POLYTECH NICE-SOPHIA

Classifieur de Bayes et approximation softmax

ANDRIEU Grégoire
GILLE Cyprien
NEGRE Philippe
YOUSSEF Alan

2020/2021

Référent : Mr FILLATRE Lionel

I. Introduction

Sujet :

L’approximation softmax, utilisée à la place de la fonction argmax, dans un classifieur de Bayes

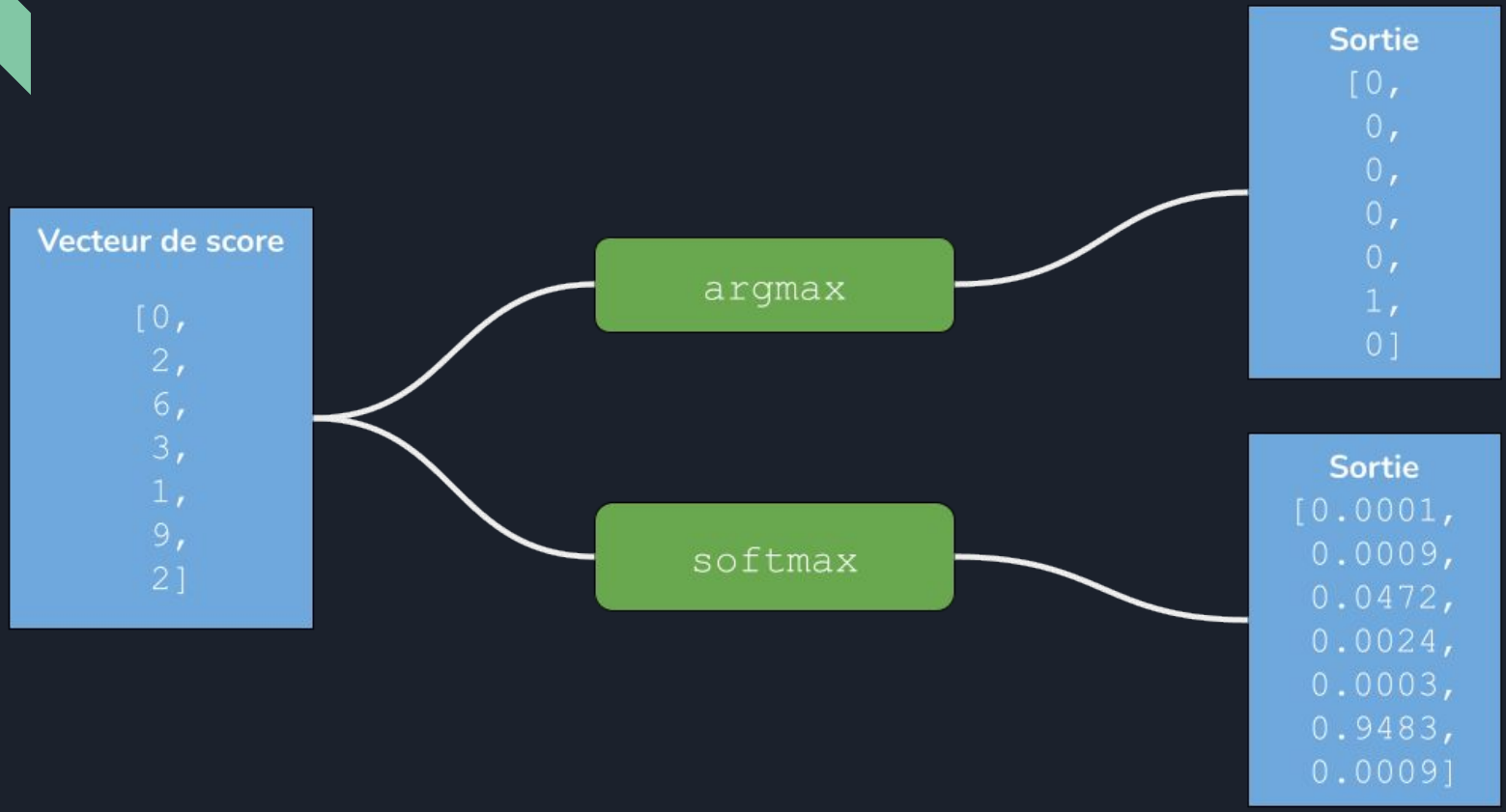
II. Présentation des objets

- Nos fonctions d'étude :

$$argmax : \mathbb{R}^n \to \{0,1\}^n$$
$$V = [v_i] \mapsto W = [w_i] \; avec \begin{cases} w_i = 1 \; si \; v_i = \max\limits_i(V) \\ 0 \; sinon \end{cases}$$

$$softmax : \mathbb{R}^n \to [0,1]^n$$
$$V = [v_i] \mapsto W = [w_i] \; avec \; w_i = \frac{\exp(v_i)}{\sum_{k=1}^{n} \exp(v_k)}$$

