Classificador logístico

Gaspar J. Machado

Departamento de Matemática, Universidade do Minho

março de 2024

Perceptron

- Contexto: classificador binário linear (determinístico).
- Base de dados $D = (x^n, y^n)_{n=1}^N$, $x^n \in \mathbb{R}^I$, $y^n \in \{-1, +1\}$.
- Espaço das hipóteses: $\mathcal{H} = \mathbb{R}^{I+1}$.
- Arquitetura da *Machine Learning*

$$\hat{y} \equiv h(x; \tilde{w}) = \mathrm{sgn}(\tilde{w} \cdot \tilde{x}), \tilde{w} \in \mathcal{H}.$$

- Se D é linearmente separável, o algoritmo *Perceptron* determina em tempo finito um classificador $\tilde{w} \in \mathbb{R}^{I+1}$ tal que o erro de treino é 0, ou seja, tal que

$$E(\tilde{w}, D) \equiv \frac{1}{N} \sum_{n=1}^{N} \frac{1}{2} |y^n - \hat{p}^n| = 0.$$

- O algoritmo *Perceptron* não usa técnicas de otimização mas antes uma regra de aprendizagem muito simples mas muito engenhosa.
- É possível aplicar técnicas de otimização para minimizar a função custo E, mas é um processo muito difícil devido à baixa regularidade da função E (a função custo pode ter o mesmo valor para classificadores muito distintos, ou seja, E não varia suavemente com uma variação dos parâmetros $\tilde{w}$).

CLog

- Nova proposta: Classificador logístico (CLog).
- Contexto: classificador binário linear (probabilístico).
- Base de dados $D = (x^n, y^n)_{n=1}^N$, $x^n \in \mathbb{R}^I$, $y^n \in \{0, 1\}$.
- Espaço das hipóteses: $\mathcal{H} = \mathbb{R}^{I+1}$.
- Arquitetura da *Machine Learning*

$$\hat{p} \equiv h(x; \tilde{w}) = \sigma(\tilde{w} \cdot \tilde{x}), \tilde{w} \in \mathcal{H}.$$

- Função logística ou função sigmóide

$$\sigma(z) = \frac{1}{1 + \exp(-z)}.$$

- Relativamente ao *Perceptron*, a otimização do CLog é mais fácil e o *output* do CLog contém mais informação ($(0, 1)$ *versus* $\{0, 1\}$).

Função logística

- Definição

$$\sigma(z) = \frac{1}{1 + \exp(-z)}.$$

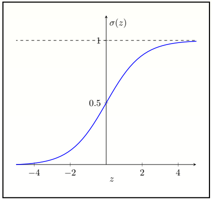
- Propriedades

$$\lim_{z \to -\infty} \sigma(z) = 0, \quad \sigma(0) = 0.5, \quad \lim_{z \to +\infty} \sigma(z) = 1.$$

$$\sigma'(z) = \frac{\exp(-z)}{(1 + \exp(-z))^2} > 0 \Rightarrow \sigma \text{ é uma função monótona crescente.}$$

$$\sigma(z) \in (0, 1).$$

- Gráfico



CLog

- Interpretação da arquitetura do CLog $\hat{p} \equiv h(x; \tilde{w}) = \sigma(\tilde{w} \cdot \tilde{x})$: a probabilidade da classe de x ser 1.
- E como funciona o classificador? Uma possibilidade para prever a classe é:

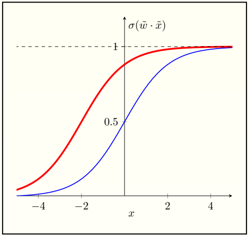
$$\hat{y} = \begin{cases} 1, & \text{se } \hat{p} > 0.5, \\ 0, & \text{se } \hat{p} \leqslant 0.5. \end{cases}$$

- Dependendo do problema, o valor de corte (*threshold*) do classificador $\varepsilon_{\text{CLog}}$ pode assumir outros valores no intervalo $(0, 1)$, tendo-se

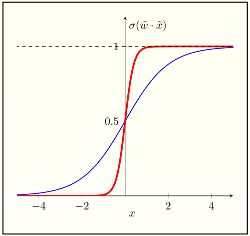
$$\hat{y} = \begin{cases} 1, & \text{se } \hat{p} > \varepsilon_{\text{CLog}}, \\ 0, & \text{se } \hat{p} \leqslant \varepsilon_{\text{CLog}}. \end{cases}$$

Arquitetura do CLog: $\hat{p} \equiv h(x; \tilde{w}) = \sigma(\tilde{w} \cdot \tilde{x}),\ \tilde{w} \in \mathbb{R}^{1+1}$

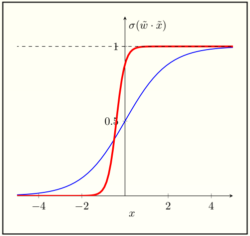
$\tilde{w} = (2, 1)^\top$



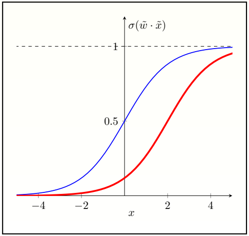
$\tilde{w} = (0, 5)^\top$



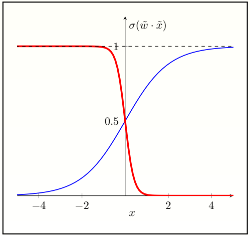
$\tilde{w} = (2, 5)^\top$



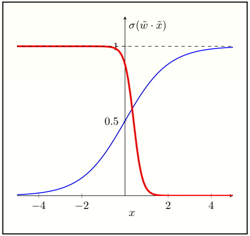
$\tilde{w} = (-2, 1)^\top$



$\tilde{w} = (0, -5)^\top$



$\tilde{w} = (2, -5)^\top$



Função custo do CLog (i)

- Ideia: o classificador é bom se atribui uma probabilidade $\hat{p}^n$ (valor predito) próxima da classe y^n, ou seja, se é grande (próximo de 1) o valor da expressão

$$L(\tilde{w}; D) = \prod_{n=1}^{N} \begin{cases} \hat{p}^n, & \text{se } y^n = 1 \\ 1 - \hat{p}^n, & \text{se } y^n = 0 \end{cases} = \prod_{n=1}^{N} \left(\hat{p}^n\right)^{y^n} \left(1 - \hat{p}^n\right)^{1-y^n}.$$

- Como a função logaritmo é monótona crescente, maximizar L é equivalente a maximizar $\ln L$

$$\ln L = \frac{1}{N} \sum_{n=1}^{N} y^n \ln(\hat{p}^n) + (1 - y^n) \ln(1 - \hat{p}^n),$$

- ou ainda a minimizar a função custo E

$$E(\tilde{w}; D) = \frac{1}{N} \sum_{n=1}^{N} E_n(\tilde{w}; x^n, y^n),$$

$$E_n(\tilde{w}; x^n, y^n) = -y^n \ln(\hat{p}^n) - (1 - y^n) \ln(1 - \hat{p}^n).$$

Função custo do CLog (ii)

- À função custo

$$E_n(\tilde{w}; x^n, y^n) = -y^n \ln(\hat{p}^n) - (1 - y^n) \ln(1 - \hat{p}^n)$$

  chama-se "entropia cruzada" (*cross entropy* ou *negative log-likelihood* ou *log loss*).
- Exemplos

y^n	1	1	1	1	1
$\hat{p}^n$	0.001	0.01	0.5	0.99	0.999
E_n	6.9078	4.6052	0.6932	0.0101	0.0010

y^n	0	0	0	0	0
$\hat{p}^n$	0.001	0.01	0.5	0.99	0.999
E_n	0.0010	0.0101	0.6932	4.6052	6.9078

Método do Gradiente | $f : \mathbb{R} \to \mathbb{R}$

Input: $(f : \mathbb{R} \to \mathbb{R},)\ f',\ x_{(0)} \in \mathbb{R},\ \eta \in \mathbb{R}^+,\ CP$
Output: $x^* \in \mathbb{R}$

$t \leftarrow 0$;
while V **do**
    $s_{(t)} \leftarrow f'(x_{(t)})$;
    $x_{(t+1)} \leftarrow x_{(t)} - \eta s_{(t)}$;
    if $CP = V$ **then**
        $x^* \leftarrow x_{(t+1)}$; **return** x^*;
    else
        $t \leftarrow t + 1$;

Método do Gradiente | $f : \mathbb{R}^M \to \mathbb{R}$

Input: $(f : \mathbb{R}^M \to \mathbb{R},)$ ∇f, $x_{(0)} \in \mathbb{R}^M$, $\eta \in \mathbb{R}^+$, CP
Output: $x^* \in \mathbb{R}^M$

$t \leftarrow 0$;
while *V* **do**
    $s_{(t)} \leftarrow \nabla f(x_{(t)})$;
    $x_{(t+1)} \leftarrow x_{(t)} - \eta s_{(t)}$;
    if CP=*V* **then**
        $x^* \leftarrow x_{(t+1)}$; **return** x^*;
    else
        $t \leftarrow t + 1$;

Método do Gradiente | CP

- Critérios de paragem (CP):
  - $t = t_{\max}$.
  - $|f(x_{(t+1)}) - f(x_{(t)})| \leqslant \varepsilon$ ou $\frac{|f(x_{(t+1)})-f(x_{(t)})|}{|f(x_{(t+1)})|} \leqslant \varepsilon$.
  - $\|x_{(t+1)} - x_{(t)}\| \leqslant \varepsilon$ ou $\frac{\|x_{(t+1)}-x_{(t)}\|}{\|x_{(t+1)}}\| \leqslant \varepsilon$.
  - $M = 1$: $|f'(x_{(t+1)})| \leqslant \varepsilon$.
  - $M > 1$: $\|\nabla f(x_{(t+1)})\| \leqslant \varepsilon$.

Método do Gradiente | ML

- Muitas vezes as funções objetivo que se encontram em ML são da forma

$$E(x) = \frac{1}{N} \sum_{n=1}^{N} E_n(x),$$

em que E_n é uma função real de M variáveis reais $E_n : \mathbb{R}^M \to \mathbb{R}$.

- Então, o método do gradiente assume três versões:
  - Método do Gradiente *batch* — algoritmo **MGB**: consideram-se **todas** as funções E_n por iteração.
  - Método do Gradiente *mini-batch* — algoritmo **MGmB**: considera-se um **subconjunto** das funções E_n com $\boldsymbol{B}$ elementos por iteração.
  - Método do Gradiente estocástico — algoritmo **MGE**: considera-se **uma função** aleatoriamente escolhida E_n por iteração.

Algoritmo MGB

Input: $(E(x) = \frac{1}{N}\sum_{n=1}^{N} E_n(x),\ E_n(x) : \mathbb{R}^M \to \mathbb{R},)\ \nabla E_1,\ \ldots,$ $\nabla E_N,\ x_{(0)} \in \mathbb{R}^M,\ \eta \in \mathbb{R}^+,\ CP$

Output: $x^* \in \mathbb{R}^M$

$t \leftarrow 0$;

while *V* **do**

$$s_{(t)} \leftarrow \frac{1}{N}\sum_{n=1}^{N} \nabla E_n(x_{(t)});$$

$x_{(t+1)} \leftarrow x_{(t)} - \eta s_{(t)}$;

if CP=*V* **then**

$x^* \leftarrow x_{(t+1)}$; **return** x^*;

else

$t \leftarrow t + 1$;

Algoritmo MGmB

Input: $(E(x) = \frac{1}{N}\sum_{n=1}^{N} E_n(x),\ E_n(x) : \mathbb{R}^M \to \mathbb{R},)\ \nabla E_1,\ \ldots,$ $\nabla E_N,\ x_{(0)} \in \mathbb{R}^M,\ \eta \in \mathbb{R}^+,\ B \in \mathbb{N},\ CP$

Output: $x^* \in \mathbb{R}^M$

```
t ← 0;
while V do
    gerar aleatoriamente um subconjunto de índices
      D_(t) ⊂ {1, …, N} com B elementos;
    s_(t) ← (1/B) Σ_{n∈D_(t)} ∇E_n(x_(t));
    x_(t+1) ← x_(t) − η s_(t);
    if CP=V then
        x* ← x_(t+1); return x*;
    else
        t ← t + 1;
```

$$s_{(t)} \leftarrow \frac{1}{B}\sum_{n\in D_{(t)}} \nabla E_n(x_{(t)})$$

Algoritmo MGE

Input: $(E(x) = \frac{1}{N}\sum_{n=1}^{N} E_n(x),\ E_n(x) : \mathbb{R}^M \to \mathbb{R},)\ \nabla E_1,\ \ldots,$ $\nabla E_N,\ x_{(0)} \in \mathbb{R}^M,\ \eta \in \mathbb{R}^+,\ CP$
Output: $x^* \in \mathbb{R}^M$

$t \leftarrow 0$;
while V **do**
    gerar aleatoriamente um índice $n \in \{1, \ldots, N\}$;
    $s_{(t)} \leftarrow \nabla E_n(x_{(t)})$;
    $x_{(t+1)} \leftarrow x_{(t)} - \eta s_{(t)}$;
    if $CP{=}V$ **then**
        $x^* \leftarrow x_{(t+1)}$; **return** x^*;
    else
        $t \leftarrow t + 1$;

Gradiente do CLog

- Para se aplicar o Método do Gradiente ao CLog tem que se calcular o gradiente da função custo E

$$E(\tilde{w}; D) = \frac{1}{N} \sum_{n=1}^{N} E_n(\tilde{w}; x^n, y^n),$$

$$E_n(\tilde{w}; x^n, y^n) = -y^n \ln(\hat{p}^n) - (1 - y^n) \ln(1 - \hat{p}^n).$$

- Pode-se mostrar (exercício) que

$$\nabla E_n(\tilde{w}; x^n, y^n) = (\hat{p}^n - y^n)\tilde{x}^n,$$

$$\nabla E(\tilde{w}; D) = \frac{1}{N} \sum_{n=1}^{N} \nabla E_n(\tilde{w}; x^n, y^n) = \frac{1}{N} \sum_{n=1}^{N} (\hat{p}^n - y^n)\tilde{x}^n.$$

- Vamos denotar por **CLog–MGB** o algoritmo do classificador logístico com o MGB a por **CLog–MGE** o algoritmo do classificador logístico com o MGE.

