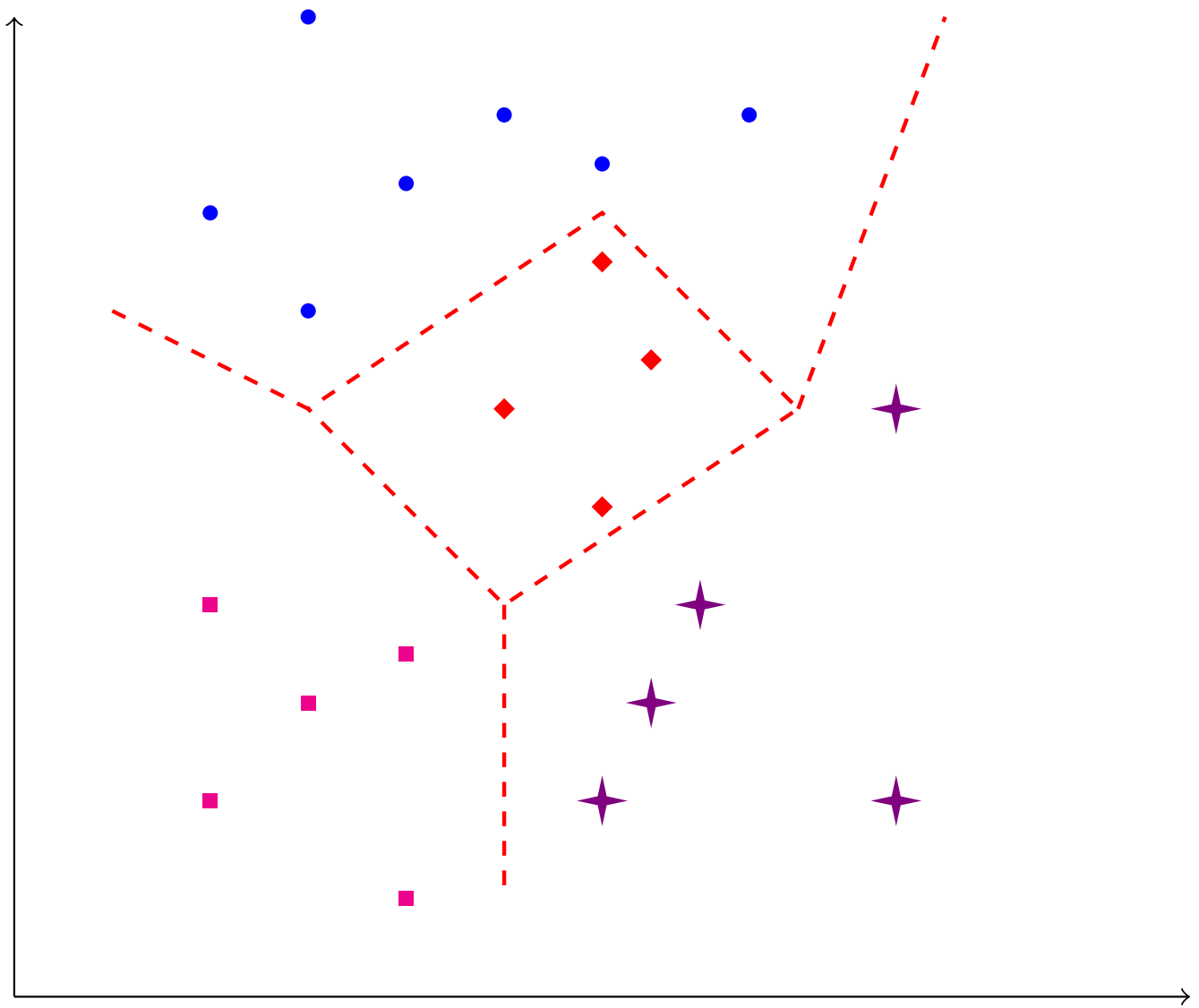
Probabilités imprécises appliquées à la classification par décomposition binaire

Sebastien Destercke and Benjamin Quost

Heuristic and Diagnosis for Complex Systems (HEUDIASYC) laboratory, Compiegne, France

Seminaire

Le problème



Plan

Definition

Incertitude: ne pas pouvoir répondre exactement à une question dans un contexte donné

- On ne sait pas si une proposition est vraie ou fausse
- On ne sait pas si un événement va se produire

par exemple

- Va-t-il pleuvoir demain?
- Ce patient est-il malade? Quelle maladie?
- Quels sont les risques que le niveau de la seine dépasse un seuil donné?
- Quel sera le score de ce match?

$\rightarrow$ informations rendent souvent certaines hypothèses plus crédibles que d'autres.

