Exame ML 2016/2017

1.

a) O modelo é $y = a + st^2$, o dataset fica:

t	t^2	y
5	25	722
10	100	1073
15	225	1178
20	400	1177

Using ridge regression: $(X^T X + \lambda I) w = X^T y$

$$X = \begin{bmatrix} 25 \\ 100 \\ 225 \\ 400 \end{bmatrix} \qquad w^T = \begin{bmatrix} a & b \end{bmatrix}, \qquad y = \begin{bmatrix} 722 \\ 1073 \\ 1178 \\ 1177 \end{bmatrix}$$

$\lambda = 0.5$

↳ Introduzindo o termo bias:

$$X = \begin{bmatrix} 1 & 25 \\ 1 & 100 \\ 1 & 225 \\ 1 & 400 \end{bmatrix}$$

Logo

$$\left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 25 & 100 & 225 & 400 \end{bmatrix} \begin{bmatrix} 1 & 25 \\ 1 & 100 \\ 1 & 225 \\ 1 & 400 \end{bmatrix} + 0{,}5\, I \right) w = X^T y$$

2×4 $\quad$ 4×2

$$\left(\begin{bmatrix} 4 & 750 \\ 750 & 221250 \end{bmatrix} + 0{,}5\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)\begin{bmatrix} a \\ b \end{bmatrix} = X^T y$$

2×2 $\qquad$ 2×1

$$\begin{bmatrix} 4{,}5 & 750 \\ 750 & 221250{,}5 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1 & 1 & 11 \\ 25 & 100 & 225-400 \end{bmatrix}\begin{bmatrix} 722 \\ 1073 \\ 1178 \\ 1177 \end{bmatrix}$$

2×2 $\qquad$ 2×1 $\qquad$ 2×4 $\qquad$ 4×1

$$\begin{bmatrix} 4{,}5a + 750b \\ 750a + 221250{,}5b \end{bmatrix} = \begin{bmatrix} 4150 \\ 861200 \end{bmatrix}$$

2×1 $\qquad$ 2×1

Convirtiendo en sistema de Eqs:

$$\begin{cases} 4{,}5a + 750b = 4150 \\ 750a + 221250{,}5b = 861200 \end{cases} \xrightarrow{(-)} \begin{cases} a = \dfrac{4150 - 750b}{4{,}5} \\ 750\left(\dfrac{4150 - 750b}{4{,}5}\right) + 221250{,}5b = 861200 \end{cases}$$

$$(-)\begin{cases} \dfrac{3112500}{4{,}5} - \dfrac{562500}{4{,}5}b + 221250{,}5b = 861200 \end{cases}$$

a)
$$\begin{cases} \\ \left(221250{,}5 - \frac{562500}{4{,}5}\right) b = 861200 - \frac{3112500}{4{,}5} \end{cases}$$

c)
$$\begin{cases} \\ 96250{,}5\, b = \frac{508600}{3} \end{cases} \Leftrightarrow \begin{cases} \\ b = \frac{508600}{3} \times \frac{1}{96250{,}5} \end{cases} \Leftrightarrow$$

$$\begin{cases} a = \frac{4150 - 250(1{,}76)}{4{,}5} \\ b = 1{,}76 \end{cases} \Leftrightarrow \begin{cases} a = 628{,}89 \\ b = 1{,}76 \end{cases}$$

Modelo: $y = 628{,}89 + 1{,}76t^2$

b)
$y_{related} = 600 + 2t^2$, logo $w = \begin{bmatrix} 600 \\ 2 \end{bmatrix}$

Since we have initializer weights that could help us how to define, the learning task should be defined not as Ridge Regression but a standard Regression that could be executed with Least-Mean-Squares update or ~~deed~~ steepest descent.

so,

for LMS: $w^{t(+1)} = w^{(t)} + (y_i - w^T x_i)\, x_i$

for steepest descent: $w^{(t+1)} = w^{(t)} + \alpha \sum_n (y_n - w^T x_n)\, x_n$

where $w^{(t)} = \begin{bmatrix} 600 \\ 2 \end{bmatrix}$, and α is a learning rate

2.

a) $p(x|y_0) = e^{2\cos(x-1)} / 2\pi(2.2796)$

$p(x|y_1) = e^{3\cos(x+0,9)} / 2\pi(4.8808)$

, if $x=0$ $\begin{cases} p(x|y_0) = 0,206 \\ p(x|y_1) = 0,210 \end{cases}$

como $p(x|y_1) > p(x|y_0)$, a predição seria y_1.