Algoritmo CLog–MGB

Input: $D = (x^n, y^n)_{n=1}^N,\ x^n \in \mathbb{R}^I,\ y^n \in \{0,1\},\ \tilde{w}_{(0)} \in \mathbb{R}^{I+1},$
$\eta \in \mathbb{R}^+,\ CP$

Output: $\tilde{w}^* \in \mathbb{R}^{I+1}$

$t \leftarrow 0$;

while *V* **do**

  $\hat{p}^n \leftarrow \sigma\left(\tilde{w}_{(t)} \cdot \tilde{x}^n\right),\ n = 1, \ldots, N;$

  $s_{(t)} \leftarrow \dfrac{1}{N} \displaystyle\sum_{n=1}^{N} \left(\hat{p}^n - y^n\right) \tilde{x}^n;$

  $\tilde{w}_{(t+1)} \leftarrow \tilde{w}_{(t)} - \eta s_{(t)};$

  if CP=*V* **then**

    $\tilde{w}^* \leftarrow \tilde{w}_{(t+1)}$; **return** $\tilde{w}^*$;

  else

    $t \leftarrow t + 1;$

Algoritmo CLog–MGE

Input: $D = (x^n, y^n)_{n=1}^{N}$, $x^n \in \mathbb{R}^I$, $y^n \in \{0,1\}$, $\tilde{w}_{(0)} \in \mathbb{R}^{I+1}$, $\eta \in \mathbb{R}^+$, CP

Output: $\tilde{w}^* \in \mathbb{R}^{I+1}$

$t \leftarrow 0$;
while V **do**
  selecionar $n \in \{1, \ldots, N\}$ aleatório;
  $\hat{p}^n \leftarrow \sigma\left(\tilde{w}_{(t)} \cdot \tilde{x}^n\right)$;
  $s_{(t)} \leftarrow (\hat{p}^n - y^n)\, \tilde{x}^n$;
  $\tilde{w}_{(t+1)} \leftarrow \tilde{w}_{(t)} - \eta s_{(t)}$;
  if $CP{=}\mathsf{V}$ **then**
    $\tilde{w}^* \leftarrow \tilde{w}_{(t+1)}$; **return** $\tilde{w}^*$;
  else
    $t \leftarrow t + 1$;

Perceptron vs CLog–MGE

- *Perceptron*

$$
\begin{aligned}
\hat{y}^n &\leftarrow \mathrm{sgn}(\tilde{w}_{(t)} \cdot \tilde{x}^n) \\
\tilde{w}_{(t+1)} &\leftarrow \tilde{w}_{(t)} + \frac{\eta}{2}(y^n - \hat{y}^n)\tilde{x}^n
\end{aligned}
$$

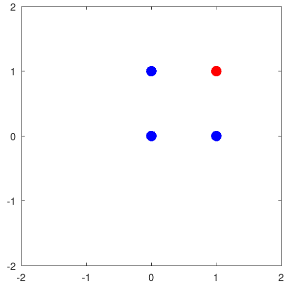
- CLog–MGE

$$
\begin{aligned}
\hat{p}^n &\leftarrow \sigma\left(\tilde{w}_{(t)} \cdot \tilde{x}^n\right) \\
\tilde{w}_{(t+1)} &\leftarrow w_{(t)} + \eta\left(y^n - \hat{p}^n\right)\tilde{x}^n
\end{aligned}
$$

ex1 (AND) | CLog–MGE

- Aplicar o algoritmo CLog–MGE à base de dados "AND":
  $D = (x^n, y^n)_{n=1}^{4}$ com
  $x^1 = (0,0)^\top$, $y^1 = 0$ (F) $\quad$ $x^2 = (0,1)^\top$, $y^2 = 0$ (F)
  $x^3 = (1,0)^\top$, $y^3 = 0$ (F) $\quad$ $x^4 = (1,1)^\top$, $y^4 = 1$ (V)

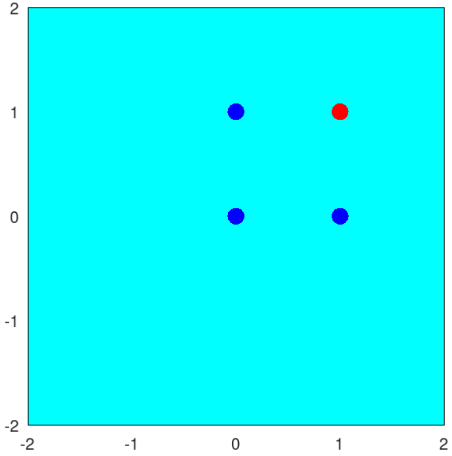
- Aplicar o algoritmo CLog–MGE à base de dados "AND":
  $D = (x^n, y^n)_{n=1}^4$ com
  $x^1 = (0,0)^\top$, $y^1 = 0$ (F) $\quad$ $x^2 = (0,1)^\top$, $y^2 = 0$ (F)
  $x^3 = (1,0)^\top$, $y^3 = 0$ (F) $\quad$ $x^4 = (1,1)^\top$, $y^4 = 1$ (V)



- CP: $t_{\text{MAX}} = 50$.
- Taxa de aprendizagem: $\eta = 0.5$.
- Aproximação inicial: $\tilde{w}_{(0)} = (0,0,0)^\top$.
- Valor de corte: $\varepsilon_{\text{CLog}} = 0.5$.

