Método Heurístico Regularizado com Regra de Normalidade e Diretrizes para Ensemble Heurístico–Discriminativo

Guia didático e objetivo

3 de setembro de 2025

1 Método Heurístico Regularizado com Regra de Normalidade

Esta seção descreve, em linguagem direta, o método que você está usando hoje: um **modelo heurístico ajustado por dados** (mantendo a interpretabilidade) e uma **regra de baixa ativação** para detectar a classe *Sem Transtorno*.

Modelo (visão curta e precisa)

Dadas as respostas $X \in [0,1]^{n\times m}$ e a matriz de pesos $W \in [0,1]^{m\times K}$,

$$S = XW, \qquad p = \mathrm{softmax}(S). \tag{1}$$

Isto é equivalente a uma **logística multinomial sem termo de viés**. Cada linha de p é uma distribuição de probabilidades entre as K classes clínicas tradicionais.

Perda (o que o treinamento minimiza)

Queremos que as probabilidades concordem com o Alvo (multirrótulo permitido). Usamos a entropia cruzada média com **Elastic Net ancorado em** W_0:

$$\mathcal{L}(W) = -\frac{1}{n}\sum Y \log p \;+\; \lambda_1 \|W - W_0\|_1 \;+\; \lambda_2 \|W - W_0\|_2^2. \tag{2}$$

Interpretação: o primeiro termo foca em acertar; os termos ℓ_1 e ℓ_2 empurram para *mudar o mínimo possível* em relação ao W_0 original (preservando o desenho do questionário).

Restrições estruturais

1. Se uma coluna j de X tem **todos** os valores iguais a 0 ($X_{\cdot j} \equiv 0$), então **não alteramos** essa coluna de pesos: ela permanece exatamente como em W_0 (coluna *congelada*).
2. Para as demais colunas (*ajustáveis*), após cada passo projetamos os pesos para o intervalo $[\,\varepsilon, 1\,]$, evitando negativos/zeros e explosões.

Pós-regra: *Sem Transtorno* por baixa ativação

Ao final, aplicamos uma regra para introduzir a classe **Sem Transtorno**, sem aumentar a dimensionalidade de W:

- Se a maior probabilidade (top-1) é *baixa* ($p^{(1)} < T_1$) e a margem entre a primeira e a segunda é *pequena* ($p^{(1)} - p^{(2)} < T_2$), alocamos $p_{\text{normal}} = \gamma$ e reescalamos as demais por $(1-\gamma)$, mantendo soma 1.

- Os parâmetros (T_1, T_2, γ) são escolhidos por *grid search* para maximizar a **macro top-3** (incluindo a nova classe).

Leitura rápida das abas no Excel

- **Pontuação_Tunada:** W após o ajuste (colunas congeladas = W_0; ajustáveis em $[\varepsilon, 1]$).
- **Resultado_Heuristica_Tunada:** probabilidades $p_{\langle classe \rangle}$, $p_{\text{Sem Transtorno}}$ e ranking top-1/top-2/top-3 (já com a classe extra).
- **Metricas_Heuristica_Tunada:** macro top-3 e tabela por classe (taxa top-3, suporte).
- **Regras_Normal:** melhores T_1, T_2, γ e taxa de acionamento.

2 Diretrizes para construir um *Ensemble* Heurístico–Discriminativo

Se desejar evoluir para ensemble, a ideia é combinar a probabilidade da heurística tunada ($p^{(H)}$) com a de um **modelo discriminativo** treinado ($p^{(M)}$), e só então aplicar a mesma regra de *Sem Transtorno*.

O que usar em $p^{(M)}$ (modelo discriminativo)

- **Recomendado:** Logística multinomial (`multi_class='multinomial'`, `solver='lbfgs'`, `class_weight='balanced'`): simples, estável com poucos dados, bem calibrável e combina bem com a heurística.
- **Alternativas:** Random Forest (boa capacidade, mas geralmente menos calibrada) ou Gradient Boosting/XGBoost/LightGBM (modelos fortes; calibrar depois).
- **Calibração:** Platt, isotônica ou temperatura antes do blending tende a melhorar a qualidade das probabilidades.

Mistura convexa e regra de normalidade

$$p \;=\; (1-\alpha)\, p^{(H)} \;+\; \alpha\, p^{(M)}, \qquad \alpha \in [0,1], \tag{3}$$

com α escolhido por validação para maximizar a *macro top-3*. Depois, aplicamos a **mesma regra** do *Sem Transtorno* (baixa ativação) sobre p.