II. Présentation des objets



II. Présentation des objets

- Classification

Associer à une observation X une classe Y :

$$D(X) = Y_i$$

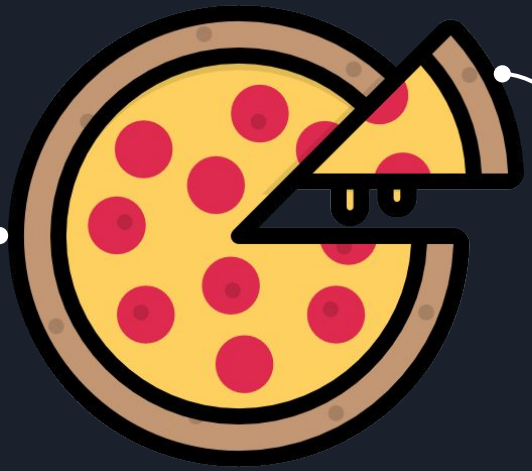
- Règle de Bayes

$$\hat{y} = \underset{k \in \{1,\dots,K\}}{argmax}\ p(Y = k) \prod_{i=1}^{n} p(X_i \mid Y = k)$$

III. Etude d’un cas particulier

3 Classes :

- Excellente
- Ok
- Honteuse



4 features :

- Degré de cuisson
- Nombre de pepperonis
- Quantité de fromage
- Rayon

III. Etude d’un cas particulier

- Répartition dans les classes :

P(Y)	0 : Excellente	1 : Ok	2 : Honteuse
	0.30	0.50	0.20

- Exemple d’une feature : le degré de cuisson (1 = pas cuite, 5 = en cendre)

DdC selon acceptabilité	1	2	3	4	5
Excellente	0.0	0.1	0.8	0.1	0.0
Ok	0.0	0.3	0.4	0.3	0.0
Honteuse	0.4	0.1	0.0	0.1	0.4

III. Etude d’un cas particulier

```
print("\nClassification d'une pizza la plus mauvaise possible")
mauvaise_pizza = [1, 0, 1, 10]
pizza_bc.classify(mauvaise_pizza)
pizza_bc.bayes_softmax(mauvaise_pizza)
```

```
Classification d'une pizza la plus mauvaise possible
Classe: 2
Classe 0. Confiance: 0.2119
Classe 1. Confiance: 0.2119
Classe 2. Confiance: 0.5761
[Finished in 0.7s]
```

III. Le risque d'erreur global

- Expressions matricielles:

Notons N le nombre d'observations (i.e. de données), K restant le nombre de classes. Notre matrice de probabilités A' est alors de dimension (K, N):

$$A' = \begin{pmatrix} P(Y = 1 \mid X_1) & \dots & P(Y = 1 \mid X_N) \\ \vdots & \ddots & \vdots \\ P(Y = K \mid X_1) & \dots & P(Y = K \mid X_N) \end{pmatrix}$$

On applique le théorème de Bayes, et on obtient la matrice A des probabilités postérieures, où apparaît la première dépendance en π.

$$A = \begin{pmatrix} P(X_1 \mid Y = 1)\pi_1 & \dots & P(X_N \mid Y = 1)\pi_1 \\ \vdots & \ddots & \vdots \\ P(X_1 \mid Y = K)\pi_k & \dots & P(X_N \mid Y = K)\pi_k \end{pmatrix}$$

III. Le risque d'erreur global

- Expressions matricielles:

- Argmax:

$$D_j^* = \underset{i}{argmax}[A_j]$$

$$d_{ij}^* = \left(\underset{k}{argmax} P(X_j \mid Y = k)\pi_k \right)_i$$

$$D^* = \underset{D \in \mathbf{D}}{argmin}\ R(D)$$

- Softmax:

$$\tilde{D}_j = softmax[A_j]$$

$$\tilde{d}_{ij} = \frac{\exp(\pi_i P(X_j \mid Y = i))}{\sum_{k=1}^{K} \exp(\pi_k P(X_j \mid Y = k))}$$