Approche classique probabiliste

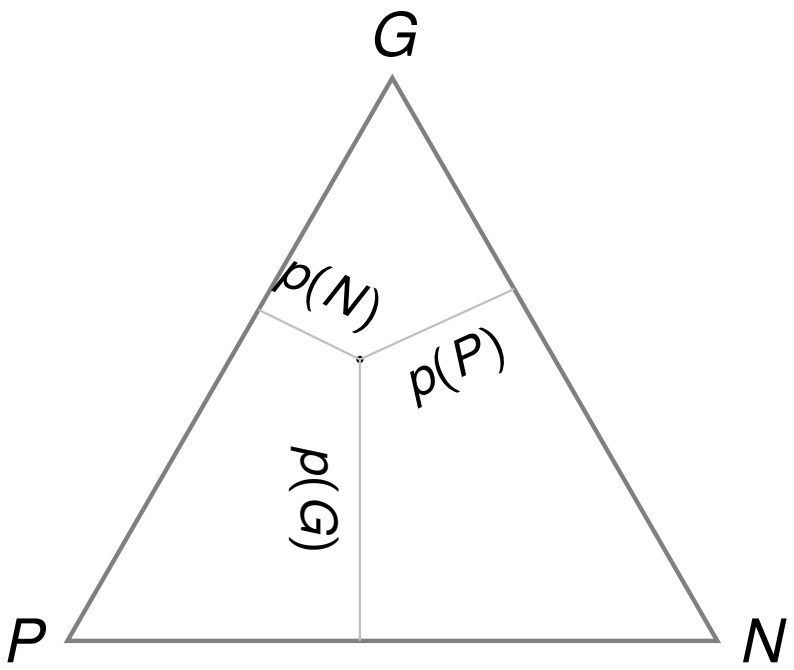
Un ensemble $\mathcal{X} = \{x_1, \ldots, x_n\}$ d'alternatives **mutuellement exclusives**

Incertitude sur $\mathcal{X}$ modélisée par une distribution de probabilité $p : \mathcal{X} \to [0, 1]$

- $p(x_i) \geq 0$ et $\sum_{x_i \mathcal{X}} p(x_i) = 1$ (axiomes)
- $P(A) = \sum_{x_i \in A} p(x_i)$ mesure la vraisemblance de A

- Interprétation fréquentiste: $p(x_i)$ nbre d'observations de x_i sur nbre total $\to$ besoin d'une population
- Interprétation subjective (De Finetti): $P(A) =$ somme qu'un agent est prêt à parier s'il reçoit 1 quand A arrive $\to$ pas besoin d'une "population"

Un petit exemple



Un match et $\mathcal{X} = \{G, N, P\}$ résultat de la première équipe.

Les contraintes

$$p(x_i) \geq 0 \text{ et } \sum_{\mathcal{X}} p(x_i) = 1$$

correspondent au simplexe.

Exemple:

$$p(G) = 0.5,$$
$$p(P) = 0.3,$$
$$p(N) = 0.2$$

Oui mais. . .

La théorie probabiliste suppose qu'en pratique, tout état d'incertitude peut se modéliser par une proba, mais souvent

- Informations sur probabilités souvent partielles
- Données imprécises, non-fiables, en faible nombre (comment calculer fréquences?)
- Loi uniforme comme modèle d'ignorance critiquable

Probabilité imprécises: outil de base

Considérer l'ensemble probabiliste induit par nos connaissances

- Résumer connaissances sur $p(x_i)$ par contraintes linéaires
- Travailler avec ensembles de solutions possibles.

Ces contraintes donnent un programme linéaire

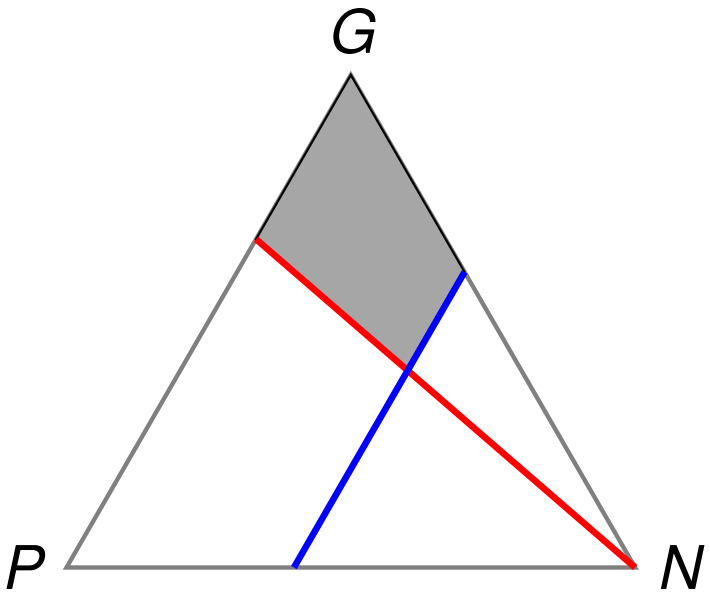
$$A\mathbf{p} \leq \mathbf{B}$$

avec

$$p(x_i) \geq 0 \text{ et } \sum_{\mathcal{X}} p(x_i) = 1$$

qui sont les variables $\rightarrow$ solution unique si information suffisante **et consistente**

Exemple: ensemble de proba.