b). Para a predição ser y_0:

$$p(x|y_0) > p(x|y_1), \text{ então:}$$

$$\frac{e^{2\cos(x-1)}}{2\pi(2.2796)} > \frac{e^{3\cos(x-1)}}{2\pi(4.8808)}$$

$$\ln\left(\frac{e^{2\cos(x-1)}}{2\pi(2.2796)}\right) > \ln\left(\frac{e^{3\cos(x-1)}}{2\pi(4.8808)}\right)$$

$$\ln\left(e^{2\cos(x-1)}\right) - \ln\left(2\pi(2.2796)\right) > \ln\left(e^{3\cos(x-1)}\right) - \ln\left(2\pi(4.8808)\right)$$

$$2\cos(x-1) - 3\cos(x-1) > \ln\left(2\pi(2.2796)\right) - \ln\left(2\pi(4.8808)\right)$$

$$+\cos(x-1) < -\left(\ln\left(2\pi(2.2796)\right) - \ln\left(2\pi(4.8808)\right)\right)$$

$$x-1 < \cos^{-1}(0,761)$$

$$x < \cos^{-1}(0,761) + 1 \quad \Leftrightarrow \quad x < 1,76 \text{ rad}$$

Infatto, the prediction is equal to $y_0 : [0, 1,76[$ rad.

c) we can use the Bayes Formula to convert the prior probability to the posterior probability:

$$P(C_j|x) = \frac{p(x|C_j)\,P(C_j)}{p(x)}$$

where $p(x) = \sum_{j=1}^{2} p(x|C_j)\,P(C_j)$

we want:

$$p(y_0|x) = \frac{p(x|y_0)\,P(y_0)}{p(x)}$$

let's assume

$$P(y_0) = P(y_1)$$

$$P(y_0) = P(y_1) = 0.5$$

$$p(x) = p(x|y_0)\,P(y_0) + p(x|y_1) * P(y_1)$$

$$= 0,5\,(p(x|y_0) + p(x|y_1))$$

Ergo

$$p(y_0|x) = \frac{p(x|y_0)\,P(y_0)}{0,5\,(p(x|y_0)+p(x|y_1))} = \frac{p(x|y_0)\,0.5}{0.5\,(p(x|y_0)+p(x|y_1))}$$

$$= \frac{p(x|y_0)}{p(x|y_0)+p(x|y_1)} = \frac{1}{\frac{p(x|y_0)+p(x|y_1)}{p(x|y_0)}} = \frac{1}{1+p(x|y_1)} \quad (A)$$

C4

$$p(x|y_1) = \frac{e^{k_1 \cos(x-\mu_1)}}{2\pi I(k_1)} = \frac{e^{k_1 \cos(x-\mu_1)}}{e^{\ln(2\pi I(k_1))}}$$

$$= e^{k_1 \cos(x-\mu_1) - \ln(2\pi I(k_1))}$$

(2) $$= \boxed{e^{-\ln(2\pi I(k_1)) + k_1 \cos(x-\mu_1)}}$$

Vota:

$$\cos(x) = \sin\left(x + \frac{\pi}{2}\right)$$

Sja $x = x+1$

$$\cos(x+1) = \sin\left(x+1+\frac{\pi}{2}\right)$$

Sja $x = x-\mu_1$

$$\cos(x-\mu_1) = \sin\left(x-\mu_1+\frac{\pi}{2}\right)$$

Logo:

$$(1) = e^{-\ln(2\pi I(k_1)) + k_1 \sin\left(x-\mu_1+\frac{\pi}{2}\right)}$$

Tal que: $-\ln(2\pi I(k_1)) = w_0$

$k_1 = w_1$

$\theta = \mu_1 + \frac{\pi}{2}$

Logo, fazendo (1) e (2)

$$p(y|x) = \frac{1}{1 + e^{w_0 + w_1 \sin(x-\theta)}}$$

Cqd

3. see midterm_sol.pdf

a)

(a) and (b) are appropriate to use in classification. In (c), there is very little penalty for extremely misclassified examples, which correspond to very negative $y F(x)$. In (d) and (e), correctly classified examples are penalized, whereas misclassified examples are not. In general, L should approximate the 0-1 loss, and it should be a non-increasing function of $y F(x)$.

b) Function (b) is more robust to outliers. For outliers, $yF(x)$ is often very negative. In (a), outliers are heavily penalized. So, the resulting classifier is largely affected by the outliers. On the other hand, in (b), the loss of outliers is bounded. So, the resulting classifier is less affected by the outliers, and thus more robust.