III. Le risque d'erreur global

- L'expression du risque :

$$R(D, \pi) = \sum_{i \in \{1,\dots,K\}} \pi_i R_i(D)$$

- Risque conditionnel:

$$R_i(D) = \sum_j \sum_{k \neq i} P(Y = k | X_j)$$
$$= \sum_j C^i D$$

Avec C^i un vecteur $(1, K)$ tel que :

$$C^i_j = \begin{cases} 1 \ si \ j \neq i \\ 0 \ sinon \end{cases}$$

III. Le risque d'erreur global

- Expressions du risque:

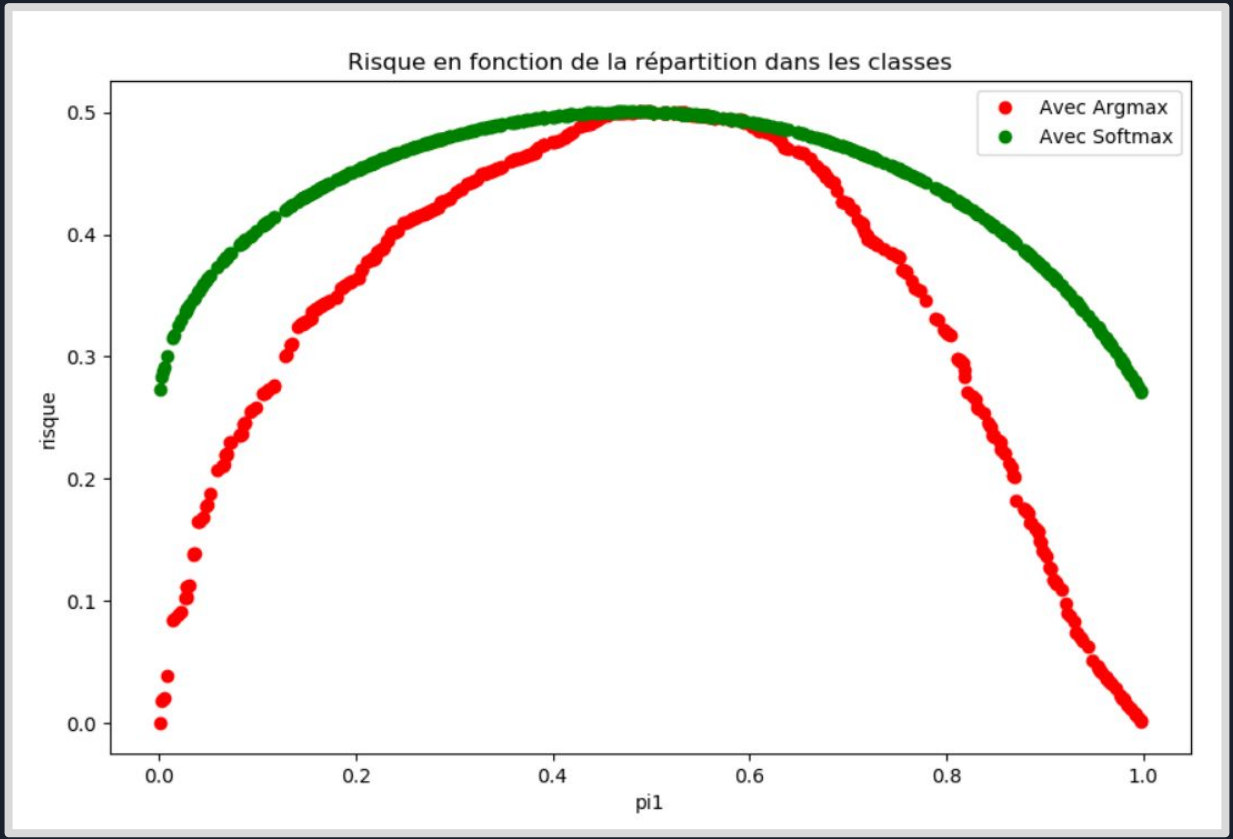
- Argmax:

$$R(D^*, \pi) = \sum_{i \in \{1,\dots,K\}} \pi_i R_i(D^*) = \sum_i \sum_j \pi_i C^i D^*$$