$\mathcal{X} = \{G, N, P\}$ avec les infos suivantes:

$$2p(P) \leq p(G)$$

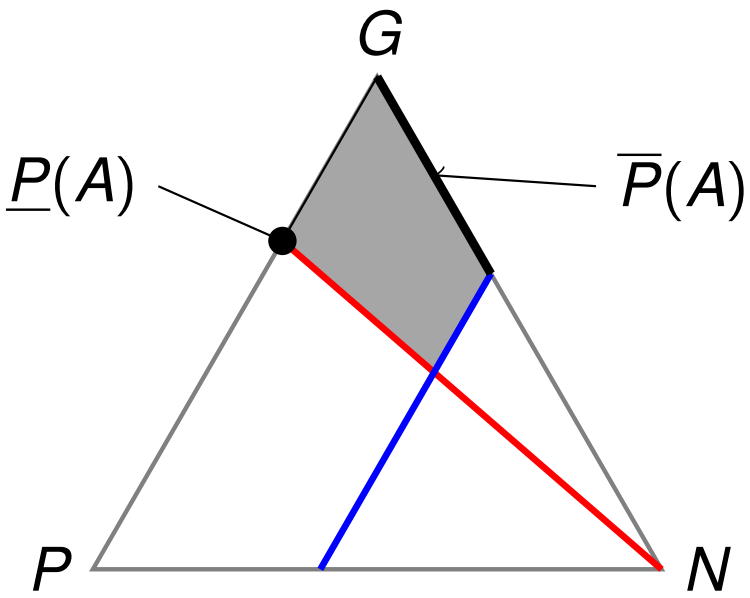
$$p(N) \leq 0.4$$

avec

$$p(x_i) \geq 0 \text{ et } \sum_{\mathcal{X}} p(x_i) = 1$$

Ens. p possibles=ensemble des solutions d'un prog. linéaire.

Exemple: probabilités sup. et inf.



Connaissance sur l'événement "ne pas perdre" $A = \{G, N\}$

$\underline{P}(A)$: minimiser $p(G) + p(N)$

$\overline{P}(A)$: maximiser $p(G) + p(N)$

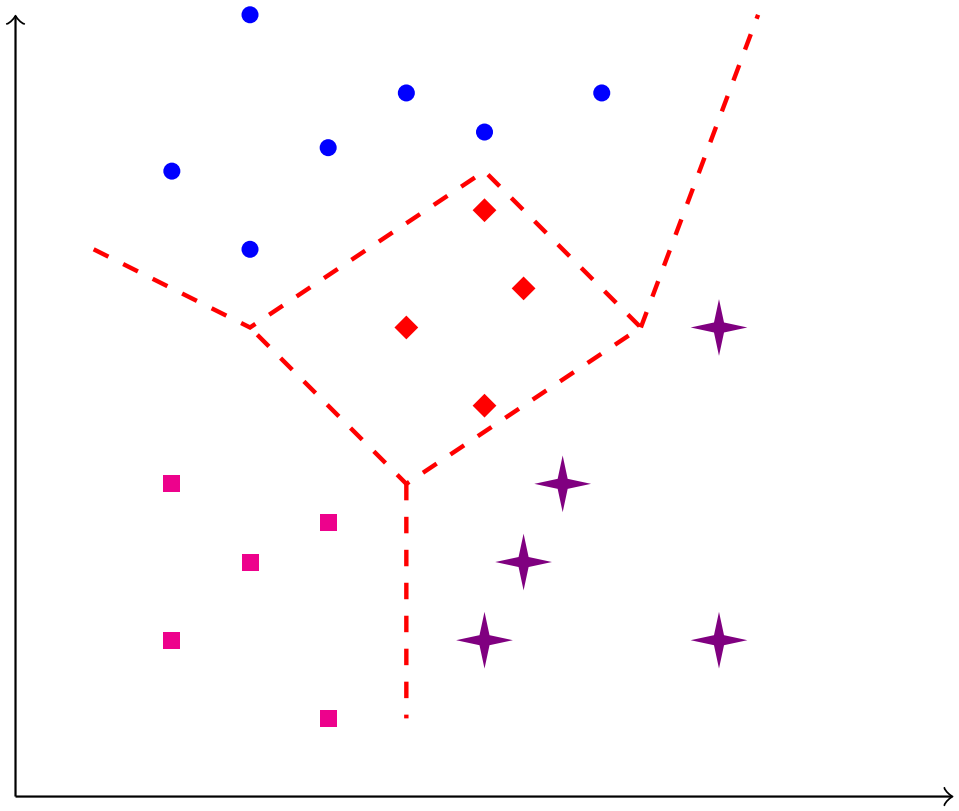
sous contraintes linéaires.