Frase pronta (para o relatório). *"Ensemble Heurístico–Discriminativo: $p = (1-\alpha)p^{(H)} + \alpha p^{(M)}$ (com α escolhido por validação para maximizar a macro top-3), seguido da regra de baixa ativação para 'Sem Transtorno'."*

A Apêndice A: Aprendizado supervisionado com entropia cruzada e Elastic Net

Nesta seção, detalhamos como o ajuste de W é feito, mantendo o método interpretável e próximo da sua heurística original W_0.

O que está sendo treinado

Temos $X \in [0,1]^{n \times m}$ (respostas) e $W \in [0,1]^{m \times K}$ (pesos). Calculamos

$$S = XW, \qquad P_{i,:} = \mathrm{softmax}(S_{i,:}). \tag{4}$$

Queremos que P concorde com os rótulos (Alvo). Para lidar com *multirrótulo*, usamos uma distribuição alvo Y por linha (soma 1 entre os rótulos positivos).

Entropia cruzada (cross-entropy)

A perda média é

$$\mathrm{CE}(W) \;=\; -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} Y_{ik} \, \log P_{ik}, \tag{5}$$

diminuindo quando o modelo atribui alta probabilidade às classes corretas.

Elastic Net ancorado em W_0

Para evitar overfitting e permanecer próximo da heurística,

$$\lambda_1 \|W - W_0\|_1 \;+\; \lambda_2 \|W - W_0\|_2^2. \tag{6}$$

O termo $L2$ favorece mudanças suaves; o termo $L1$ zera muitos desvios $(W - W_0)$, preservando a estrutura.

Função objetivo completa

$$\mathcal{L}(W) \;=\; \mathrm{CE}(W) \;+\; \lambda_1 \|W - W_0\|_1 \;+\; \lambda_2 \|W - W_0\|_2^2. \tag{7}$$

Passo de atualização (ideia)

1. **Gradiente da CE:** $G = \frac{1}{n} X^\top (P - Y)$.
2. **Parte suave (CE + L2):** $W_{\text{tent}} = W - \eta \left(G + 2\lambda_2 (W - W_0) \right)$.
3. **Prox L1 no desvio:** $\Delta = W_{\text{tent}} - W_0 \Rightarrow \Delta \leftarrow \text{soft-threshold}(\Delta, \eta\lambda_1)$; então $W_{\text{novo}} = W_0 + \Delta$.
4. **Projeção de restrições:** colunas sem sinal em X ficam exatamente como W_0 (congeladas); demais são *clipadas* para $[\varepsilon, 1]$.

Escolha de λ_1, λ_2 e dicas práticas

- Aumente λ_2 se notar pesos "explodindo" (suaviza).
- Aumente λ_1 se quiser mudar menos (mais desvios zerados $\rightarrow$ mais perto de W_0).
- Faça um grid pequeno e selecione por *macro top-3* e inspeção de interpretabilidade.
- Se houver desequilíbrio de classes, é possível ponderar a CE por classe.

Diagnóstico e parada

Monitore a *macro top-3* ao longo das iterações; use *early stopping* quando estabilizar. Inspecione colunas antes/depois para verificar coerência clínica.

B Apêndice B: Softmax detalhado

A função **softmax** transforma um vetor de escores (logits) $s = (s_1, \ldots, s_K)$ em probabilidades $p = (p_1, \ldots, p_K)$, com $p_k \in (0, 1)$ e $\sum_k p_k = 1$:

$$\text{softmax}(s)_k \;=\; \frac{e^{s_k}}{\sum_{t=1}^{K} e^{s_t}}\,. \tag{8}$$

Para estabilidade numérica, subtrai-se $\max(s)$ (não altera o resultado final):

$$\text{softmax}(s)_k \;=\; \frac{e^{\,s_k - \max(s)}}{\sum_{t=1}^{K} e^{\,s_t - \max(s)}}\,. \tag{9}$$

Intuição. As exponenciais ampliam diferenças; a normalização força a soma = 1 e torna as classes comparáveis.

Temperatura (opcional).

$$\text{softmax}_T(s)_k \;=\; \frac{e^{s_k/T}}{\sum_t e^{s_t/T}}, \tag{10}$$

com $T < 1$ deixando a distribuição mais "dura" (top-1 maior) e $T > 1$ mais "suave".