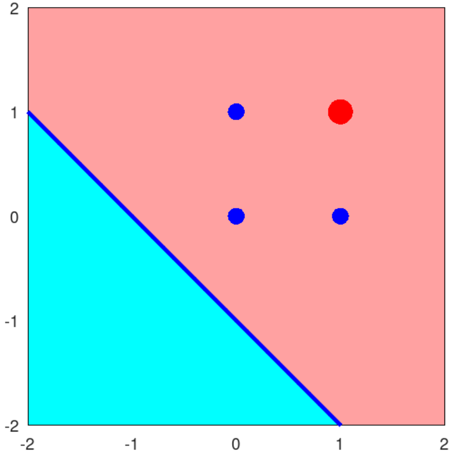
ex1 (AND) | CLog–MGE | $t = 0$

- $\tilde{w}_{(0)} = (0, 0, 0)^\top$, $E_{(0)} = 0.69315$



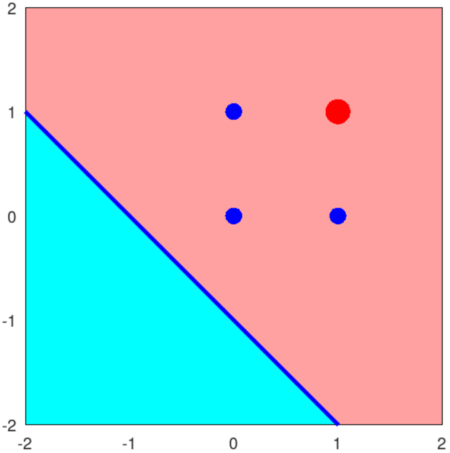
ex1 (AND) | CLog–MGE | $t = 1$

- $\tilde{w}_{(1)} = (0.25, 0.25, 0.25)^\top, E_{(1)} = 0.79024$



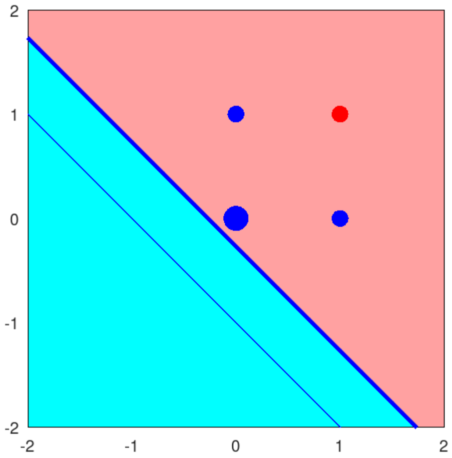
ex1 (AND) | CLog–MGE | $t = 2$

- $\tilde{w}_{(2)} = (0.41041, 0.41041, 0.41041)^\top, E_{(2)} = 0.88661$



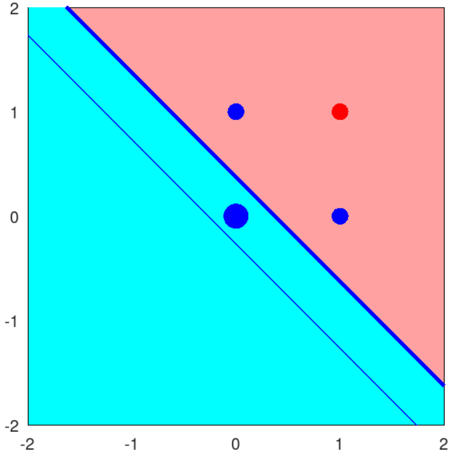
ex1 (AND) | CLog–MGE | $t = 3$

- $\tilde{w}_{(3)} = (0.10982, 0.41041, 0.41041)^\top, E_{(3)} = 0.76385$



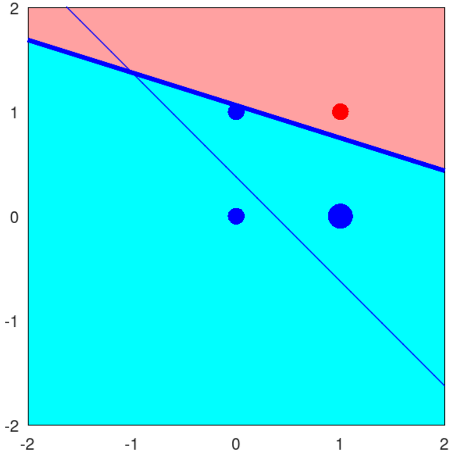
ex1 (AND) | CLog–MGE | $t = 4$

- $\tilde{w}_{(4)} = (-0.1539, 0.41041, 0.41041)^\top, E_{(4)} = 0.67316$

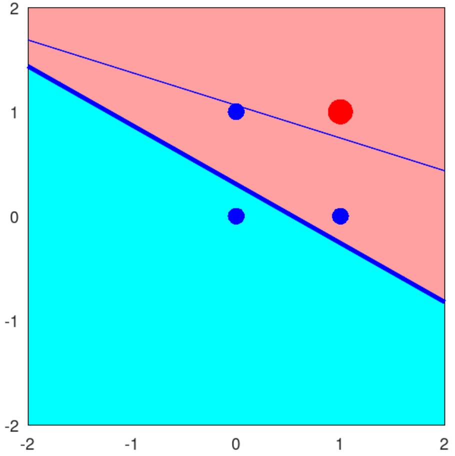


ex1 (AND) | CLog–MGE | $t = 5$

- $\tilde{w}_{(5)} = (-0.43579, 0.12852, 0.41041)^\top, E_{(5)} = 0.59338$

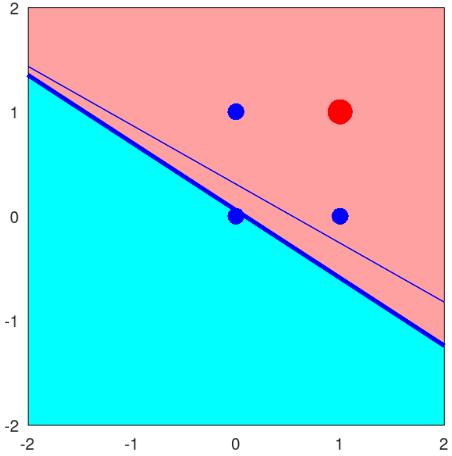


- $\tilde{w}_{(6)} = (-0.19867, 0.36564, 0.64753)^\top, E_{(6)} = 0.67201$

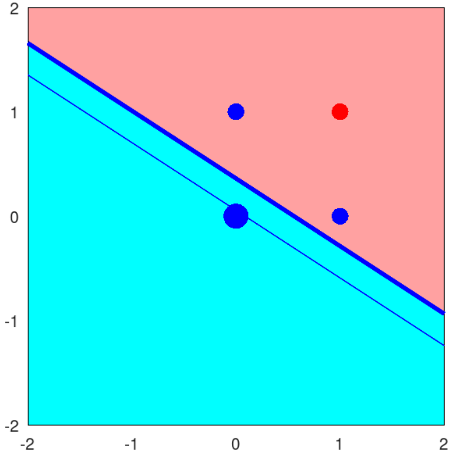


ex1 (AND) | CLog–MGE | $t = 7$

- $\tilde{w}_{(7)} = (-0.045201, 0.51911, 0.80099)^\top, E_{(7)} = 0.75399$

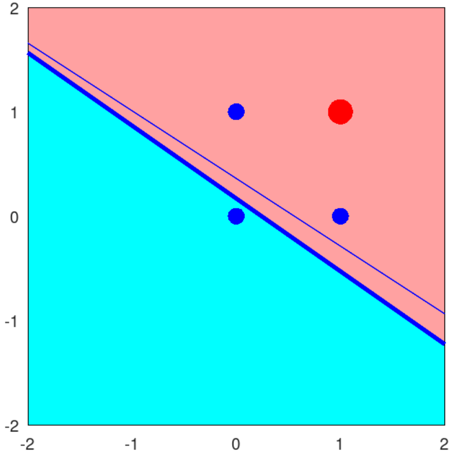


- $\tilde{w}_{(8)} = (-0.28955, 0.51911, 0.80099)^\top, E_{(8)} = 0.66492$



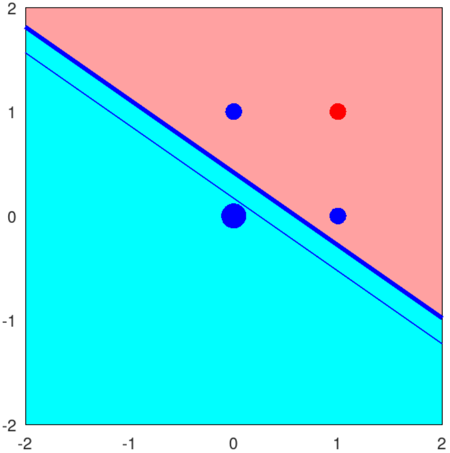
ex1 (AND) | CLog–MGE | $t = 9$

- $\tilde{w}_{(9)} = (-0.15806, 0.65059, 0.93248)^\top, E_{(9)} = 0.73893$



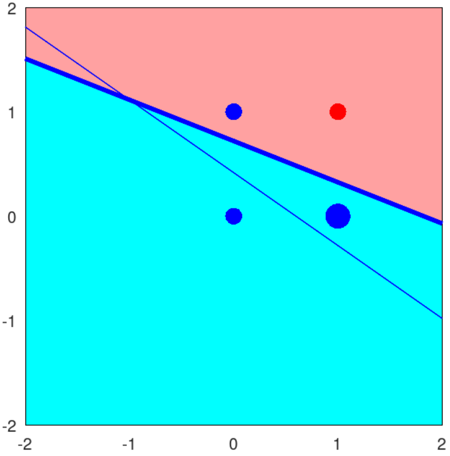
ex1 (AND) | CLog–MGE | $t = 10$

- $\tilde{w}_{(10)} = (-0.38835, 0.65059, 0.93248)^\top, E_{(10)} = 0.65421$



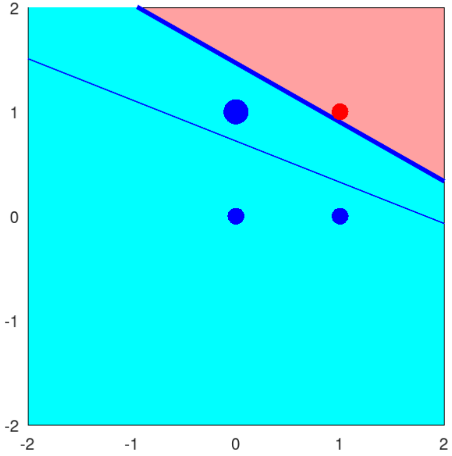
ex1 (AND) | CLog–MGE | $t = 11$

- $\tilde{w}_{(11)} = (-0.67094, 0.368, 0.93248)^\top, E_{(11)} = 0.5564$

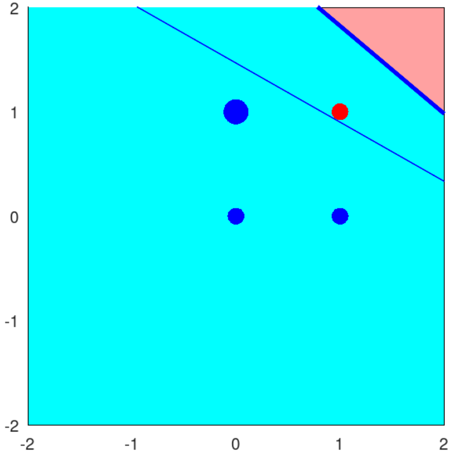


ex1 (AND) | CLog–MGE | $t = 12$

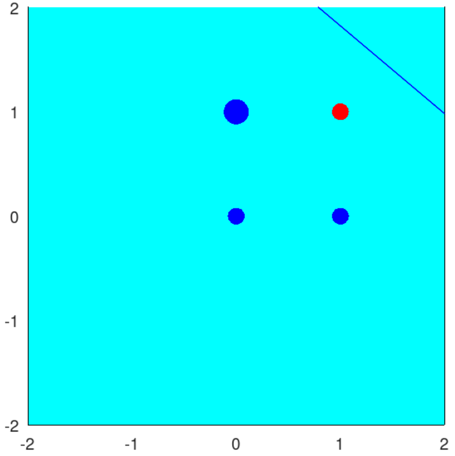
- $\tilde{w}_{(12)} = (-0.95345, 0.368, 0.64998)^\top, E_{(12)} = 0.49574$