$$\underline{P}(A) = 2/3 \; ; \; \overline{P}(A) = 1$$

Quelques propriétés générales

- $\underline{P}(\mathcal{X}) = 1$
- $\underline{P}(A \cup B) \geq \underline{P}(A) + \underline{P}(B)$ si $A \cap B = \emptyset$
- $\underline{P}(A) = 1 - \underline{P}(A^c)$

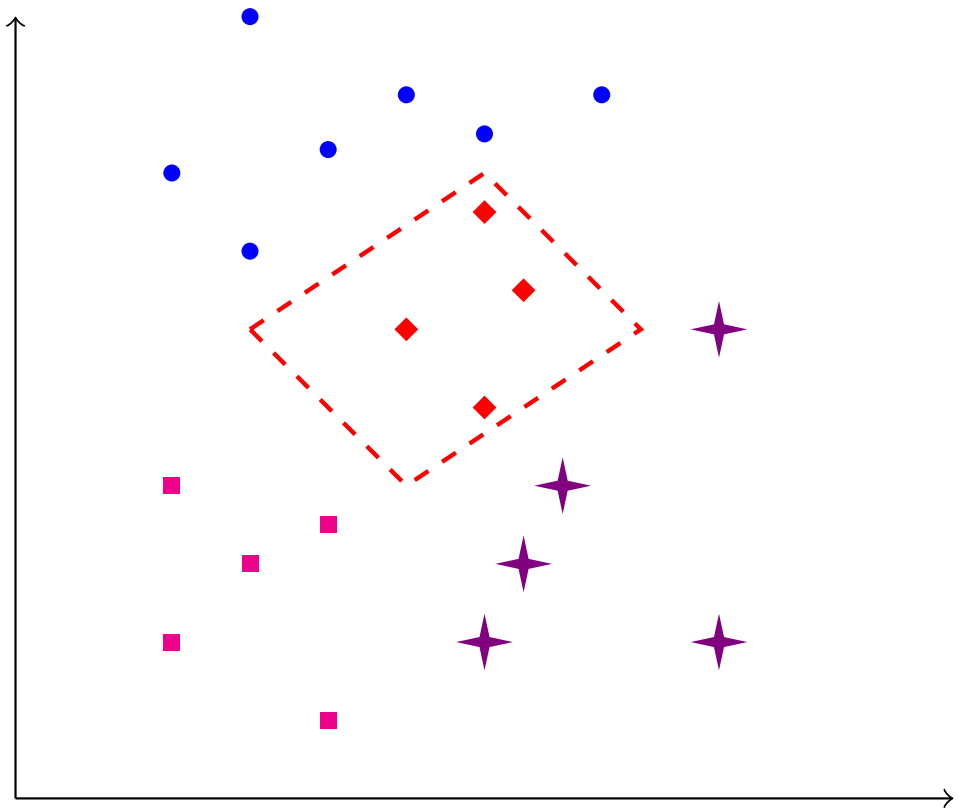
Pourquoi décomposer?



Classification classique

- **Pour** : un seul classifieur
- **Contre** : frontière difficile à apprendre

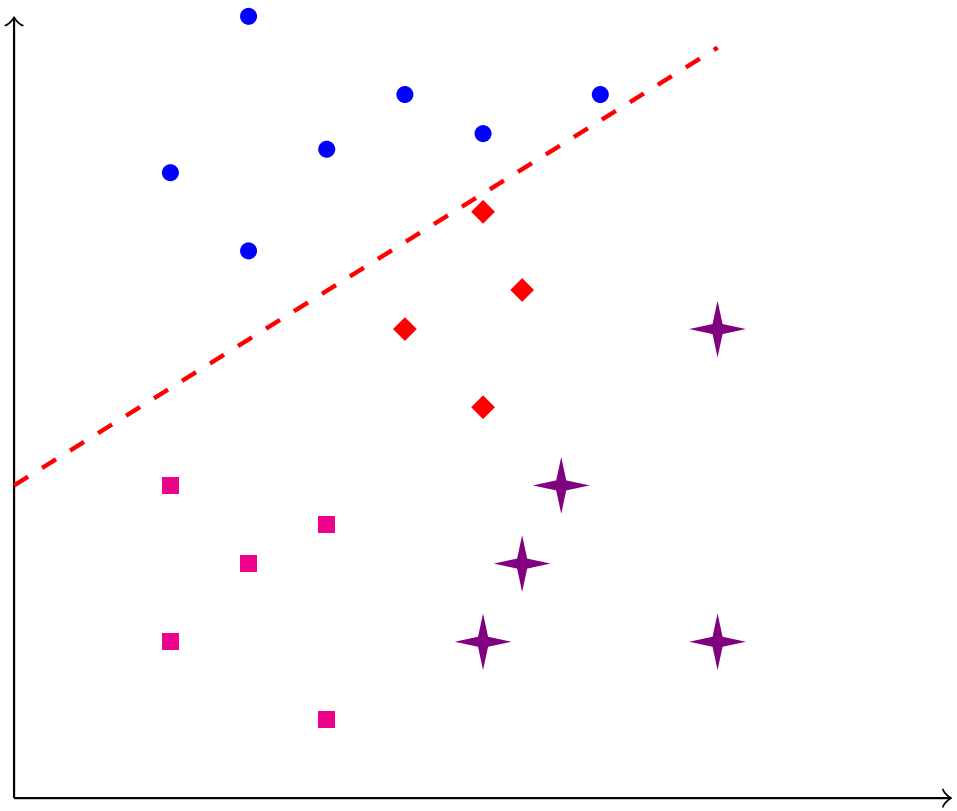
Un contre tous (◆ contre tous)