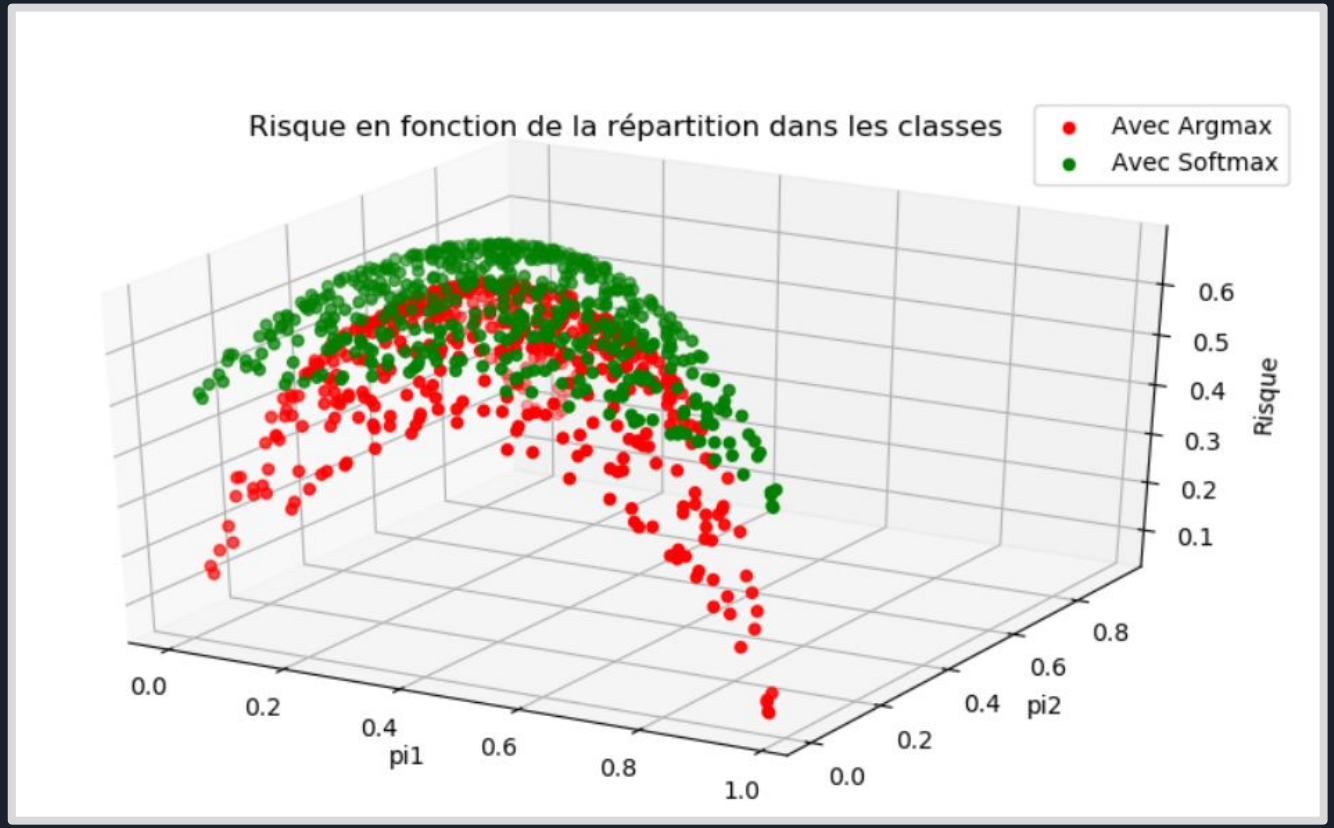
- Softmax:

$$R(\tilde{D}, \pi) = \sum_{i=1}^{K} \sum_{j=1}^{N} \pi_i (1 - \delta_{ij}) \frac{\exp(\pi_i P(X_j \mid Y = i))}{\sum_{k=1}^{K} \exp(\pi_k P(X_j \mid Y = k))}$$

III. Le risque d'erreur global



III. Le risque d'erreur global



III. Le risque d'erreur global

- Concavité
  - Risque Argmax
  - Risque Softmax

- Gradient

$$\frac{\partial R(\tilde{D}, \pi)}{\partial \pi_i} = \sum_{j=1}^{N} \left[(1 - \delta_{ij}) \left(S_i + \pi_i S_i (1 - S_i) \right) + \sum_{k \neq i}^{K} (1 - \delta_{kj})(-S_k S_i) \right]$$