- $\tilde{w}_{(13)} = (-1.1658, 0.368, 0.43762)^\top, E_{(13)} = 0.48159$

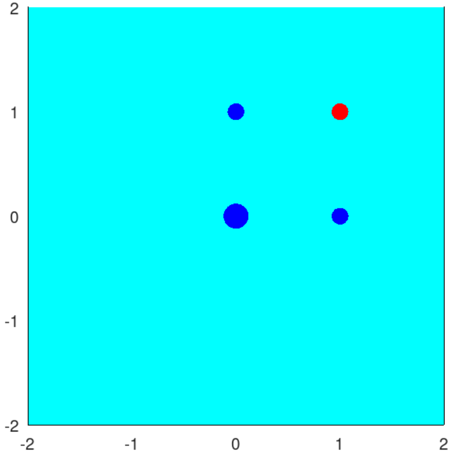


- $\tilde{w}_{(14)} = (-1.3286, 0.368, 0.27482)^\top, E_{(14)} = 0.48794$

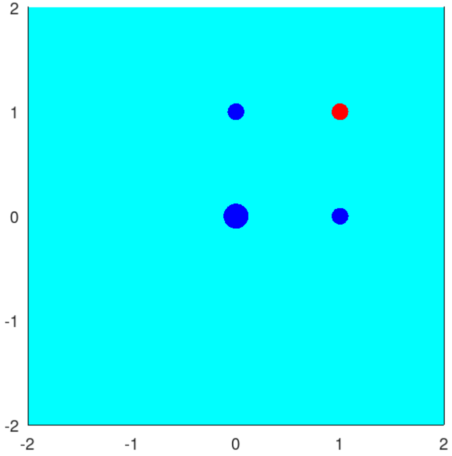


ex1 (AND) | CLog–MGE | $t = 15$

- $\tilde{w}_{(15)} = (-1.4333, 0.368, 0.27482)^\top$, $E_{(15)} = 0.4869$

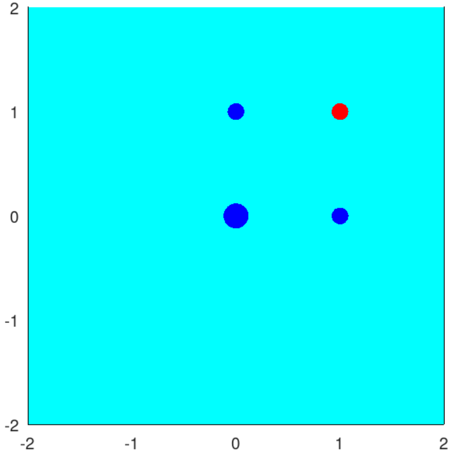


- $\tilde{w}_{(16)} = (-1.5296, 0.368, 0.27482)^\top$, $E_{(16)} = 0.48775$



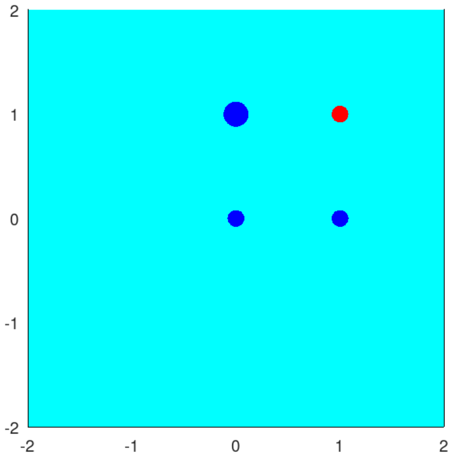
ex1 (AND) | CLog–MGE | $t = 17$

- $\tilde{w}_{(17)} = (-1.6186, 0.368, 0.27482)^\top$, $E_{(17)} = 0.49$



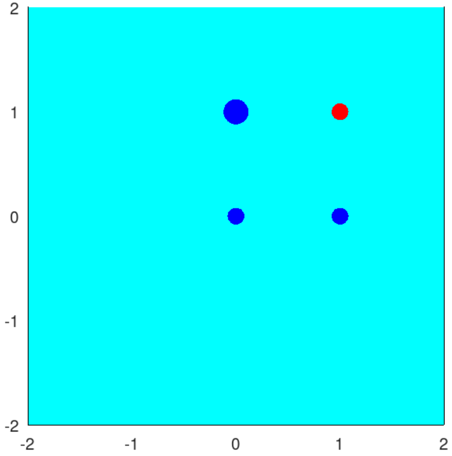
ex1 (AND) | CLog–MGE | $t = 18$

- $\tilde{w}_{(18)} = (-1.7221, 0.368, 0.17138)^\top$, $E_{(18)} = 0.50911$



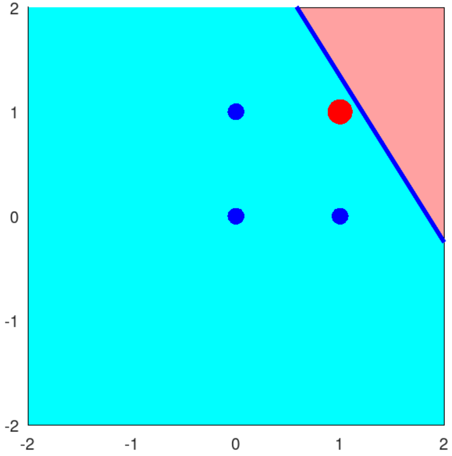
ex1 (AND) | CLog–MGE | $t = 19$

- $\tilde{w}_{(19)} = (-1.8096, 0.368, 0.083887)^\top, E_{(19)} = 0.52861$



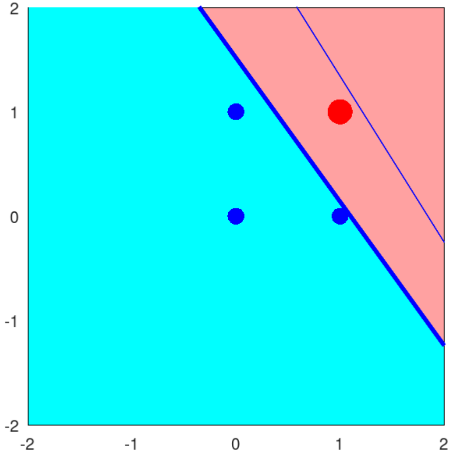
ex1 (AND) | CLog–MGE | $t = 20$

- $\tilde{w}_{(20)} = (-1.4119, 0.76569, 0.48158)^\top, E_{(20)} = 0.43769$



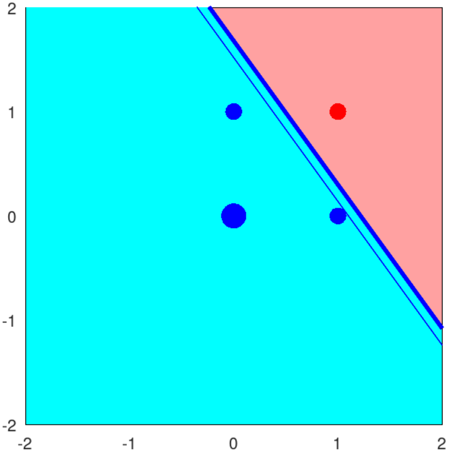
ex1 (AND) | CLog–MGE | $t = 21$

- $\tilde{w}_{(21)} = (-1.1413, 1.0362, 0.75211)^\top, E_{(21)} = 0.4644$

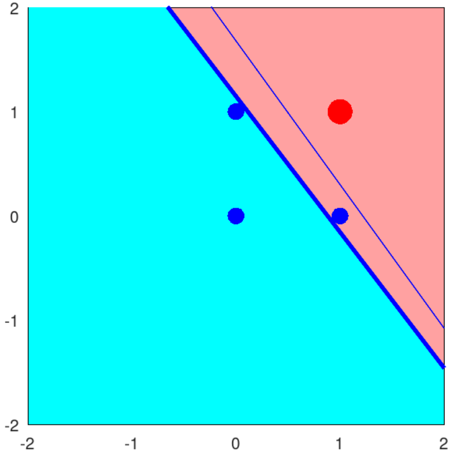


ex1 (AND) | CLog–MGE | $t = 22$

- $\tilde{w}_{(22)} = (-1.2624, 1.0362, 0.75211)^\top, E_{(22)} = 0.44255$

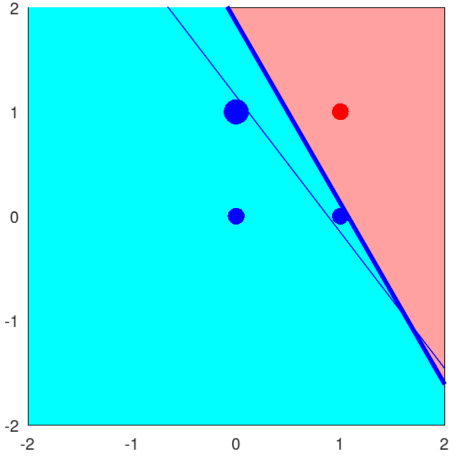


- $\tilde{w}_{(23)} = (-1.0766, 1.2219, 0.93784)^\top, E_{(23)} = 0.49485$

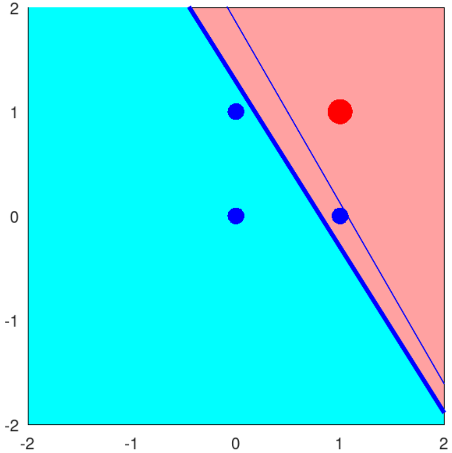


ex1 (AND) | CLog–MGE | $t = 24$

- $\tilde{w}_{(24)} = (-1.3093, 1.2219, 0.70516)^\top, E_{(24)} = 0.43917$

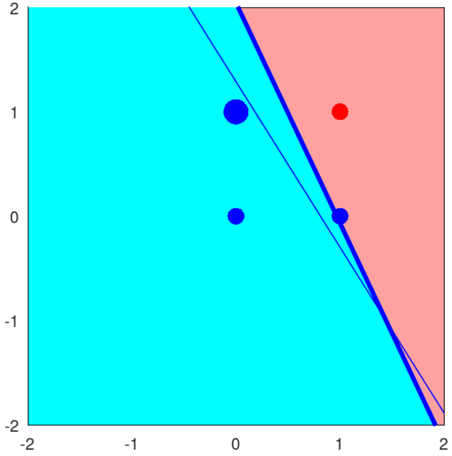


- $\tilde{w}_{(25)} = (-1.1342, 1.3971, 0.8803)^\top, E_{(25)} = 0.49077$



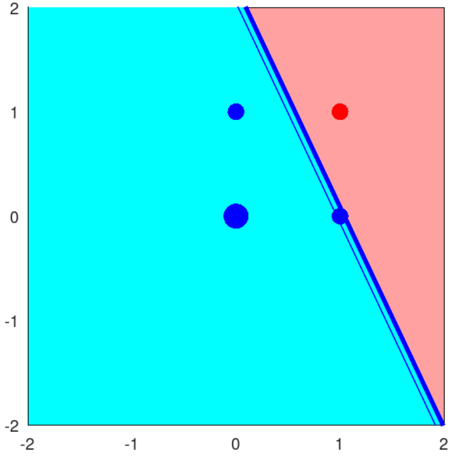
ex1 (AND) | CLog–MGE | $t = 26$

- $\tilde{w}_{(26)} = (-1.3526, 1.3971, 0.66187)^\top, E_{(26)} = 0.43824$



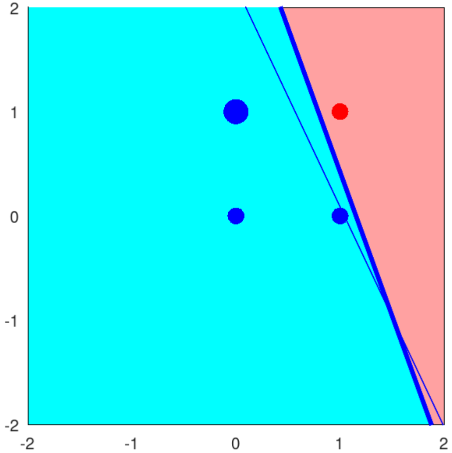
ex1 (AND) | CLog–MGE | $t = 27$

- $\tilde{w}_{(27)} = (-1.4553, 1.3971, 0.66187)^\top, E_{(27)} = 0.42087$



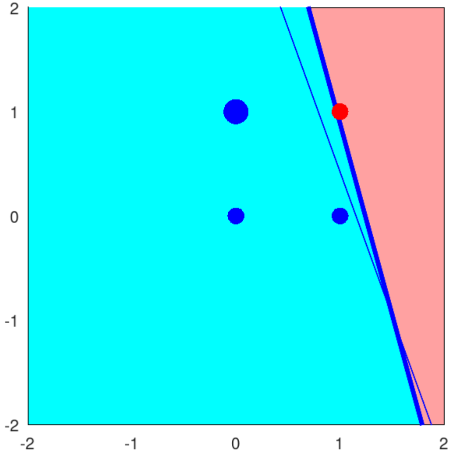
ex1 (AND) | CLog–MGE | $t = 28$

- $\tilde{w}_{(28)} = (-1.611, 1.3971, 0.50615)^\top$, $E_{(28)} = 0.40443$



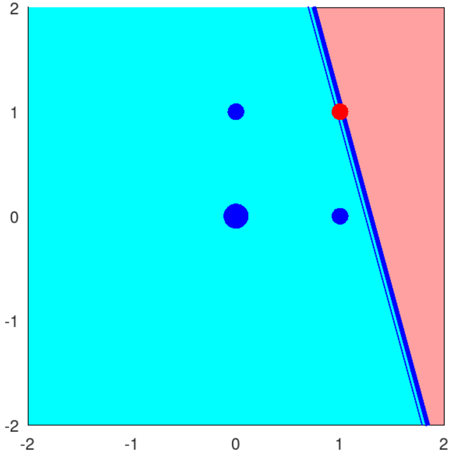
ex1 (AND) | CLog–MGE | $t = 29$

- $\tilde{w}_{(29)} = (-1.7355, 1.3971, 0.38174)^\top, E_{(29)} = 0.40051$



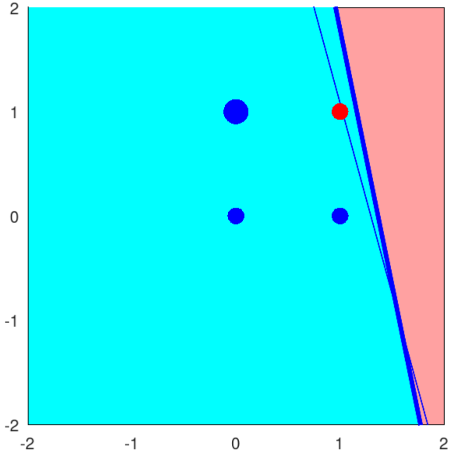
ex1 (AND) | CLog–MGE | $t = 30$

- $\tilde{w}_{(30)} = (-1.8104, 1.3971, 0.38174)^\top$, $E_{(30)} = 0.39577$



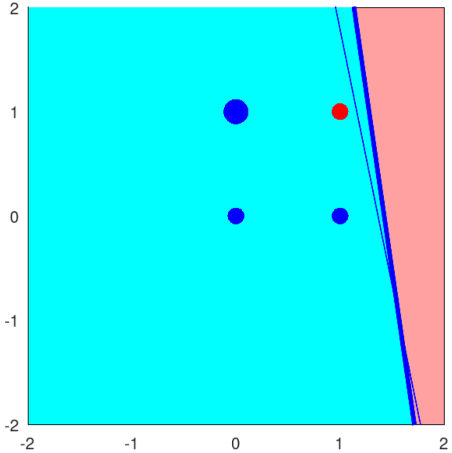
ex1 (AND) | CLog–MGE | $t = 31$

- $\tilde{w}_{(31)} = (-1.9071, 1.3971, 0.28509)^\top, E_{(31)} = 0.40023$

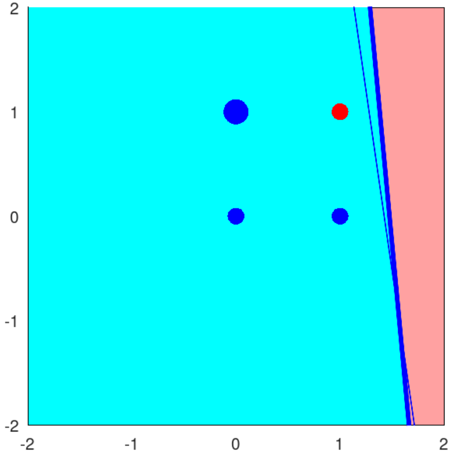


ex1 (AND) | CLog–MGE | $t = 32$

- $\tilde{w}_{(32)} = (-1.9895, 1.3971, 0.20262)^\top$, $E_{(32)} = 0.40753$

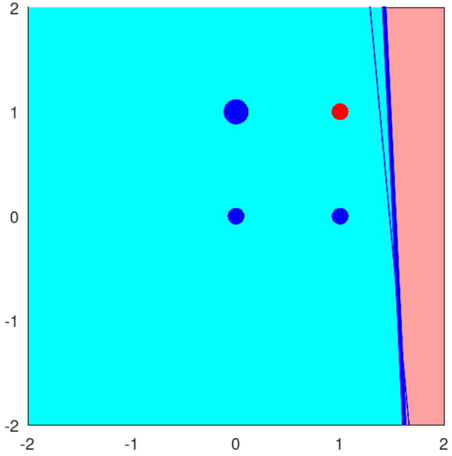


- $\tilde{w}_{(33)} = (-2.0613, 1.3971, 0.13089)^\top$, $E_{(33)} = 0.41636$



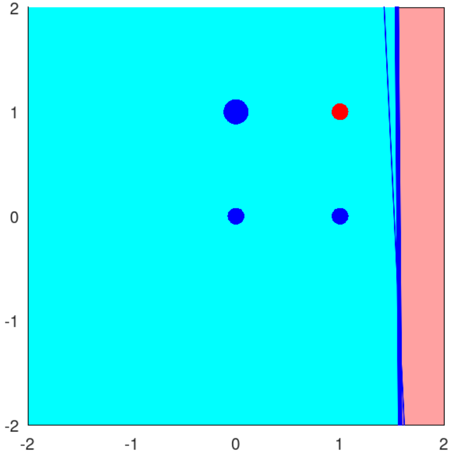
ex1 (AND) | CLog–MGE | $t = 34$

- $\tilde{w}_{(34)} = (-2.1246, 1.3971, 0.067539)^\top, E_{(34)} = 0.42598$



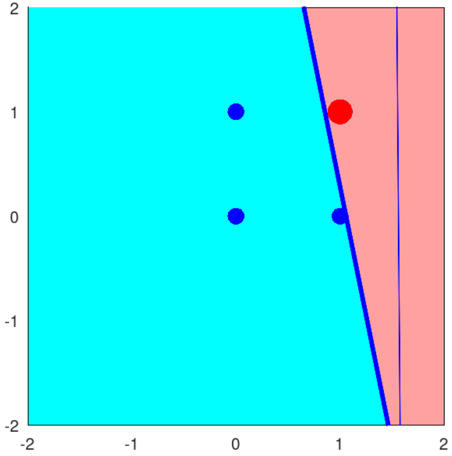
ex1 (AND) | CLog–MGE | $t = 35$

- $\tilde{w}_{(35)} = (-2.1813, 1.3971, 0.010869)^\top, E_{(35)} = 0.43596$



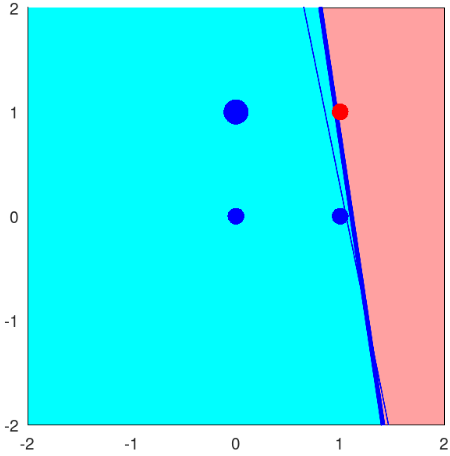
ex1 (AND) | CLog–MGE | $t = 36$

- $\tilde{w}_{(36)} = (-1.8392, 1.7392, 0.35299)^\top$, $E_{(36)} = 0.39262$

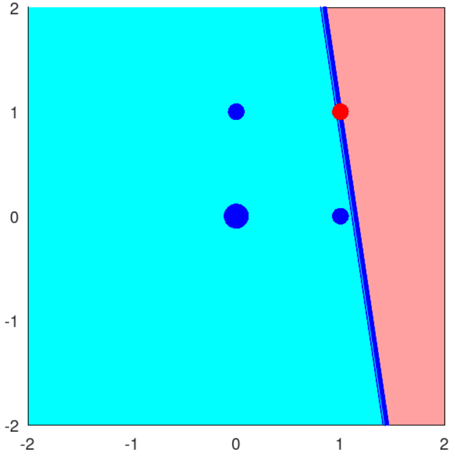


ex1 (AND) | CLog–MGE | $t = 37$

- $\tilde{w}_{(37)} = (-1.9314, 1.7392, 0.26074)^\top$, $E_{(37)} = 0.39221$

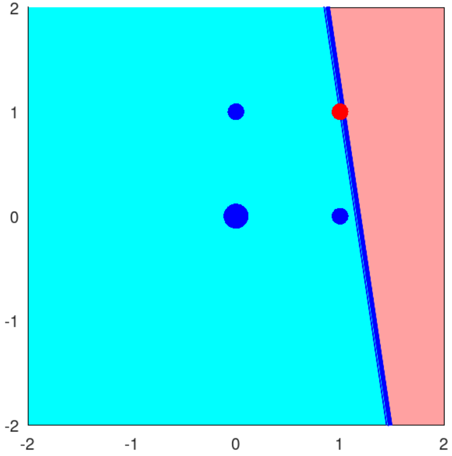


- $\tilde{w}_{(38)} = (-1.9947, 1.7392, 0.26074)^\top$, $E_{(38)} = 0.38856$

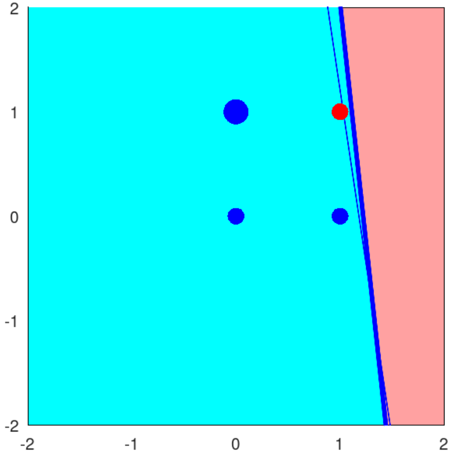


ex1 (AND) | CLog–MGE | $t = 39$

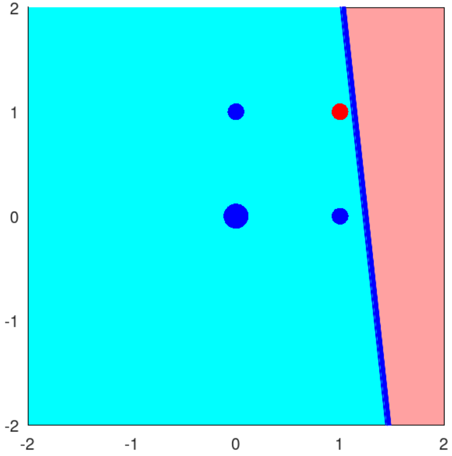
- $\tilde{w}_{(39)} = (-2.0546, 1.7392, 0.26074)^\top, E_{(39)} = 0.38578$



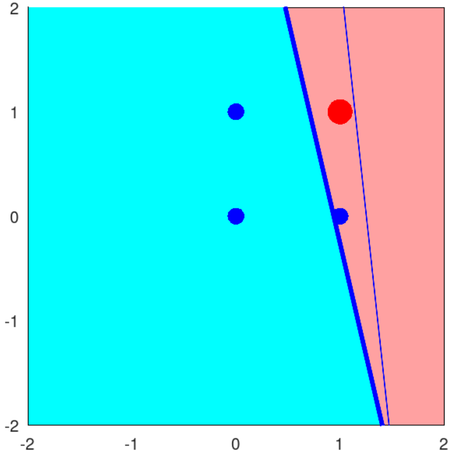
- $\tilde{w}_{(40)} = (-2.1259, 1.7392, 0.18944)^\top$, $E_{(40)} = 0.39061$



- $\tilde{w}_{(41)} = (-2.1792, 1.7392, 0.18944)^\top$, $E_{(41)} = 0.38969$

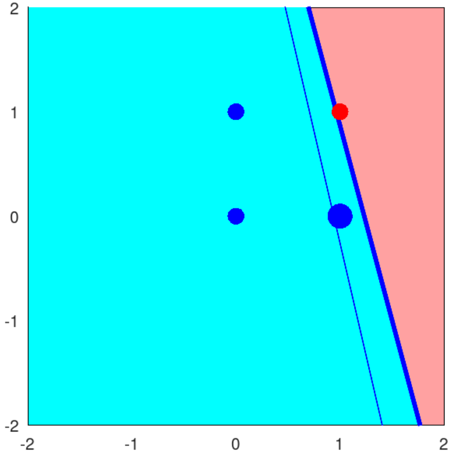


- $\tilde{w}_{(42)} = (-1.898, 2.0204, 0.47059)^\top$, $E_{(42)} = 0.38772$

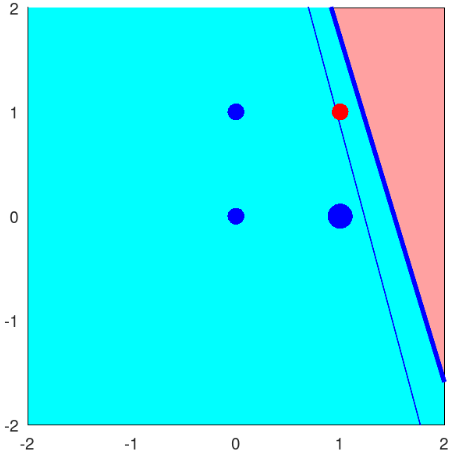


ex1 (AND) | CLog–MGE | $t = 43$

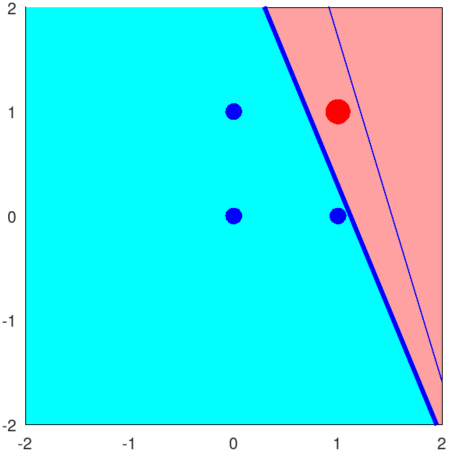
- $\tilde{w}_{(43)} = (-2.1633, 1.7551, 0.47059)^\top$, $E_{(43)} = 0.36247$