Un contre tous

- **Pour** : frontière plus facile et nbre classifieurs= nbre de classes
- **Contre** : frontière reste complexe et plusieurs classifieurs

Un contre un (◆ contre ●)



Un contre un

- **Pour** : frontières les plus simples possibles
- **Contre** : + grand nombre de classifieurs

Le problème: notation et introduction

Classique

- Un ensemble d'exemples $(\mathbf{x_i}, y_i)$, $i = 1, \ldots, N$ avec $y_i \in \mathcal{W} = \{w_1, \ldots, w_M\}$ des classes observées.
- Apprendre un classifieur $\mathcal{C}$ prédisant, à partir d'une nouvelle entrée $\mathbf{x}$ donnée, la classe y de sortie.
- En général, apprendre $\mathcal{C}$ est difficile.

Décomposition

- Apprendre un classifieur $\mathcal{C}_{ij}$ pour chaque paire de sorties $w_i, w_j \in \mathcal{W}$ prédisant laquelle des deux est la plus vraisemblable
- $M(M-1)/2$ classifieurs à apprendre sur des jeux de données "réduits"
- Combiner toutes prédictions pour obtenir résultat sur $\mathcal{W}$

Approche probabiliste

Classique

Le classifieur $\mathcal{C}$ fournit, pour $\mathbf{x}$, une probabilité $p_{\mathbf{x}}$ sur $\mathcal{W}$. On choisit

$$y = \arg\max_{w \in \mathcal{W}} p_{\mathbf{x}}(w)$$

Décomposition

Le classifieur $\mathcal{C}_{ij}$ fournit, pour $\mathbf{x}$ et chaque paire w_i, w_j, une probabilité $p_{\mathbf{x}}(w_i|\{w_i, w_j\}) = \alpha_{ij}$ sur $\{w_i, w_j\}$, qu'on transforme

$$p_{\mathbf{x}}(w_i|\{w_i, w_j\}) = \frac{p_{\mathbf{x}}(w_i)}{p_{\mathbf{x}}(w_i) + p_{\mathbf{x}}(w_j)} = \alpha_{ij} \rightarrow p(w_i) = \frac{\alpha_{ij}}{1 - \alpha_{ij}} p(w_j)$$

Approche probabiliste: problèmes

- pas/peu de prise en compte de la "compétence" des classifieurs
- problème "sur-contraint" souvent sans solutions: M variables et $M(M-1)/2 + M + 1$ contraintes

Un exemple simple

Soit $\mathcal{W} = \{w_1, w_2, w_3\}$ avec:

$$p(w_1|\{w_1, w_2\}) = 0.2, p(w_1|\{w_1, w_3\}) = 1/3, p(w_2|\{w_2, w_3\}) = 0.8.$$

transformées en

$$p(w_1) = 1/4 p(w_2),\ p(w_1) = 1/2 p(w_3),\ p(w_2) = 4p(w_3),$$

qui sont inconsistantes avec $p(w_1) + p(w_2) + p(w_3) = 1$.

En général, recherche de la meilleure solution approchée au sens d'une distance.

Approche probabiliste imprécise

Chaque classifieur $\mathcal{C}_{ij}$ peut fournir une évaluation imprécise

$$\alpha_{ij} \leq p(w_i|\{w_i, w_j\}) \leq \beta_{ij}$$

qui peut se transformer en deux contraintes

$$\frac{\alpha_{ij}}{1-\alpha_{ij}} p(w_j) \leq p(w_i) \text{ et } p(w_i) \leq \frac{\beta_{ij}}{1-\beta_{ij}} p(w_j)$$

en utilisant $p(w_i|\{w_i, w_j\}) = {}^{p(w_i)}/_{p(w_i)+p(w_j)}$. On a donc $M(M-1)+M+1$ contraintes (mais une seule d'égalité).