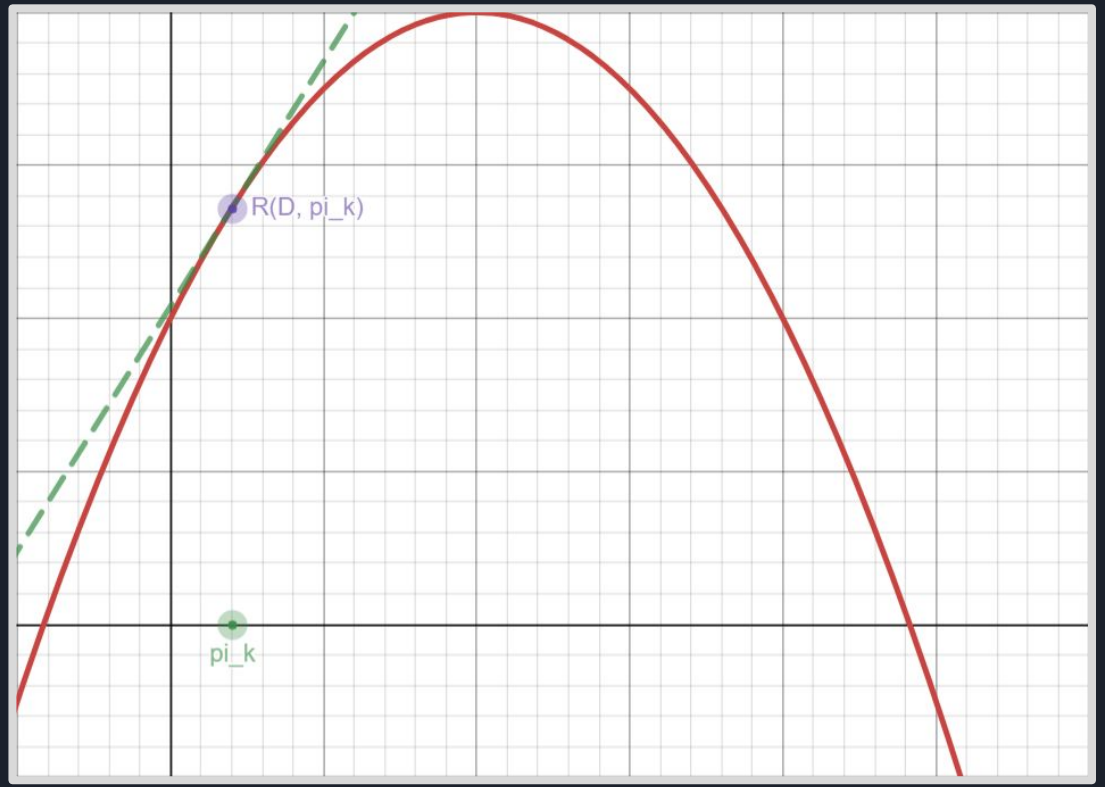
$$S_i = \frac{\exp(a_i)}{\sum_{k=1}^{K} \exp(a_k)}$$

IV. Maximisation du risque : méthode du gradient projeté

- Montée du gradient :
  - Paramètres :
    - Nombre d'étapes
    - π_0 (Point de départ)
    - η (learning rate, ou facteur mutliplicatif du pas) .

IV. Maximisation du risque : méthode du gradient projeté

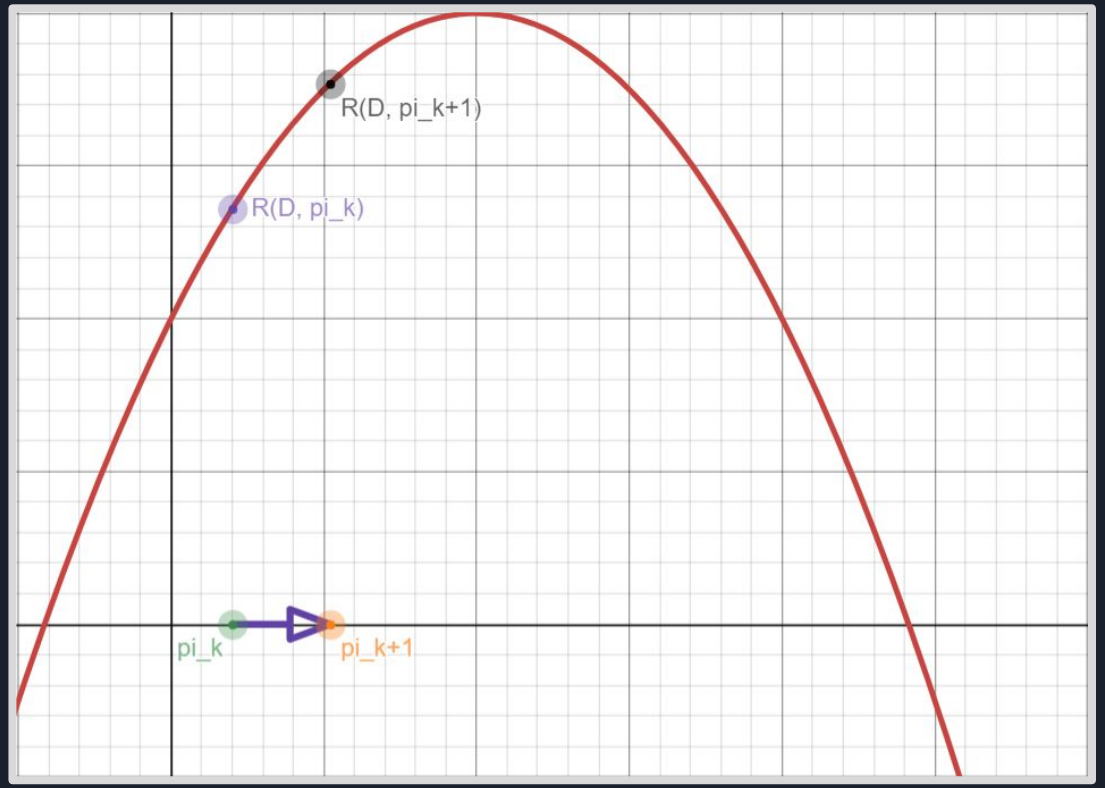
- Déroulé d'une étape:
  1. Evaluation du gradient en $\boldsymbol{\pi}_k$



IV. Maximisation du risque : méthode du gradient projeté

- Déroulé d'une étape:

1. Evaluation du gradient en $\boldsymbol{\pi}_k$
2. $\boldsymbol{\pi}_{k+1} = \boldsymbol{\pi}_k + \eta * \text{gradient}$