- $\tilde{w}_{(44)} = (-2.363, 1.5554, 0.47059)^\top, E_{(44)} = 0.36872$

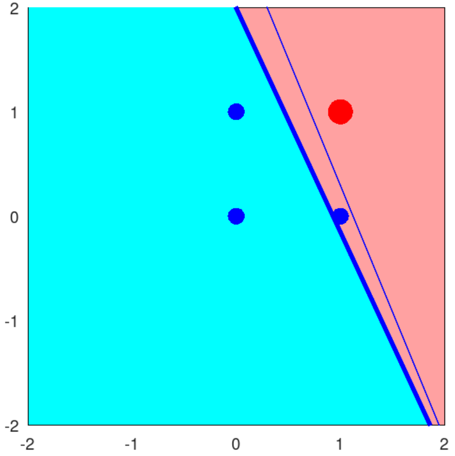


- $\tilde{w}_{(45)} = (-2.0712, 1.8471, 0.76232)^\top, E_{(45)} = 0.35124$



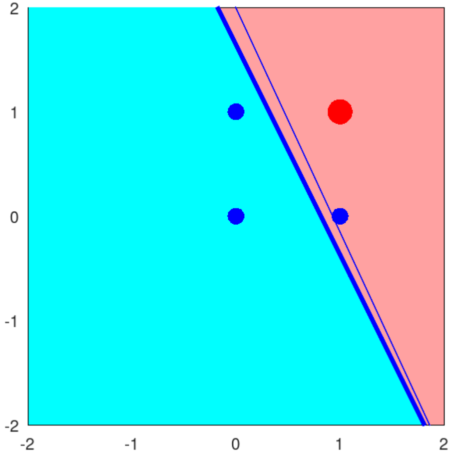
ex1 (AND) | CLog–MGE | $t = 46$

- $\tilde{w}_{(46)} = (-1.8869, 2.0314, 0.94662)^\top, E_{(46)} = 0.38208$

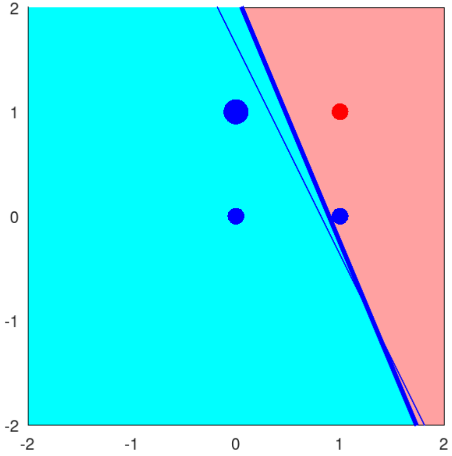


ex1 (AND) | CLog–MGE | $t = 47$

- $\tilde{w}_{(47)} = (-1.7612, 2.1572, 1.0723)^\top, E_{(47)} = 0.42082$

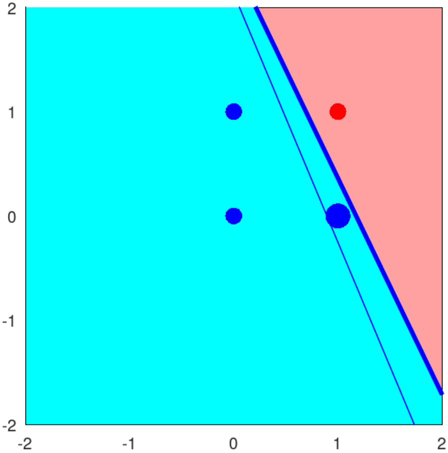


- $\tilde{w}_{(48)} = (-1.9284, 2.1572, 0.90519)^\top, E_{(48)} = 0.38396$



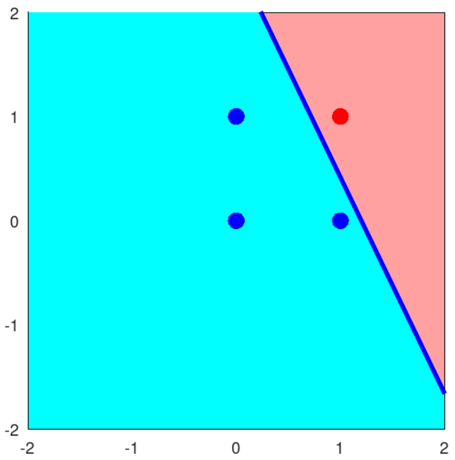
ex1 (AND) | CLog–MGE | $t = 49$

- $\tilde{w}_{(49)} = (-2.2069, 1.8787, 0.90519)^\top, E_{(49)} = 0.3333$

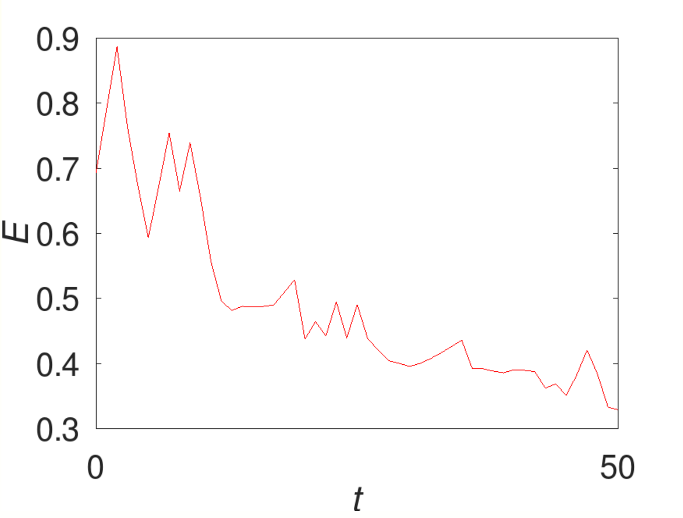


ex1 (AND) | CLog–MGE | $t = 50$

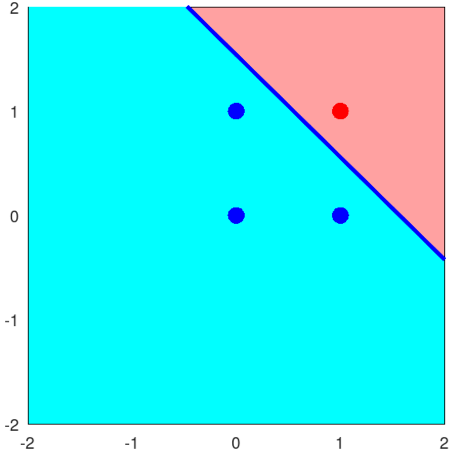
- $\tilde{w}_{(0)} = (0, 0, 0)^\top, E_{(0)} = 0.69315$
- $\tilde{w}_{(50)} = (-2.2564, 1.8787, 0.90519)^\top, E_{(50)} = 0.32892$



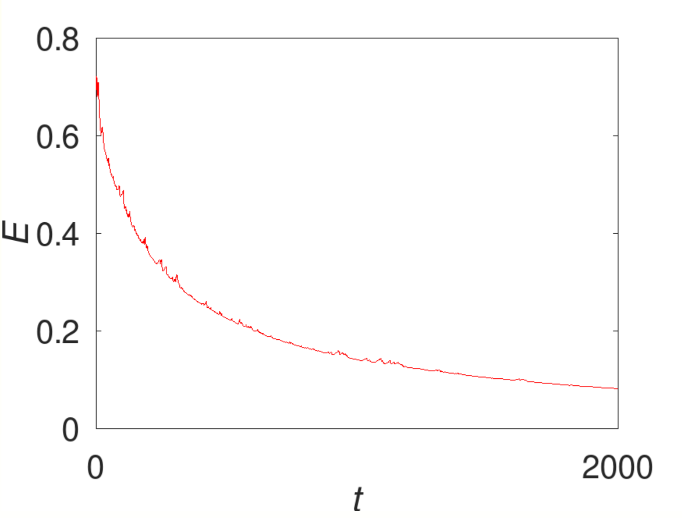
ex1 (AND) | CLog–MGE | E



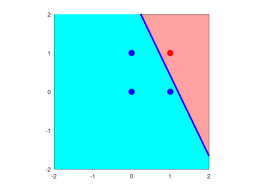
- $\tilde{w}_{(0)} = (0,0,0)^\top, E_{(0)} = 0.69315$
- $\tilde{w}_{(2000)} = (-6.5701, 4.1876, 4.2653)^\top, E_{(2000)} = 0.081614$

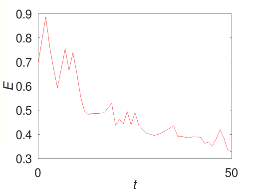


ex1 (AND) | CLog–MGE | E



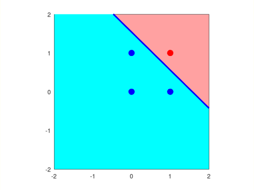
$t_{\mathrm{MAX}} = 50,\ \eta = 0.5$
$E = 0.3289,\ acc = 1$

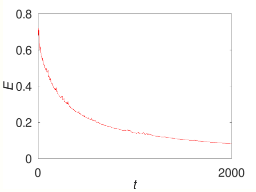




y^n	0	0	0	1
$\hat{p}^n$	0.0948	0.2057	0.4067	0.6289
$\hat{y}^n$	0	0	0	1

$t_{\mathrm{MAX}} = 2000,\ \eta = 0.1$
$E = 0.0816,\ acc = 1$

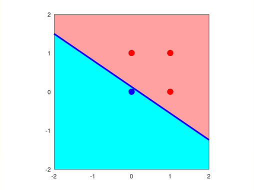


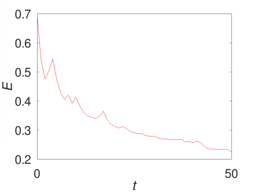


y^n	0	0	0	1
$\hat{p}^n$	0.0014	0.0907	0.0845	0.8679
$\hat{y}^n$	0	0	0	1

ex2 (OR) | CLog–MGE

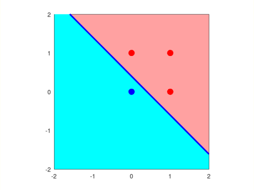
$t_{\mathrm{MAX}} = 50,\ \eta = 0.5$
$E = 0.2248,\ acc = 1$

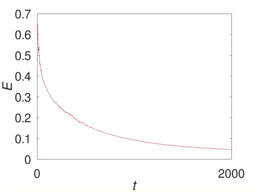




y^n	0	1	1	1
$\hat{p}^n$	0.4223	0.8981	0.8006	0.9798
$\hat{y}^n$	0	1	1	1

$t_{\mathrm{MAX}} = 2000,\ \eta = 0.1$
$E = 0.0471,\ acc = 1$

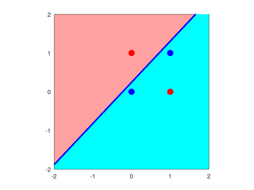


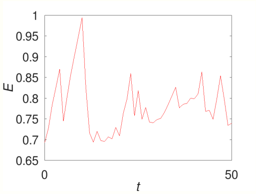


y^n	0	1	1	1
$\hat{p}^n$	0.1048	0.9614	0.9624	0.9998
$\hat{y}^n$	0	1	1	1

ex3 (XOR) | CLog–MGE

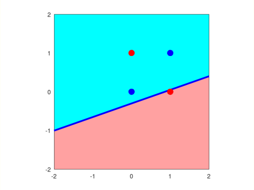
$t_{\mathrm{MAX}} = 50,\ \eta = 0.5$
$E = 0.7398,\ acc = 0.75$

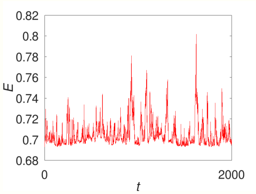




y^n	0	1	1	0
$\hat{p}^n$	0.4518	0.6466	0.2610	0.4395
$\hat{y}^n$	0	1	0	0

$t_{\mathrm{MAX}} = 2000,\ \eta = 0.1$
$E = 0.6959,\ acc = 0.75$

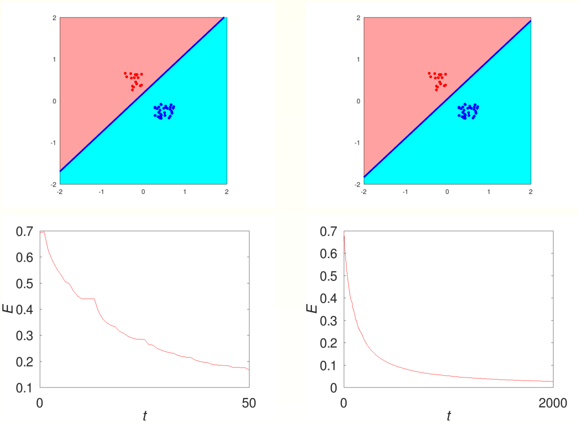


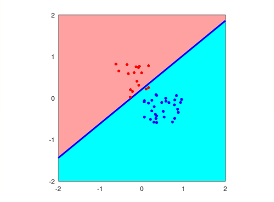
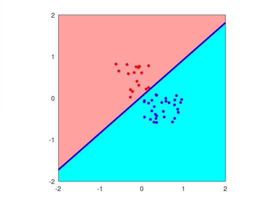
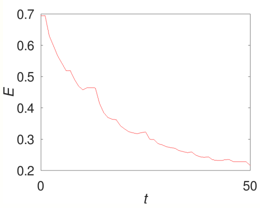
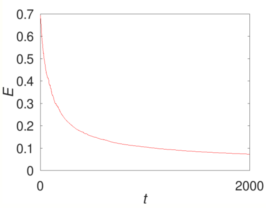


y^n	0	1	1	0
$\hat{p}^n$	0.4863	0.4412	0.5022	0.4569
$\hat{y}^n$	0	0	1	0

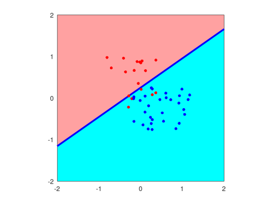
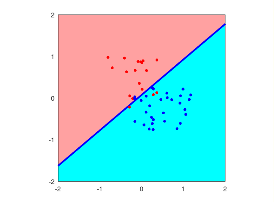
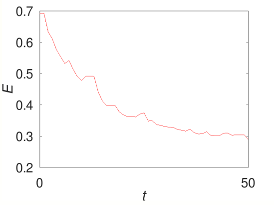
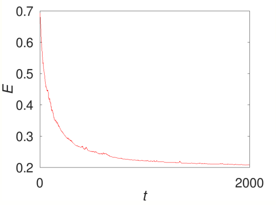
ex4 | CLog–MGE