N.B.: on retrouve le cas classique quand $\alpha_{ij} = \beta_{ij}$

Approche probabiliste imprécise: exemple

Soit $\mathcal{W} = \{w_1, w_2, w_3\}$ et les estimations:

$$0.1 \leq p(w_1|\{w_1, w_2\}) \leq 1/3; \quad 1/6 \leq p(w_1|\{w_1, w_3\}) \leq 0.4;$$

$$2/3 \leq p(w_2|\{w_2, w_3\}) \leq 0.8.$$

Donnant

$$1/9 p(w_2) \leq p(w_1) \leq 1/2 p(w_2), \; 1/5 p(w_3) \leq p(w_1) \leq 2/3 p(w_3),$$

$$2p(w_3) \leq p(w_2) \leq 4p(w_3),$$

Les probabilités inférieures et supérieures pour chaque classe w_i sont

$$p(w_1) \in [0.067, 0.182], \; p(w_2) \in [0.545, 0.735], \; p(w_3) \in [0.176, 0.31].$$

et on choisit w_2

Gestion de l'inconsistence

problème

Les contraintes $\alpha_{ij} \leq p(w_i|\{w_i, w_j\}) \leq \beta_{ij}$ ajoutées aux contraintes de positivité et de somme unitaire peuvent aboutir à un programme linéaire sans solutions.

solution

- Idée: affaiblir les contraintes pour rendre le problème faisable
- Etant donné $\epsilon \in [0, 1]$ affaiblir contraintes en

$$(1 - \epsilon)\alpha_{ij} \leq p(w_i|\{w_i, w_j\}) \leq \epsilon + (1 - \epsilon)\beta_{ij}$$

- Choisir le plus petit ϵ tel que le problème devient faisable (ait des solutions) $\rightarrow$ utiliser les nouvelles contraintes pour choisir la classe

Décision et probabilités imprécises

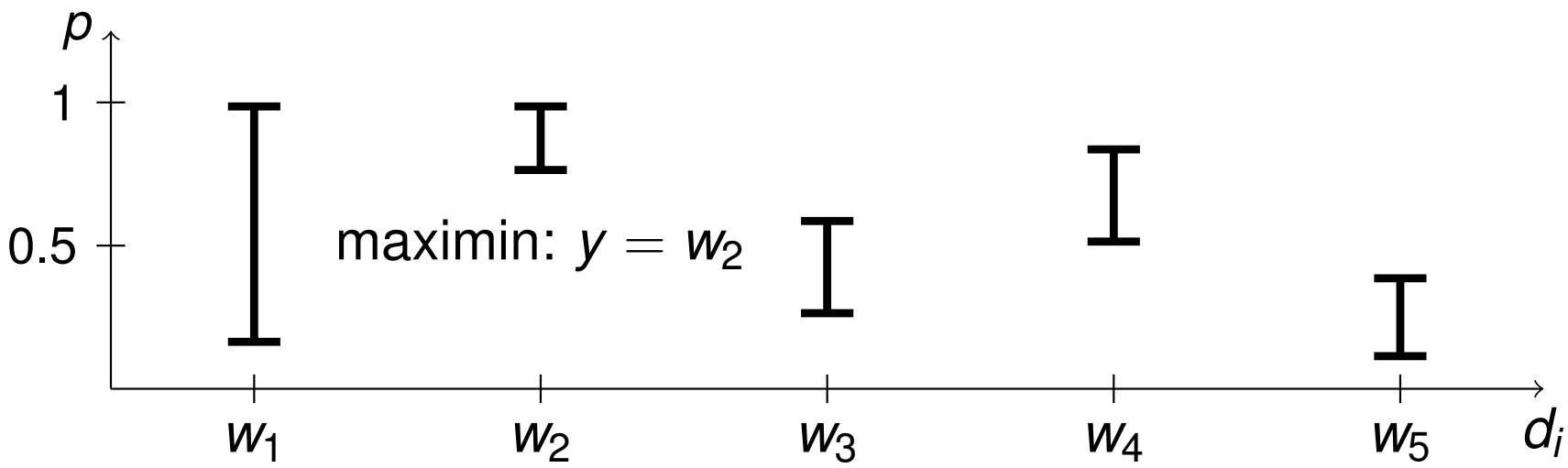
Deux critères "classiques"

- Précis: maximin

$$\rightarrow y := \arg\max_{w_i \in \mathcal{W}} \underline{P}(\{w_i\})$$

- Imprécis: dominance d'intervalles

$$\rightarrow Y := \{w_i \in \mathcal{W} |\ \nexists w_j \text{ s.t. } \overline{P}(\{w_i\}) \leq \underline{P}(\{w_j\})\}.$$



