad\qquad \mathbb{P}[\boldsymbol{X} > a] < \frac{\mathbb{E}[\boldsymbol{X}]}{a}.$$

Demonstração. Provamos apenas a primeira desigualdade, pois a segunda tem uma prova análoga. Se pusermos $p \stackrel{\text{def}}{=} \mathbb{P}[\boldsymbol{X} \geq a]$, então

$$\mathbb{E}[\boldsymbol{X}] = \mathbb{E}[\boldsymbol{X} \mid \boldsymbol{X} \geq a] \cdot p + \mathbb{E}[\boldsymbol{X} \mid \boldsymbol{X} < a] \cdot (1 - p) \geq a \cdot p,$$

donde $p \leq \mathbb{E}[\boldsymbol{X}]/a$. ■

Nota 1.1.13. Tipicamente a desigualdade de Markov (Proposição 1.1.12) é aplicada com a sendo um múltiplo (maior que 1) do valor esperado: $a \stackrel{\text{def}}{=} (1 + c) \cdot \mathbb{E}[\boldsymbol{X}]$ para $c > 0$.

Corolário 1.1.14. *No Problema 1.1.9, para $c > 0$, temos*

$$\mathbb{P}\Big[\boldsymbol{T} \leq (1 + c) \cdot n \cdot (\ln(n) + 1)\Big] > 1 - \frac{1}{1 + c}.$$

Demonstração. Segue aplicando a Desigualdade de Markov (Proposição 1.1.12) ao resultado da Proposição 1.1.11:

$$\mathbb{P}\Big[\boldsymbol{T} > (1 + c) \cdot n \cdot (\ln(n) + 1)\Big] < \frac{\mathbb{E}[\boldsymbol{T}]}{(1 + c) \cdot n \cdot (\ln(n) + 1)} \leq \frac{1}{1 + c}.$$ ■

Corolário 1.1.15. *No Problema 1.1.8, o algoritmo $\mathcal{A}$ de* (1.1) *é uma solução com*

$$m^{\text{PCunif}}_{X,\mathcal{A}}(\delta) = \left\lceil \frac{|X| \cdot (\ln(|X|) + 1)}{\delta} \right\rceil.$$

Demonstração. Se $m \geq m^{\text{PCunif}}_{X,\mathcal{A}}(\delta)$, então pelo Corolário 1.1.14 com

$$c \stackrel{\text{def}}{=} \frac{1}{\delta} - 1,$$

temos que $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\} = X$ com probabilidade maior que

$$1 - \frac{1}{1 + c} = 1 - \delta.$$

Por sua vez, quando $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\} = X$, o algoritmo $\mathcal{A}$ retorna F. ■

Será que conseguimos provar alguma cota inferior? Isto é, conseguiríamos provar que qualquer algoritmo $\mathcal{A}$ que resolve o Problema 1.1.8 necessariamente deve possuir $m_{X,\mathcal{A}}(\delta)$ maior que alguma certa função de δ? Intuitivamente, $m_{X,\mathcal{A}}(\delta)$ muito menor que $|X|$ deve ser impossível, mas conseguiríamos provar isso?

Para isso usaremos o que às vezes é conhecida como Desigualdade de Markov Reversa:

Proposição 1.1.16 (Markov Reversa)**.** *Se $\boldsymbol{X}$ é uma variável aleatória e $a, c \in \mathbb{R}$ são tais que $\boldsymbol{X} \leq c$ e $a < c$, então*

$$\mathbb{P}[\boldsymbol{X} \leq a] \leq \frac{c - \mathbb{E}[\boldsymbol{X}]}{c - a}, \qquad \mathbb{P}[\boldsymbol{X} < a] < \frac{c - \mathbb{E}[\boldsymbol{X}]}{c - a}.$$

Demonstração. Provamos apenas a primeira desigualdade, pois a segunda tem uma prova análoga. Se pusermos $\boldsymbol{Y} \stackrel{\text{def}}{=} c - \boldsymbol{X}$, então $\boldsymbol{Y} \geq 0$, daí a Desigualdade de Markov (Proposição 1.1.12), temos

$$\mathbb{P}[\boldsymbol{X} \leq a] = \mathbb{P}[\boldsymbol{Y} \geq c - a] \leq \frac{\mathbb{E}[\boldsymbol{Y}]}{c - a} = \frac{c - \mathbb{E}[\boldsymbol{X}]}{c - a}.$$