$t_{\text{MAX}} = 50,\ \eta = 0.5$	$t_{\text{MAX}} = 2000,\ \eta = 0.1$
$E = 0.1670,\ acc = 1$	$E = 0.0272,\ acc = 1$



$t_{\text{MAX}} = 50,\ \eta = 0.5$ $E = 0.2159,\ acc = 0.96$	$t_{\text{MAX}} = 2000,\ \eta = 0.1$ $E = 0.0730,\ acc = 1$
	
	

ex6 | CLog–MGE

$t_{\mathrm{MAX}} = 50,\ \eta = 0.5$ $E = 0.2891,\ acc = 0.90$	$t_{\mathrm{MAX}} = 2000,\ \eta = 0.1$ $E = 0.2083,\ acc = 0.90$
	
	

CLog | versão dual (i)

- Base de dados $D = (x^n, y^n)_{n=1}^N$, $x^n \in \mathbb{R}^I$, $y^n \in \{0, 1\}$.
- Espaço das hipóteses: $\mathcal{H} = \mathbb{R}^N$.
- Relação entre o vetor das variáveis primais $\tilde{w} \in \mathbb{R}^{I+1}$ e o vetor das variáveis duais $\alpha \in \mathbb{R}^N$

$$\tilde{w} = \sum_{n=1}^{N} \alpha_n \tilde{x}^n.$$

- Arquitetura da *Machine Learning*

$$\hat{p} \equiv h(x; \alpha) = \sigma\left(\sum_{n=1}^{N} \alpha_n (\tilde{x}^n \cdot \tilde{x})\right), \alpha \in \mathcal{H}.$$

- Função custo

$$E(\alpha; D) = \frac{1}{N} \sum_{n=1}^{N} E_n(\alpha; x^n, y^n),$$

$$E_n(\alpha; x^n, y^n) = -y^n \ln(\hat{p}^n) - (1 - y^n) \ln(1 - \hat{p}^n).$$

CLog | versão dual (ii)

- Para se aplicar o Método do Gradiente ao CLog tem que se calcular o gradiente da função custo E, podendo-se mostrar (exercício) que

$$\nabla E_n(\alpha; x^n, y^n) = (\hat{p}^n - y^n) \begin{bmatrix} \tilde{x}^1 \cdot \tilde{x}^n \\ \vdots \\ \tilde{x}^N \cdot \tilde{x}^n \end{bmatrix},$$

$$\nabla E(\alpha; D) = \frac{1}{N} \sum_{n=1}^{N} \nabla E_n(\alpha; x^n, y^n) = \frac{1}{N} \sum_{n=1}^{N} (\hat{p}^n - y^n) \begin{bmatrix} \tilde{x}^1 \cdot \tilde{x}^n \\ \vdots \\ \tilde{x}^N \cdot \tilde{x}^n \end{bmatrix}.$$

- Vamos denotar por **CLogD–MGE** o algoritmo do classificador logístico na versão dual com o MGE.

Algoritmo CLogD–MGB

Input: $D = (x^n, y^n)_{n=1}^N$, $x^n \in \mathbb{R}^I$, $y^n \in \{0,1\}$, $\eta \in \mathbb{R}^+$, $d \in \mathbb{N}$, CP
Output: $\alpha^* \in \mathbb{R}^N$

$t \leftarrow 0$;

$\alpha_{(0)} = (0, \ldots, 0)^\top \in \mathbb{R}^N$;

while V **do**

$$\hat{p}^n \leftarrow \sigma\left(\sum_{\ell=1}^N \alpha_{(t),\ell}(\tilde{x}^\ell \cdot \tilde{x}^n)\right),\ n = 1, \ldots, N;$$

$$s_{(t)} \leftarrow \frac{1}{N}\sum_{n=1}^N (\hat{p}^n - y^n)\begin{bmatrix} \tilde{x}^1 \cdot \tilde{x}^n \\ \vdots \\ \tilde{x}^N \cdot \tilde{x}^n \end{bmatrix};$$

$\alpha_{(t+1)} \leftarrow \alpha_{(t)} - \eta s_{(t)}$;

if CP=V **then**

$\alpha^* \leftarrow \alpha_{(t+1)}$; **return** α^*;

else

$t \leftarrow t + 1$;

Input: $D = (x^n, y^n)_{n=1}^N$, $x^n \in \mathbb{R}^I$, $y^n \in \{0,1\}$, $\eta \in \mathbb{R}^+$, $d \in \mathbb{N}$, CP
Output: $\alpha^* \in \mathbb{R}^N$

$t \leftarrow 0$;

$\alpha_{(0)} = (0, \ldots, 0)^\top \in \mathbb{R}^N$;

while V **do**

  selecionar $n \in \{1, \ldots, N\}$ aleatório;

  $$\hat{p}^n \leftarrow \sigma\left(\sum_{\ell=1}^N \alpha_{(t),\ell}(\tilde{x}^\ell \cdot \tilde{x}^n)\right);$$

  $$s_{(t)} \leftarrow (\hat{p}^n - y^n)\begin{bmatrix} \tilde{x}^1 \cdot \tilde{x}^n \\ \vdots \\ \tilde{x}^N \cdot \tilde{x}^n \end{bmatrix};$$

  $\alpha_{(t+1)} \leftarrow \alpha_{(t)} - \eta s_{(t)}$;

  if CP=V **then**

    $\alpha^* \leftarrow \alpha_{(t+1)}$; **return** α^*;

  else

    $t \leftarrow t + 1$;

CLog | versão dual com *kernel* polinomial

- Arquitetura da *Machine Learning*

$$\hat{p} \equiv h(x;\alpha) = \sigma\left(\sum_{n=1}^{N} \alpha_n K(\tilde{x}^n, \tilde{x})\right), \alpha \in \mathcal{H}.$$

- Para se aplicar o Método do Gradiente ao CLog com *kernel* polinomial de grau d

$$K(\tilde{x}^n, \tilde{x}) = (\tilde{x}^n \cdot \tilde{x})^d,$$

tem que se calcular o gradiente da função custo E, podendo-se mostrar (exercício) que

$$\nabla E(\alpha; D) = \frac{1}{N}\sum_{n=1}^{N} \nabla E_n(\alpha; x^n, y^n) = \frac{1}{N}\sum_{n=1}^{N} (\hat{p}^n - y^n)\begin{bmatrix} (\tilde{x}^1 \cdot \tilde{x}^n)^d \\ \vdots \\ (\tilde{x}^N \cdot \tilde{x}^n)^d \end{bmatrix}.$$

- Vamos denotar por **CLogDKPd–MGE** o algoritmo do classificador logístico na versão dual com *kernel* polinomial de grau d com o MGE e por **CLogDKPd–MGE** o algoritmo do classificador logístico na versão dual com *kernel* polinomial de grau d com o MGE.

Algoritmo CLogDKPd–MGB

Input: $D = (x^n, y^n)_{n=1}^N$, $x^n \in \mathbb{R}^I$, $y^n \in \{0,1\}$, $\eta \in \mathbb{R}^+$, $d \in \mathbb{N}$, CP
Output: $\alpha^* \in \mathbb{R}^N$

$t \leftarrow 0$;

$\alpha_{(0)} = (0, \ldots, 0)^\top \in \mathbb{R}^N$;

while V **do**

$$\hat{p}^n \leftarrow \sigma\left(\sum_{\ell=1}^N \alpha_{(t),\ell}(\tilde{x}^\ell \cdot \tilde{x}^n)\right),\ n = 1, \ldots, N;$$

$$s_{(t)} \leftarrow \frac{1}{N}\sum_{n=1}^N (\hat{p}^n - y^n)\begin{bmatrix}(\tilde{x}^1 \cdot \tilde{x}^n)^d \\ \vdots \\ (\tilde{x}^N \cdot \tilde{x}^n)^d\end{bmatrix};$$

$\alpha_{(t+1)} \leftarrow \alpha_{(t)} - \eta s_{(t)}$;

if CP=V **then**

$\alpha^* \leftarrow \alpha_{(t+1)}$; **return** α^*;

else

$t \leftarrow t + 1$;

Algoritmo CLogDKPd–MGE

Input: $D = (x^n, y^n)_{n=1}^N$, $x^n \in \mathbb{R}^I$, $y^n \in \{0,1\}$, $\eta \in \mathbb{R}^+$, $d \in \mathbb{N}$, CP
Output: $\alpha^* \in \mathbb{R}^N$

$t \leftarrow 0$;

$\alpha_{(0)} = (0, \ldots, 0)^\top \in \mathbb{R}^N$;

while V **do**

  selecionar $n \in \{1, \ldots, N\}$ aleatório;

  $$\hat{p}^n \leftarrow \sigma\left(\sum_{\ell=1}^N \alpha_{(t),\ell}(\tilde{x}^\ell \cdot \tilde{x}^n)\right);$$

  $$s_{(t)} \leftarrow (\hat{p}^n - y^n)\begin{bmatrix}(\tilde{x}^1 \cdot \tilde{x}^n)^d \\ \vdots \\ (\tilde{x}^N \cdot \tilde{x}^n)^d\end{bmatrix};$$

  $\alpha_{(t+1)} \leftarrow \alpha_{(t)} - \eta s_{(t)}$;

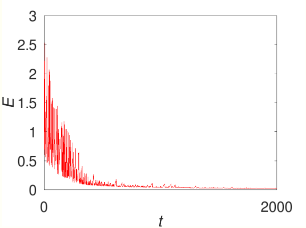
  if $CP = V$ **then**

    $\alpha^* \leftarrow \alpha_{(t+1)}$; **return** α^*;

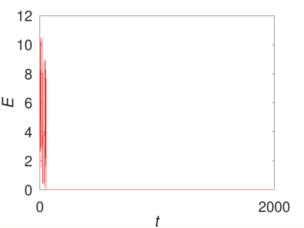
  else

    $t \leftarrow t + 1$;

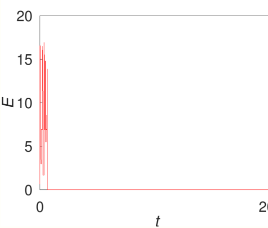
CLog–MGE, $\eta = 0.5$
$E = 0.0226,\ acc = 1$



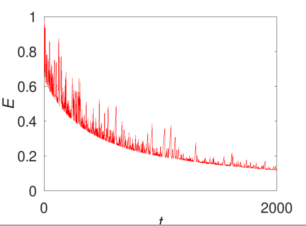
CLogDKP2–MGE, $\eta = 0.5$
$E = 0.0016,\ acc = 1$



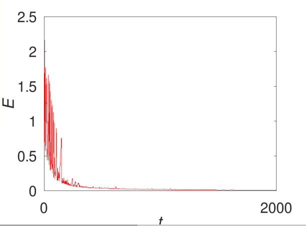
CLogDKP3–MGE, $\eta =$
$E = 0.0000,\ acc = 1$



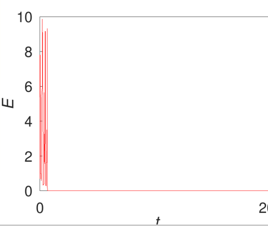
CLog–MGE, $\eta = 0.1$
$E = 0.1186,\ acc = 1$



CLogDKP2–MGE, $\eta = 0.1$
$E = 0.0102,\ acc = 1$

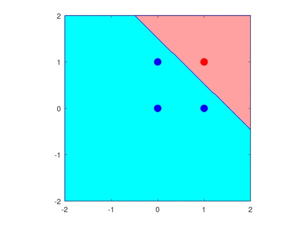


CLogDKP3–MGE, $\eta =$
$E = 0.0020,\ acc = 1$

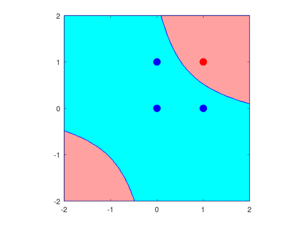


ex1 (AND) | CLog–MGE

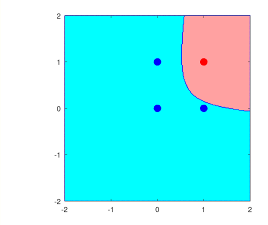
CLog–MGE, $\eta = 0.5$
$E = 0.0226,\ acc = 1$



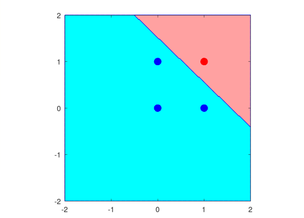
CLogDKP2–MGE, $\eta = 0.5$
$E = 0.0016,\ acc = 1$



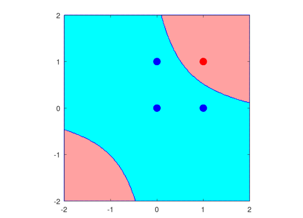
CLogDKP3–MGE, $\eta =$
$E = 0.0000,\ acc = 1$



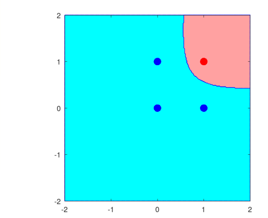
CLog–MGE, $\eta = 0.1$
$E = 0.1186,\ acc = 1$



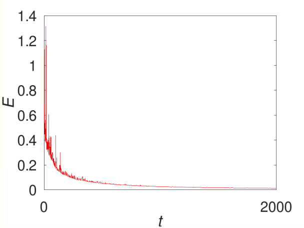
CLogDKP2–MGE, $\eta = 0.1$
$E = 0.0102,\ acc = 1$