Décision et probabilités imprécises

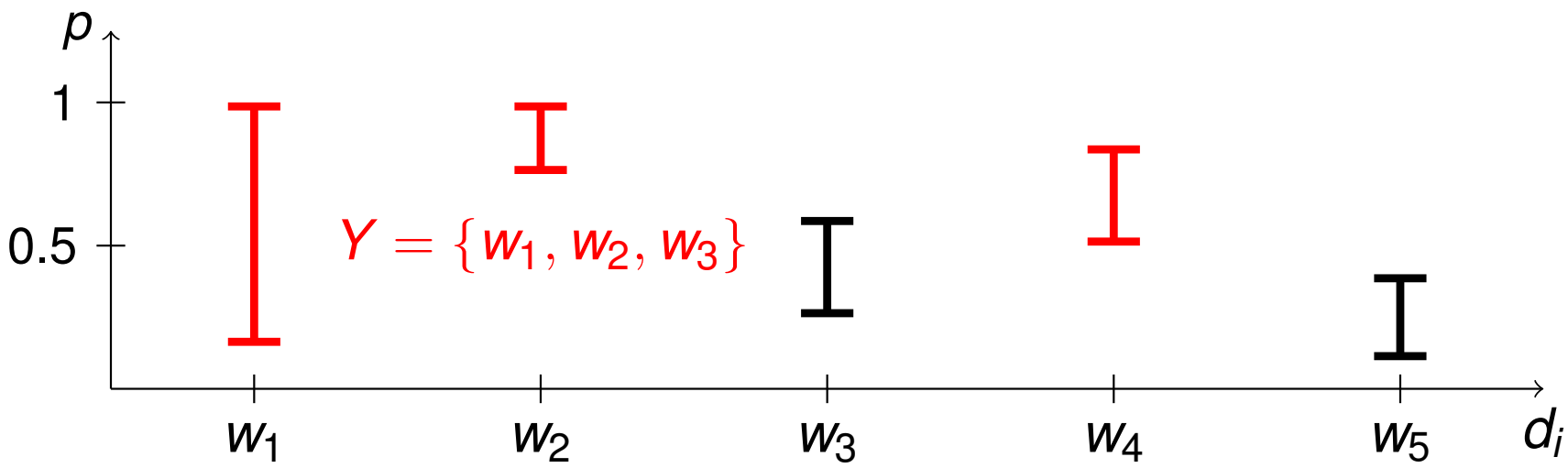
Deux critères "classiques"

- Précis: maximin

$$\rightarrow y := \arg\max_{w_i \in \mathcal{W}} \underline{P}(\{w_i\})$$

- Imprécis: dominance d'intervalles

$$\rightarrow Y := \{w_i \in \mathcal{W} |\ \nexists w_j \text{ s.t. } \overline{P}(\{w_i\}) \leq \underline{P}(\{w_j\})\}.$$



Mesurer la qualité d'un classifieur

Classiquement, on compare des sorties "test" aux sorties "prédites". Si la classification est imprécise, on utilise la notion de précision affaiblie en pénalisant l'imprécision:

$$\mathcal{D}_{acc} = \frac{\Delta}{|Y|},$$

avec Y les classes prédites et $\Delta = 1$ si la vraie classe y dans Y, $\Delta = 0$ autrement.

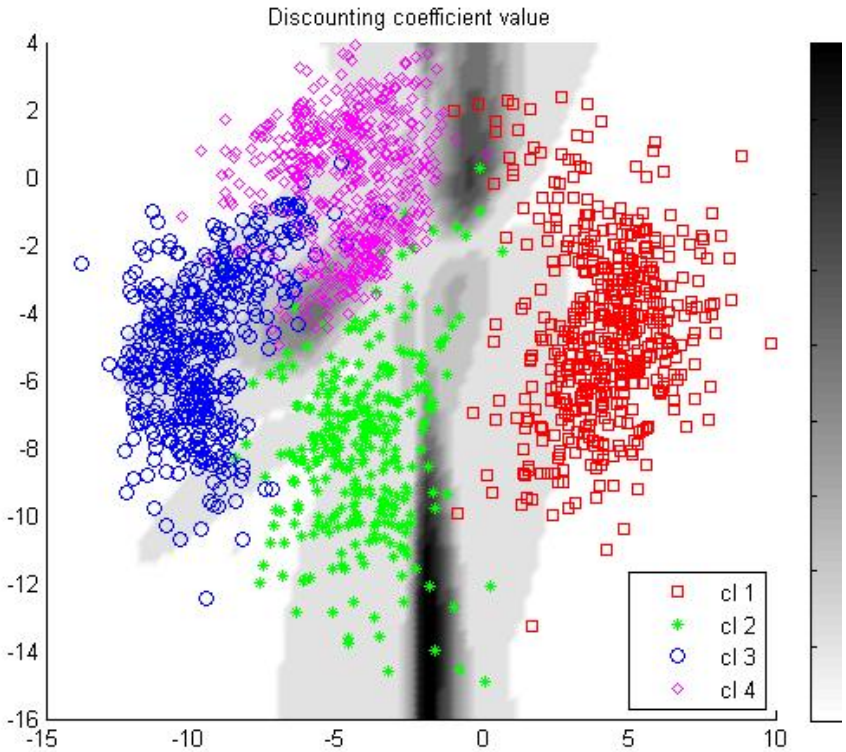
Dans le cas précis, $\mathcal{D}_{acc}$ est la précision habituelle