■

Proposição 1.1.17. *No Problema 1.1.8, para todo algoritmo $\mathcal{A}$ e todo $m < |X|$, existe uma função $F\colon X \to Y$ tal que*

$$\mathbb{P}_{\boldsymbol{x} \sim \mu^m}\Big[\mathcal{A}\Big((\boldsymbol{x}_i, F(\boldsymbol{x}_i))_{i=1}^m\Big) = F\Big] < \frac{|X| - (|X| - m)/2}{|X| - 1}.$$

Demonstração. Queremos provar que

$$\max_{F\colon X \to Y} \mathbb{P}_{\boldsymbol{x} \sim \mu^m}\Big[\mathcal{A}\Big((\boldsymbol{x}_i, F(\boldsymbol{x}_i))_{i=1}^m\Big) = F\Big] < \frac{|X| - (|X| - m)/2}{|X| - 1}.$$

Para isso, vamos criar uma variável aleatória $\boldsymbol{L}$ que mede quantas entradas da função H retornada pelo algoritmo estão erradas:

$$L_{\mathcal{A}}(F, (x_i)_{i=1}^m) \stackrel{\text{def}}{=} \sum_{z \in X} \mathbb{1}\Big[\mathcal{A}\Big((x_i, F(x_i))_{i=1}^m\Big)(z) \neq F(z)\Big].$$

Nosso objetivo então é provar que

$$\max_{F\colon X \to Y} \mathbb{P}_{\boldsymbol{x} \sim \mu^m}[L(F, \boldsymbol{x}) \geq 1] > 1 - \frac{|X| - (|X| - m)/2}{|X| - 1}. \tag{1.3}$$

Para isso, vamos primeiro investigar o valor esperado: seja ν a probabilidade uniforme sobre todas as funções $X \to Y$ e note que

$$\begin{aligned} \max_{F\colon X \to Y} \mathbb{E}_{\boldsymbol{x} \sim \mu^m}\Big[L(F, \boldsymbol{x})\Big] &\geq \mathbb{E}_{\boldsymbol{F} \sim \nu}\Big[\mathbb{E}_{\boldsymbol{x} \sim \mu^m}\Big[L(\boldsymbol{F}, \boldsymbol{x})\Big]\Big] = \mathbb{E}_{\boldsymbol{x} \sim \mu^m}\Big[\mathbb{E}_{\boldsymbol{F} \sim \nu}\Big[L(\boldsymbol{F}, \boldsymbol{x})\Big]\Big] \\ &\geq \min_{x \in X^m} \mathbb{E}_{\boldsymbol{F} \sim \nu}\Big[L(\boldsymbol{F}, x)\Big]. \end{aligned} \tag{1.4}$$

Fixando $x \in X^m$, observamos que

$$\begin{aligned}\mathbb{E}_{\boldsymbol{F}\sim\nu}\Big[L_{\mathcal{A}}(\boldsymbol{F},x)\Big] &= \sum_{z\in X}\mathbb{P}\Big[\mathcal{A}\Big((x_i,\boldsymbol{F}(x_i))_{i=1}^m\Big)(z)\neq\boldsymbol{F}(z)\Big]\\ &\geq \sum_{\substack{z\in X\\ \forall i\in[m],z\neq x_i}}\mathbb{P}_{\boldsymbol{F}\sim\nu}\Big[\mathcal{A}\Big((x_i,\boldsymbol{F}(x_i))_{i=1}^m\Big)(z)\neq\boldsymbol{F}(z)\Big]\\ &= \sum_{\substack{z\in X\\ \forall i\in[m],z\neq x_i}}\frac{1}{2}\geq\frac{|X|-m}{2},\end{aligned}$$

que junto a (1.4) prova que

$$\max_{F\colon X\to Y}\mathbb{E}_{\boldsymbol{x}\sim\mu^m}\Big[L(F,\boldsymbol{x})\Big]\geq\frac{|X|-m}{2}.$$

Já que $L\leq|X|$, usando a Desigualdade de Markov Reversa (Proposição 1.1.16), temos

$$\begin{aligned}\max_{F\colon X\to Y}\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\Big[L(F,\boldsymbol{x})\geq 1\Big] &= \max_{F\colon X\to Y}\Big(1-\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\Big[L(F,\boldsymbol{x})<1\Big]\Big)\\ &> \max_{F\colon X\to Y}\left(1-\frac{|X|-\mathbb{E}_{\boldsymbol{x}\sim\mu^m}[L_{\mathcal{A}}(F,\boldsymbol{x})]}{|X|-1}\right)\\ &\geq 1-\frac{|X|-(|X|-m)/2}{|X|-1}\end{aligned}$$

que conclui a prova. ■