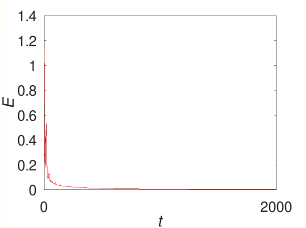
CLogDKP3–MGE, $\eta =$
$E = 0.0020,\ acc = 1$



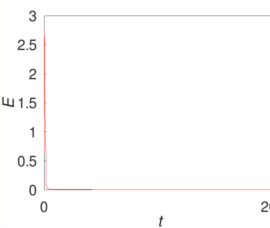
CLog–MGE, $\eta = 0.5$
$E = 0.0128,\ acc = 1$



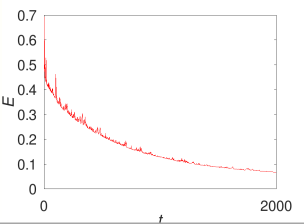
CLogDKP2–MGE, $\eta = 0.5$
$E = 0.0030,\ acc = 1$



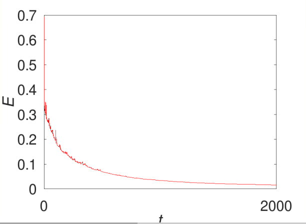
CLogDKP3–MGE, $\eta =$
$E = 0.0013,\ acc = 1$



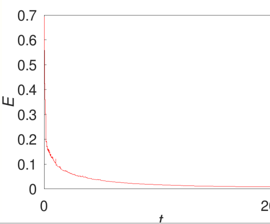
CLog–MGE, $\eta = 0.1$
$E = 0.0674,\ acc = 1$



CLogDKP2–MGE, $\eta = 0.1$
$E = 0.0159,\ acc = 1$

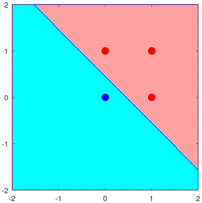


CLogDKP3–MGE, $\eta =$
$E = 0.0078,\ acc = 1$

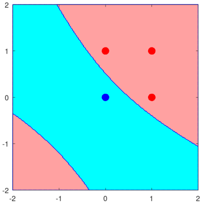


ex2 (OR) | CLog–MGE

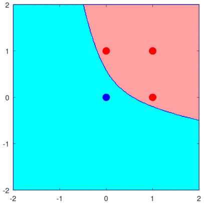
CLog–MGE, $\eta = 0.5$
$E = 0.0128,\ acc = 1$



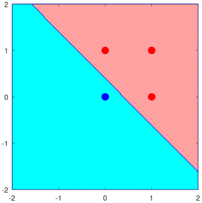
CLogDKP2–MGE, $\eta = 0.5$
$E = 0.0030,\ acc = 1$



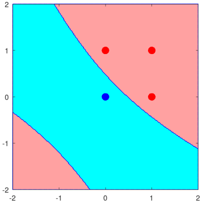
CLogDKP3–MGE, $\eta =$
$E = 0.0013,\ acc = 1$



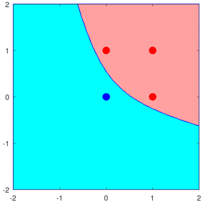
CLog–MGE, $\eta = 0.1$
$E = 0.0674,\ acc = 1$



CLogDKP2–MGE, $\eta = 0.1$
$E = 0.0159,\ acc = 1$

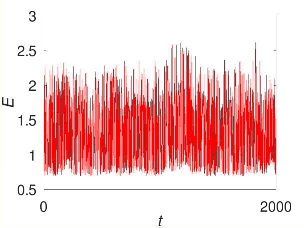


CLogDKP3–MGE, $\eta =$
$E = 0.0078,\ acc = 1$

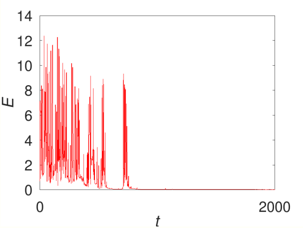


ex3 (XOR) | CLog–MGE

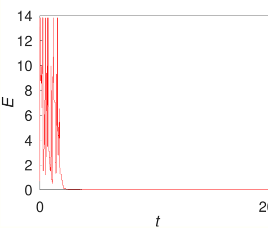
CLog–MGE, $\eta = 0.5$
$E = 1.2297,\ acc = 0.50$



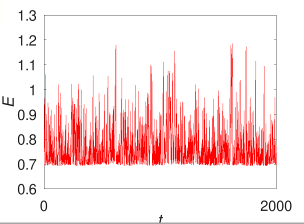
CLogDKP2–MGE, $\eta = 0.5$
$E = 0.0187,\ acc = 1$



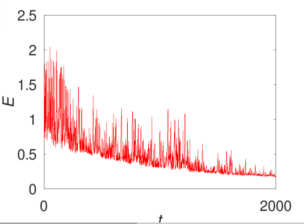
CLogDKP3–MGE, $\eta =$
$E = 0.0016,\ acc = 1$



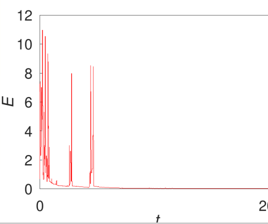
CLog–MGE, $\eta = 0.1$
$E = 0.6958,\ acc = 0.75$



CLogDKP2–MGE, $\eta = 0.1$
$E = 0.1714,\ acc = 1$

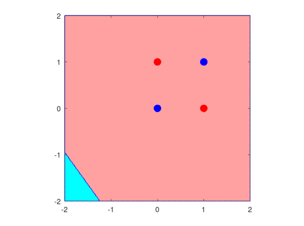


CLogDKP3–MGE, $\eta =$
$E = 0.0207,\ acc = 1$

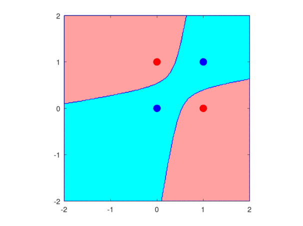


ex3 (XOR) | CLog–MGE

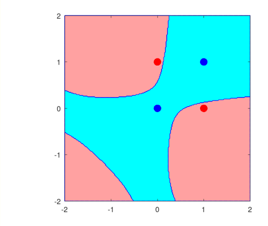
CLog–MGE, $\eta = 0.5$
$E = 1.2297,\ acc = 0.50$



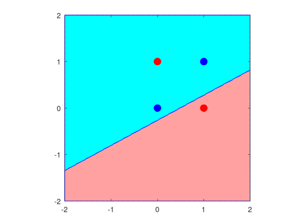
CLogDKP2–MGE, $\eta = 0.5$
$E = 0.0187,\ acc = 1$



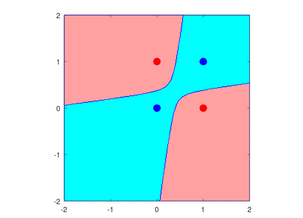
CLogDKP3–MGE, $\eta =$
$E = 0.0016,\ acc = 1$



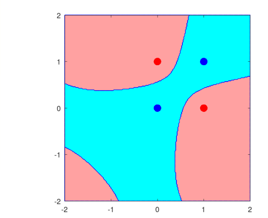
CLog–MGE, $\eta = 0.1$
$E = 0.6958,\ acc = 0.75$



CLogDKP2–MGE, $\eta = 0.1$
$E = 0.1714,\ acc = 1$

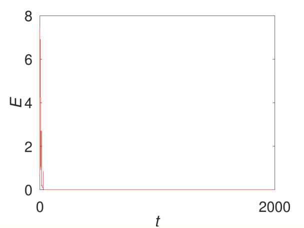


CLogDKP3–MGE, $\eta =$
$E = 0.0207,\ acc = 1$

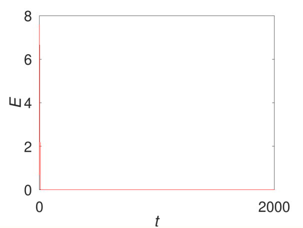


ex4 | CLog–MGE

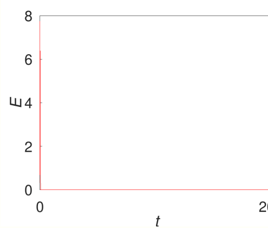
CLog–MGE, $\eta = 0.5$
$E = 0.0000,\ acc = 1$



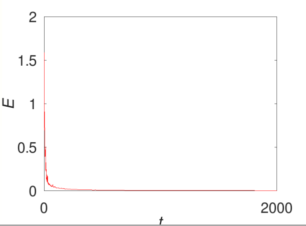
CLogDKP2–MGE, $\eta = 0.5$
$E = 0.0000,\ acc = 1$



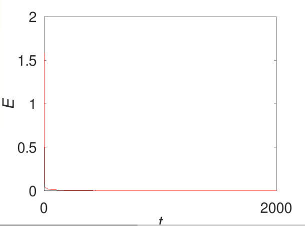
CLogDKP3–MGE, $\eta =$
$E = 0.0000,\ acc = 1$



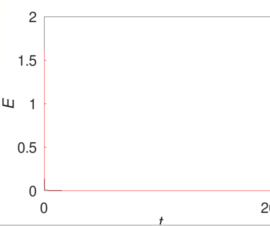
CLog–MGE, $\eta = 0.1$
$E = 0.0026,\ acc = 1$



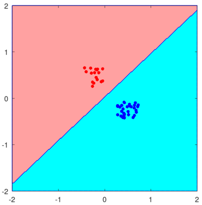
CLogDKP2–MGE, $\eta = 0.1$
$E = 0.0007,\ acc = 1$



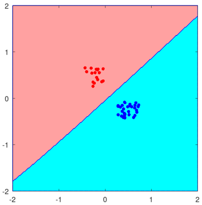
CLogDKP3–MGE, $\eta =$
$E = 0.0003,\ acc = 1$



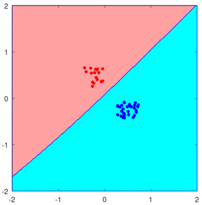
CLog–MGE, $\eta = 0.5$
$E = 0.0000,\ acc = 1$



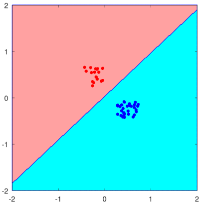
CLogDKP2–MGE, $\eta = 0.5$
$E = 0.0000,\ acc = 1$



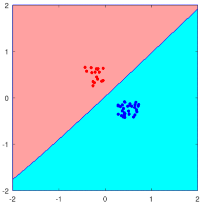
CLogDKP3–MGE, $\eta =$
$E = 0.0000,\ acc = 1$



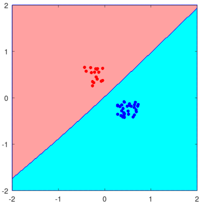
CLog–MGE, $\eta = 0.1$
$E = 0.0026,\ acc = 1$



CLogDKP2–MGE, $\eta = 0.1$
$E = 0.0007,\ acc = 1$

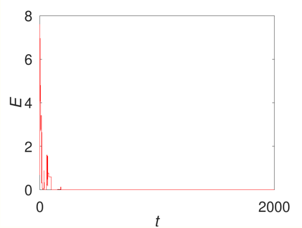


CLogDKP3–MGE, $\eta =$
$E = 0.0003,\ acc = 1$

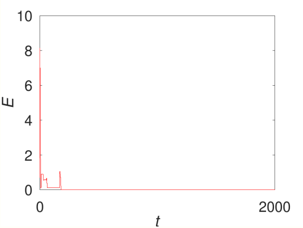


ex5 | CLog–MGE

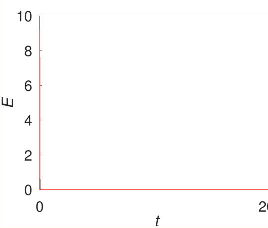
CLog–MGE, $\eta = 0.5$
$E = 0.0018, \ acc = 1$



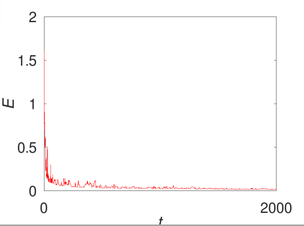
CLogDKP2–MGE, $\eta = 0.5$
$E = 0.0001, \ acc = 1$



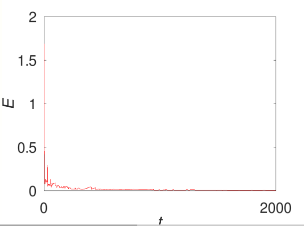
CLogDKP3–MGE, $\eta =$
$E = 0.0002, \ acc = 1$



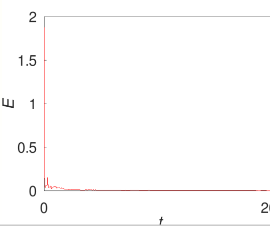
CLog–MGE, $\eta = 0.1$
$E = 0.0166, \ acc = 1$



CLogDKP2–MGE, $\eta = 0.1$
$E = 0.0054, \ acc = 1$

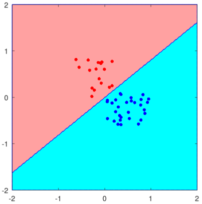


CLogDKP3–MGE, $\eta =$
$E = 0.0026, \ acc = 1$

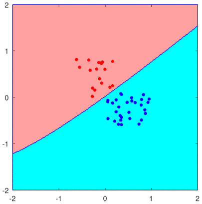


ex5 | CLog–MGE

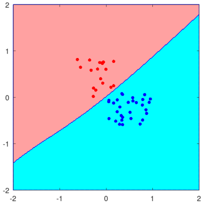
CLog–MGE, $\eta = 0.5$
$E = 0.0018,\ acc = 1$



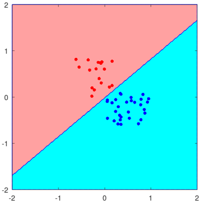
CLogDKP2–MGE, $\eta = 0.5$
$E = 0.0001,\ acc = 1$



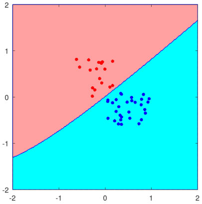
CLogDKP3–MGE, $\eta =$
$E = 0.0002,\ acc = 1$



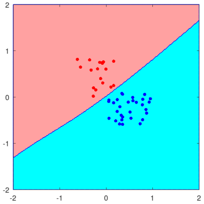
CLog–MGE, $\eta = 0.1$
$E = 0.0166,\ acc = 1$



CLogDKP2–MGE, $\eta = 0.1$
$E = 0.0054,\ acc = 1$

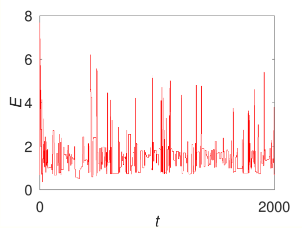


CLogDKP3–MGE, $\eta =$
$E = 0.0026,\ acc = 1$

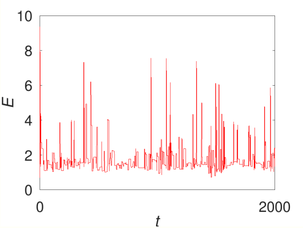


ex6 | CLog–MGE

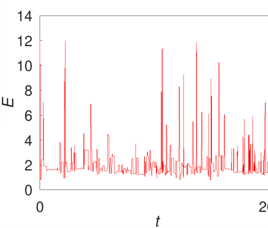
CLog–MGE, $\eta = 0.5$
$E = 0.7325,\ acc = 0.90$



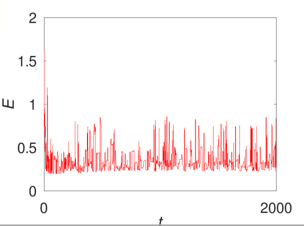
CLogDKP2–MGE, $\eta = 0.5$
$E = 1.3338,\ acc = 0.92$



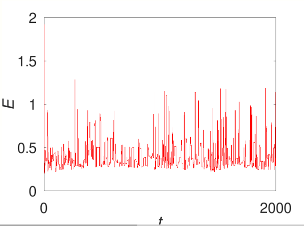
CLogDKP3–MGE, $\eta =$
$E = 1.3176,\ acc = 0.9$



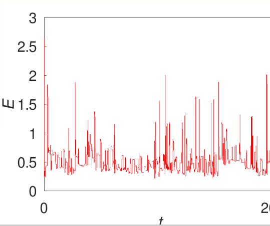
CLog–MGE, $\eta = 0.1$
$E = 0.3103,\ acc = 0.90$



CLogDKP2–MGE, $\eta = 0.1$
$E = 0.3833,\ acc = 0.90$

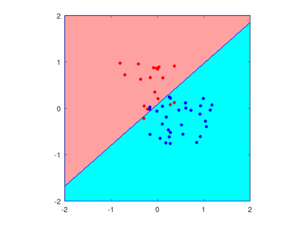


CLogDKP3–MGE, $\eta =$
$E = 0.5252,\ acc = 0.9$

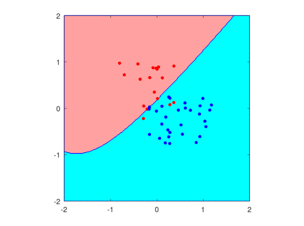


ex6 | CLog–MGE

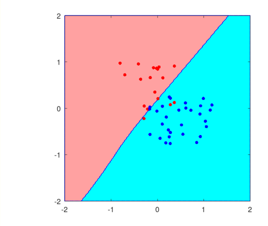
CLog–MGE, $\eta = 0.5$
$E = 0.7325,\ acc = 0.90$



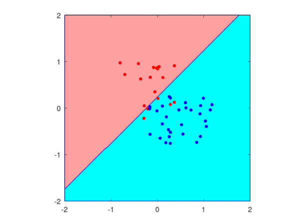
CLogDKP2–MGE, $\eta = 0.5$
$E = 1.3338,\ acc = 0.92$



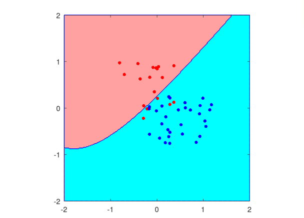
CLogDKP3–MGE, $\eta =$
$E = 1.3176,\ acc = 0.9$



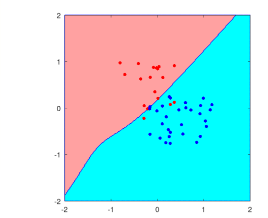
CLog–MGE, $\eta = 0.1$
$E = 0.3103,\ acc = 0.90$



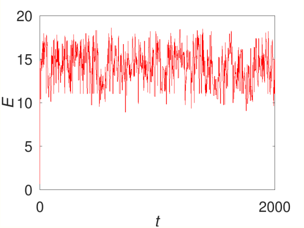
CLogDKP2–MGE, $\eta = 0.1$
$E = 0.3833,\ acc = 0.90$



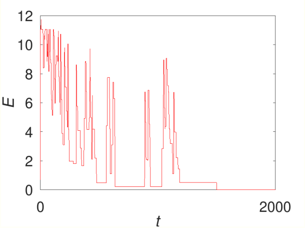
CLogDKP3–MGE, $\eta =$
$E = 0.5252,\ acc = 0.9$



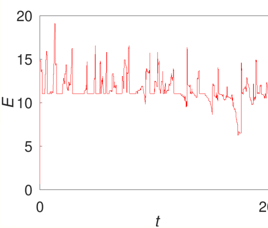
CLog–MGE, $\eta = 0.5$
$E = 11.0524,\ acc = 0.60$



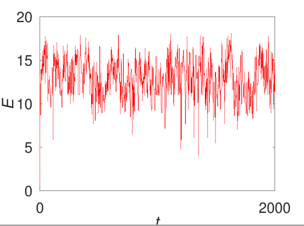
CLogDKP2–MGE, $\eta = 0.5$
$E = 0.0000,\ acc = 1$



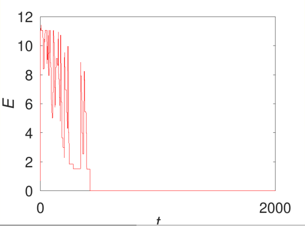
CLogDKP3–MGE, $\eta =$
$E = 10.7208,\ acc = 0.6$



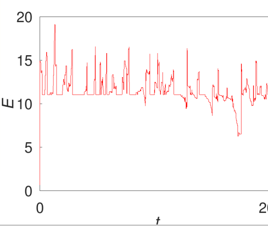
CLog–MGE, $\eta = 0.1$
$E = 11.1456,\ acc = 0.58$



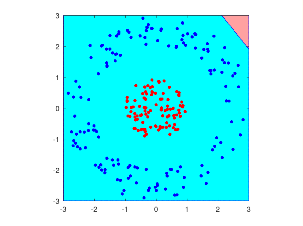
CLogDKP2–MGE, $\eta = 0.1$
$E = 0.0000,\ acc = 1$



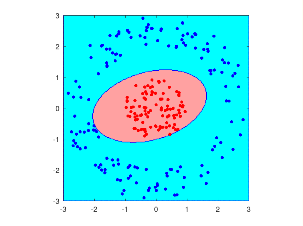
CLogDKP3–MGE, $\eta =$
$E = 10.7208,\ acc = 0.6$



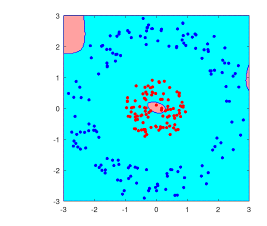
CLog–MGE, $\eta = 0.5$
$E = 11.0524,\ acc = 0.60$



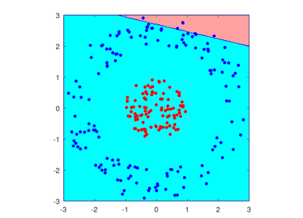
CLogDKP2–MGE, $\eta = 0.5$
$E = 0.0000,\ acc = 1$



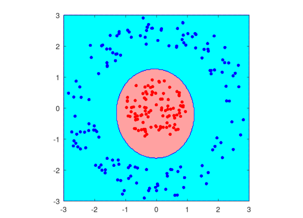
CLogDKP3–MGE, $\eta =$
$E = 10.7208,\ acc = 0.6$



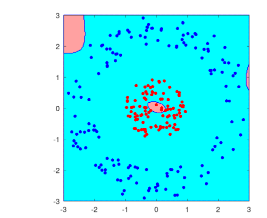
CLog–MGE, $\eta = 0.1$
$E = 11.1456,\ acc = 0.58$



CLogDKP2–MGE, $\eta = 0.1$
$E = 0.0000,\ acc = 1$



CLogDKP3–MGE, $\eta =$
$E = 10.7208,\ acc = 0.6$



Exercícios

Exercício 1. Considere um classificador logístico linear com $\varepsilon_{\text{CLog}} = 0.5$.

(a) Sejam $I = 1$ e $\tilde{w} = (1, 2)^\top$. Represente graficamente os pontos cuja classe predita seja 1.

(b) Sejam $I = 2$ e $\tilde{w} = (1, 2, 3)^\top$. Represente graficamente os pontos cuja classe predita seja 1.

(c) Sejam $I = 2$ e $\tilde{w} = (-1, -2, -3)^\top$. Represente graficamente os pontos cuja classe predita seja 1.

Exercícios

Exercício 2. Seja a base de dados $D = (x^n, y^n)_{n=1}^N$, $x^n \in \mathbb{R}^I$, $y^n \in \{0, 1\}$. O objetivo deste exercício é deduzir a expressão do gradiente da função custo E da Classificador logístico

$$
\begin{aligned}
&E(\tilde{w}; D) = \frac{1}{N} \sum_{n=1}^{N} E_n(\tilde{w}; x^n, y^n), \\
&E_n(\tilde{w}; x^n, y^n) = -y^n \ln(\hat{p}^n) - (1 - y^n) \ln(1 - \hat{p}^n), \\
&\hat{p}^n = \sigma(\tilde{w} \cdot \tilde{x}^n).
\end{aligned}
$$

Seja $n \in \{1, \ldots, N\}$.

(a) Mostre que

$$\sigma'(z) = \sigma(z)(1 - \sigma(z)).$$

(b) Mostre que

$$\frac{\partial}{\partial \tilde{w}_j} \left(y^n \ln(\hat{p}^n)\right) = y^n (1 - \hat{p}^n) \tilde{x}_j^n.$$

Exercício 2. (cont.)

(c) Mostre que

$$\frac{\partial}{\partial \tilde{w}_j}\left((1-y^n)\ln(1-\hat{p}^n)\right) = (y^n - 1)\hat{p}^n \tilde{x}_j^n.$$

(d) Mostre que

$$\frac{\partial}{\partial \tilde{w}_j} E_n(\tilde{w}; x^n, y^n) = (\hat{p}^n - y^n)\tilde{x}_j^n$$

(e) Mostre que

$$\nabla E(\tilde{w}; D) = \frac{1}{N}\sum_{n=1}^{N} \nabla E_n(\tilde{w}; x^n, y^n) = \frac{1}{N}\sum_{n=1}^{N}(\hat{p}^n - y^n)\tilde{x}^n.$$

Exercícios

Exercício 3. Seja a base de dados $D = (x^n, y^n)_{n=1}^N$, $x^n \in \mathbb{R}^I$, $y^n \in \{0,1\}$. O objetivo deste exercício é deduzir a expressão do gradiente da função custo E da Classificador logístico na versão dual

$$E(\alpha; D) = \frac{1}{N}\sum_{n=1}^{N} E_n(\alpha; x^n, y^n),$$

$$E_n(\alpha; x^n, y^n) = -y^n \ln(\hat{p}^n) - (1-y^n)\ln(1-\hat{p}^n),$$

$$\hat{p}^n = \sigma\left(\sum_{\ell=1}^{N} \alpha_\ell (\tilde{x}^\ell \cdot \tilde{x}^n)\right).$$

Sejam $n, j \in \{1, \dots, N\}$.

(a) Mostre que

$$\frac{\partial}{\partial \alpha_j}\left(y^n \ln(\hat{p}^n)\right) = y^n(1-\hat{p}^n)(\tilde{x}^j \cdot \tilde{x}^n).$$

(b) Mostre que

$$\frac{\partial}{\partial \alpha_j}\left((1-y^n)\ln(1-\hat{p}^n)\right) = (y^n-1)\hat{p}^n(\tilde{x}^j \cdot \tilde{x}^n).$$

Exercício 3. (cont.)

(c) Mostre que

$$\frac{\partial E_n}{\partial \alpha_j} = (\hat{p}^n - y^n)(\tilde{x}^j \cdot \tilde{x}^n).$$

(d) Mostre que

$$\nabla E(\alpha; D) = \frac{1}{N}\sum_{n=1}^{N} \nabla E_n(\alpha; x^n, y^n) = \frac{1}{N}\sum_{n=1}^{N} (\hat{p}^n - y^n) \begin{bmatrix} \tilde{x}^1 \cdot \tilde{x}^n \\ \tilde{x}^2 \cdot \tilde{x}^n \\ \vdots \\ \tilde{x}^N \cdot \tilde{x}^n \end{bmatrix}.$$

Exercício 4. Seja a base de dados $D = (x^n, y^n)_{n=1}^N$, $x^n \in \mathbb{R}^I$, $y^n \in \{0, 1\}$. Mostre que a expressão do gradiente da função custo E da Classificador logístico com *kernel* polinomial de grau d na versão dual é

$$\nabla E(\alpha; D) = \frac{1}{N} \sum_{n=1}^{N} \nabla E_n(\alpha; x^n, y^n) = \frac{1}{N} \sum_{n=1}^{N} (\hat{p}^n - y^n) \begin{bmatrix} \tilde{(}x^1 \cdot \tilde{x}^n)^d \\ \tilde{(}x^2 \cdot \tilde{x}^n)^d \\ \vdots \\ \tilde{(}x^N \cdot \tilde{x}^n)^d \end{bmatrix}.$$

Exercícios

Exercício 5. Considere a base de dados binária $D = (x^n, y^n)_{n=1}^{6}$ $(I = 2)$ com

$$\begin{aligned} x^1 &= (-1, 1)^\top & y^1 &= 0 \\ x^2 &= (-1, -1)^\top & y^2 &= 1 \\ x^3 &= (0, 0)^\top & y^3 &= 1 \\ x^4 &= (1, 1)^\top & y^4 &= 1 \\ x^5 &= (-1, 0)^\top & y^5 &= 0 \\ x^6 &= (1, -1)^\top & y^6 &= 0 \end{aligned}$$

(a) Aplique o algoritmo CLog–MGE com $\tilde{w}_{(0)} = (0, 0, 0)^\top$, $t_{\max} = 6$, $\eta = 1$ e n dado pela sequência 3, 2, 5, 1, 1, 6 à base de dados D, indicando a *accuracy* que se obteve.

(b) Aplique o algoritmo CLogDKP2–MGE com $t_{\max} = 3$, $\eta = 0.1$ e n dado pela sequência 3, 2, 5 à base de dados D, indicando a *accuracy* que se obteve.

Exercício 6.

(a) Implemente o algoritmo CLog–MGB.

(b) Implemente o algoritmo CLog–MGE.

(c) Implemente o algoritmo CLogDKPd–MGE.

(d) Implemente o algoritmo CLogDKPd–MGE.