Supposons $\mathcal{W} = \{w_1, w_2, w_3\}$ et la réponse attendue est $w = w_1$, alors

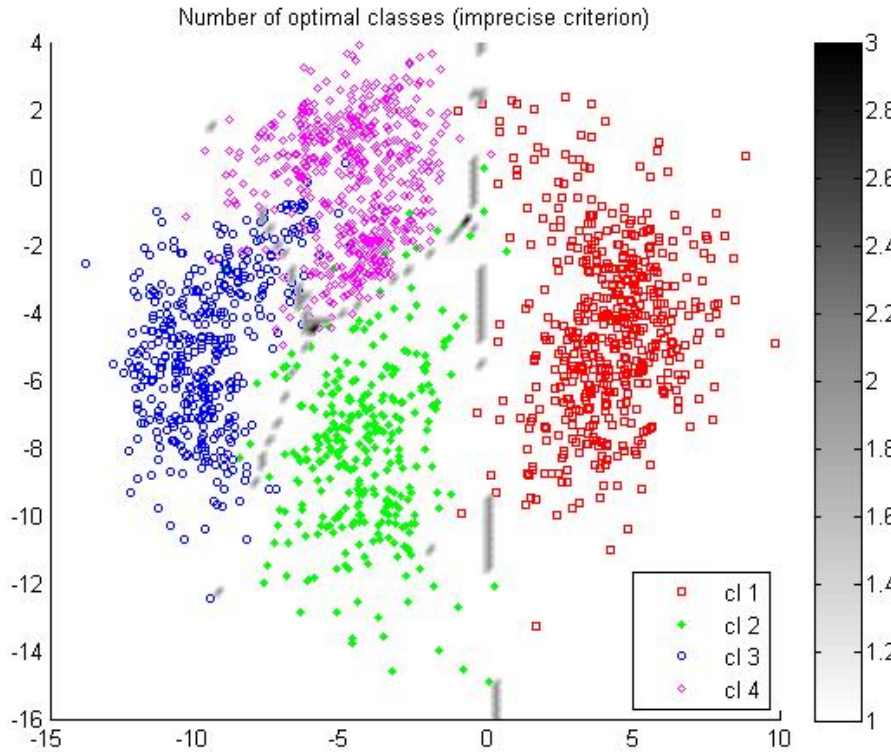
$\widehat{Y} = \{w_1\},$	$\widehat{Y} = \{w_1, w_3\},$	$\widehat{Y} = \{w_2, w_3\},$
$\mathcal{D}_{acc} = 1$	$\mathcal{D}_{acc} = 1/2$	$\mathcal{D}_{acc} = 0$

Un exemple jouet

$x \in \mathbb{R}^2$, 4 classes, binary classifiers: logistic regression



Levels of discounting



Imprecise decisions

Quelques tests

Table: erreur moyenne avec k plus proches voisins évidentiels.

dataset	single	PComp1	PComb2	maximin	$\mathcal{S}_{acc}$	$\mathcal{D}_{acc}$
synthetic	**4.67**	**4.80**	**4.80**	**4.80**	4.67	4.76
glass	**48.00**	**44.00**	**44.00**	**44.00**	42.67	45.42
pageblocks	**5.16**	**5.25**	**5.39**	**5.39**	4.75	9.87
satimage	**10.53**	**10.61**	**10.57**	**10.57**	9.72	10.20
segment	**8.57**	**9.12**	**9.12**	**9.12**	8.24	14.96
vowel	**39.39**	**38.96**	**39.18**	**39.18**	33.77	36.74
waveform	**16.44**	**16.47**	**16.47**	**16.47**	8.44	18.85
yeast	**37.37**	**37.88**	**37.88**	**37.88**	35.52	38.30

$\mathcal{S}_{acc}$: précision non-affaiblie avec décision imprécises
$\mathcal{D}_{acc}$: précision affaiblie

Conclusions

- Approche probabiliste imprécise est "compétitive" par rapport aux approches classiques

Intérêts des probabilités imprécises:

- résoud (en partie) le problème de la compétence: classifieur + imprécis sur zones de l'espace d'entrée contenant peu de données;
- décision imprécises plus robuste en présence d'ambiguité.

Perspectives

- Considérer des approches d'affaiblissement "locales" (actuellement, tous les classifieurs sont affaiblis de la même manière)
- Etendre l'approche au problème de label ranking en exploitant des structures d'ordre partiel
- Explorer d'autres règles de décision et leur effet sur les résultats
- Explorer en quoi les probabilités imprécises peuvent répondre à des problèmes "classiques" (compétence des classifieurs, erreur de position, etc. . .)