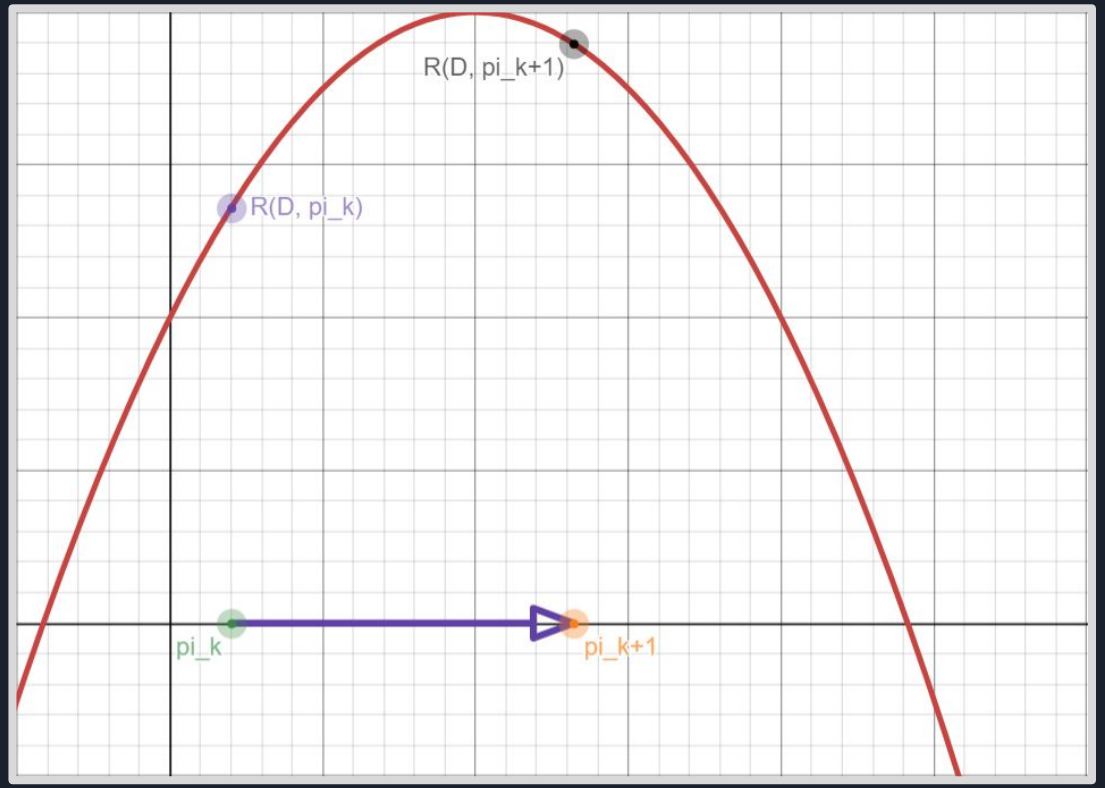


IV. Maximisation du risque : méthode du gradient projeté

- Déroulé d'une étape:

  1. Evaluation du gradient en $\boldsymbol{\pi}_k$
  2. $\boldsymbol{\pi}_{k+1} = \boldsymbol{\pi}_k + \eta * \text{gradient}$

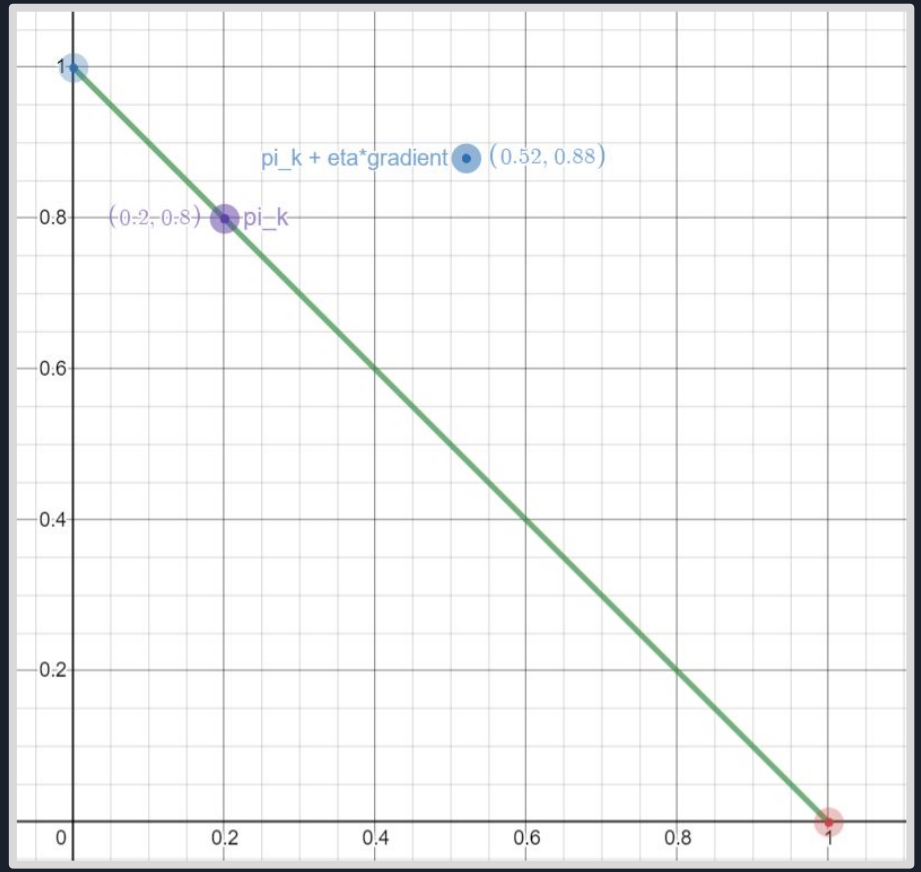
Influence du choix de η :



IV. Maximisation du risque : méthode du gradient projeté

- Problème:

Perte des propriétés probabilistes!



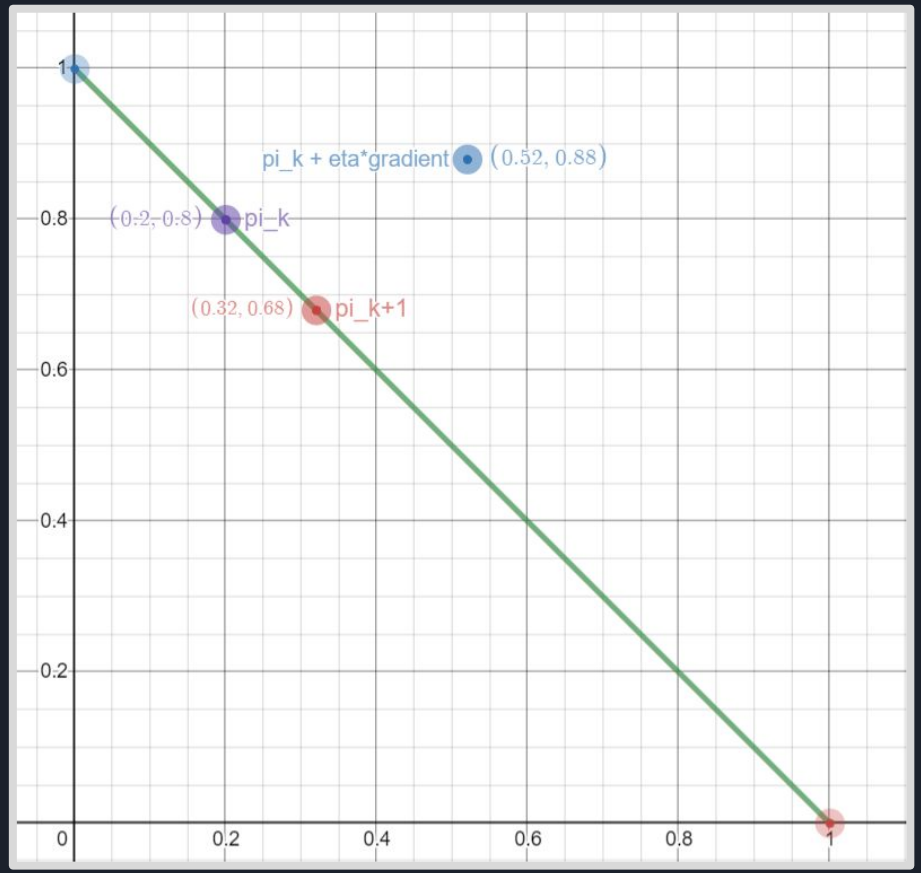
IV. Maximisation du risque : méthode du gradient projeté

- Problème:

Perte des propriétés probabilistes!

- Solution:

Projection sur le simplex



IV. Maximisation du risque : méthode du gradient projeté

- Principe de la projection sur le simplex:

Le point du K-simplex le plus proche de π_i est:

$$t_i = \max_i \{\pi_i - \Delta_i, 0\}$$

Avec :

$$\Delta_i = \frac{\left(\sum_{j=i+1}^{K} \pi_j\right) - 1}{K - i}$$

Avec les π_i triés par ordre croissant.

IV. Maximisation du risque : méthode du gradient projeté

```
Etape: 0 | [0.1, 0.2, 0.7]
Etape: 20 | [0.24258121 0.27990927 0.47750952]
Etape: 40 | [0.29918982 0.31194929 0.38886089]
Etape: 60 | [0.32166766 0.32480849 0.35352385]
Etape: 80 | [0.33059316 0.32997157 0.33943526]
Etape: 100 | [0.33413718 0.33204496 0.33381786]
[Finished in 0.7s]
```

Convergence vers la situation attendue : l'équiprobabilité à priori

V. Pistes futures

- Différence entre les risques argmax et softmax (méthode alpha)
- Influence du choix argmax/softmax sur l'entraînement d'un classifieur non-idéal ?
- Je manque d'idées aled

Bibliographie

Classification bayésienne et risque, expressions et calculs:
ECE531 Lecture 2a: A Mathematical Model for Hypothesis Testing (wpi.edu)
ECE531 Lecture 2b: Bayesian Hypothesis Testing (wpi.edu)
ECE531 Lecture 3: Minimax Hypothesis Testing (wpi.edu)

Projection sur le simplex:
http://www.gipsa-lab.fr/~laurent.condat/download/proj_simplex_l1ball.m
https://arxiv.org/pdf/1101.6081.pdf

Concavité de la fonction softmax:
https://arxiv.org/pdf/1502.04635.pdf