c) To obtain the parameters in $F(x)$, we need to ~~maximize~~ minimize

$$\sum_i L(y^i F(x^i)) = \sum_i \frac{1}{1+\exp(y^i F(x^i))}$$

$$= \sum_i \frac{1}{1+\exp(y^i (w_0 + \sum_{j=1}^{d} w_j x_j^i))}$$

Gradient Descent, logo:

$$\frac{\partial}{\partial w_0} \sum_i L(y^i F(x^i)) = \sum_i \frac{\partial}{\partial w_0} \left(\frac{1}{1 + \exp(y^i (w_0 + \sum_{j=1}^{d} w_j x_j^i))} \right)$$

$$= -\sum_i \frac{y^i \exp(y^i F(x^i))}{(1 + \exp(y^i F(x^i)))^2} ;$$

Ajou, $\forall k = 1, \ldots, d$

$$\frac{\partial}{\partial w_k} \sum_i L(y^i F(x^i)) = \sum_i \frac{\partial}{\partial w_k} \left(\frac{1}{1 + \exp(y^i (w_0 + \sum_{j=1}^{d} w_j x_j^i))} \right)$$

$$= -\sum_i \frac{y^i x_k^i \exp(y^i F(x^i))}{(1 + \exp(y^i F(x^i)))^2}$$

Therefore, the update rules are as follows:

$$w_0^{(t+1)} = w_0^{(t)} - \eta \frac{\partial}{\partial w_0} \sum_i L(y^i F(x^i)) =$$

$$w_0^{(t+1)} = w^{(t)} + \eta \sum_i \frac{y^i \exp(y^i F(x^i))}{(1 + \exp(y^i F(x^i)))^2}$$

Para $\forall k = 1, \ldots, d$

$$w_k^{(t+1)} = w_k^{(t)} - \mu \frac{\partial}{\partial w_k} \sum_i L(y^i F(x^i))$$

$$w_k^{(t+1)} = w_k^{(t)} + \mu \sum_i \frac{y^i x_k^i \exp(y^i F(x^i))}{(1 + \exp(y^i F(x^i)))^2}$$

4.

$P(A) = \frac{1}{2}$; $P(B) = \mu$; $P(C) = 3\mu$; $P(D) = \frac{1}{2} - 4\mu$

c students get a "(C)"

d students get a "D"

h students get a "(A)" or "(B)"

$a + b = h$

Total class

$c + d + h$

$c + d + a + b$

EM algorithm to estimate μ : See final 2008f - solution.pdf

a)

$$\hat{a} = \frac{P(A)}{P(A) + P(B)} h \quad ; \quad \hat{b} = \frac{P(B)}{P(A) + P(B)} h$$

c) $\hat{a} = \frac{\frac{1}{2}}{\frac{1}{2} + \mu} h$; $\hat{b} = \frac{\mu}{\frac{1}{2} + \mu} h$

5)

$\hat{\mu}$: ~~$h-a+c$~~ / $6(h-a+c+d)$ $\quad$ $\hat{\mu}$: $\frac{h-a+c}{(3\times 4)(h-a+c+d)}$ $\quad$: $\frac{h-a+c}{12(h-a+c+d)}$

See example on:

200Bf-solution.pdf

5 HMM (see midterm-solution.pdf)

a) Primo ponto, matrizes

Transition matrix:

	S	V
S	0.8	0.2
V	0.4	0.6

Emission Matrix:

	Grim	Frow
S	0.5	0.5
V	0.8	0.2

p_{i0} =

S	V
$\frac{1}{2}$	$\frac{1}{2}$

Viterbi algorithm: (slide 43 Capítulo II)

Input: 6 6 F 6

Inicialização

V_i^K: $P(x_i | y_i^K = 1)\,\pi_K$

Iteração

$V_t^K = P(x_t | y_t = 1) \max_i a_{i,K} V_{t-1}^i$

$P_{tr}(K,t) = \arg\max_i a_{i,K} V_{t+1}^i$

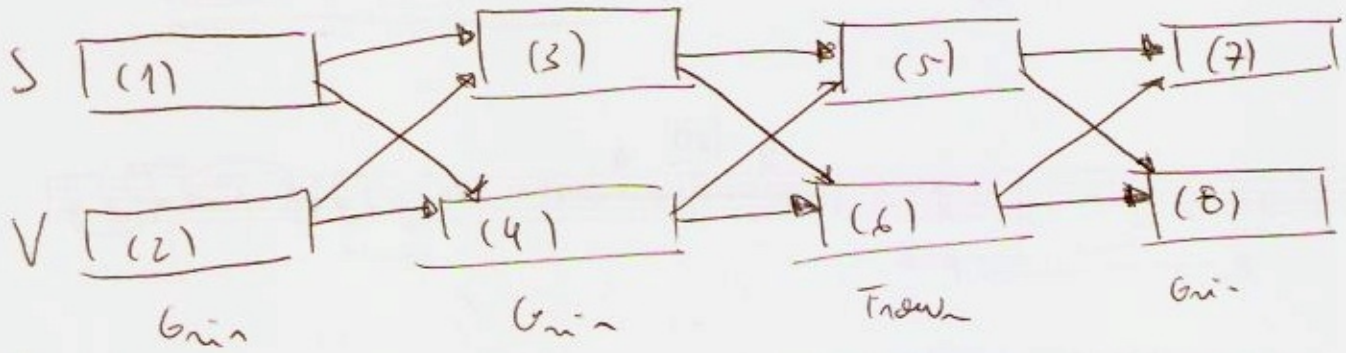
Fim

$P(x, y^*) = \max_K V_T^K$

Traceback

$$y_t^* = \text{argmax}_k \, v_t^k \quad ; \quad y_{t+1}^* = ptr(y_t^*, t)$$

Rebus:



Calculus

1. $p_S^0 \times p(\text{Grün} \mid S) = \frac{1}{2} \times 0.5 = \frac{1}{2} \times \frac{1}{2} = \boxed{\frac{1}{4}}$

2. $p_V^0 \times p(\text{Grün} \mid V) = \frac{1}{2} \times 0.8 = \frac{1}{2} \times \frac{4}{5} = \frac{4}{10} = \boxed{\frac{2}{5}}$

3. $p(\text{Grün} \mid S) \times \max \begin{cases} (1) \times p(S \mid S) = \frac{1}{4} \times 0.8 = \boxed{0.2} \Rightarrow \text{the max!} \\ (2) \times 0.4 = \frac{2}{5} \times p(V \mid S) = 0.16 \end{cases}$

↳ $0.5 \times 0.2 = \boxed{0.1}$

4. $p(\text{Grün} \mid V) \times \max \begin{cases} (1) \times p(S \mid V) = \frac{1}{4} \times 0.2 = 0.05 \\ (2) \times p(V \mid V) = \frac{2}{5} \times 0.6 = \boxed{0.24} \Rightarrow \text{the max!} \end{cases}$

↳ $0.8 \times 0.24 =$

$= \boxed{0,192}$

5.

$p(\text{Frown} \mid S) \times \max \begin{cases} (3) \times p(S|S) = 0.1 \times 0.8 = \boxed{0.08} \text{ max!} \\ (4) \times p(S|V) = 0.192 \times 0.4 = 0.0768 \end{cases}$

↳ $0.5 \times 0.08 = \boxed{0.04}$

6. $p(\text{Frown} \mid V) \times \max \begin{cases} (3) \times p(V|S) = 0.1 \times 0.2 = 0.02 \\ (4) \times p(V|V) = 0.192 \times 0.6 = \boxed{0.1152} \text{ max!} \end{cases}$

↳ $0.2 \times 0.1152 = \boxed{0.02304}$

7. $p(\text{Grin} \mid S) \times \max \begin{cases} (5) \times p(S|S) = 0.04 \times 0.8 = \boxed{0.032} \text{ max!} \\ (6) \times p(S|V) = 0.02304 \times 0.4 = \boxed{0.009216} \end{cases}$

↳ $0.5 \times 0.032 = \boxed{0.016}$

8. $p(\text{Grin} \mid V) \times \max \begin{cases} (5) \times p(V|S) = 0.04 \times 0.2 = 8 \times 10^{-3} \\ (6) \times p(V|V) = 0.02304 \times 0.6 = \boxed{0.013824} \text{ max} \end{cases}$

↳ $0.8 \times 0.013824 = \boxed{0.0110592}$

A probabilidade do caminho mais provável termina em $Y_t = K$; logo no último retângulo (7) é mais que (8), onde o caminho mais provável é "SSSS".

5) Para mudar, há que contar:

Nº de 6 sabendo S: 3

Nº de 6 sabendo V: 8

Nº de F sabendo S: 8

Nº de F sabendo V: 1

Nº de Transições S → S: 6

Nº de Transições V → S: 4

Nº Transições S → V: 5

Nº de Transições V → V: 4

Transition Matrix

$$\begin{array}{c} \\ S \\ V \end{array} \begin{array}{c} \begin{array}{cc} S & V \end{array} \\ \begin{bmatrix} \frac{6}{11} & \frac{5}{11} \\ \frac{4}{8} & \frac{4}{8} \end{bmatrix} \end{array} = \begin{bmatrix} \frac{6}{11} & \frac{5}{11} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Aqui o smoothing

$\lambda = 0$

Se fosse $\lambda = 0.1$

ia só adicionar 0.1 ao numerador.

Emission Matrix:

$$\begin{array}{c} \\ S \\ V \end{array} \begin{array}{c} \begin{array}{cc} 6 & F \end{array} \\ \begin{bmatrix} \frac{3}{11} & \frac{8}{11} \\ \frac{8}{9} & \frac{1}{9} \end{bmatrix} \end{array}$$

Estado

E	$p(E \mid X = S)$
Sim	3/11
Não	8/11

E	$p(E \mid X = V)$
Sim	8/9
Não	1/9

HMM: