Participation de l'IRISA à DEFT

Apprentissage par *boosting* et *lazy-learning*

Christian Raymond, Vincent Claveau

IRISA, INSA, CNRS - Rennes

2 juillet 1815

Introduction

Participation de l'IRISA

- première participation
- participation aux deux tâches

Contexte

- *background* en apprentissage et en RI
- délais d'entraînement très courts

$\Rightarrow$ méthodes apprentissage et RI simples, non informées

Outline

Outline

Attributs de description utilisés pour la classification

1. le texte (mots sauf ponctuation) + une étiquette
   - étiquettes morpho-syntaxiques
   - + liste de connaissances (*i.e.* villes, pays, titres de noblesse, grade militaire...)
   - − tout ce qui n'est pas d'une catégorie précédente ou morpho-syntaxique (noms, adjectifs, verbes)
2. fréquence d'apparition dans le texte des catégories précédentes
3. idem que le premier, mais sans (1.3)→ figures de style ?

Outline

Arbre de décision : critère entropie

- critère automatique d'arrêt (MDL) :
  ↪pas de développement
- pas de critères vraiment discriminants
  ↪mauvaise généralisation
- performance tâche 1 $S \approx 0.15$

la résolution ne doit pas être envisagée par une classification brutale en année→pas de prise en compte de l'erreur relative (1810 est aussi différent de 1809 que 1900)

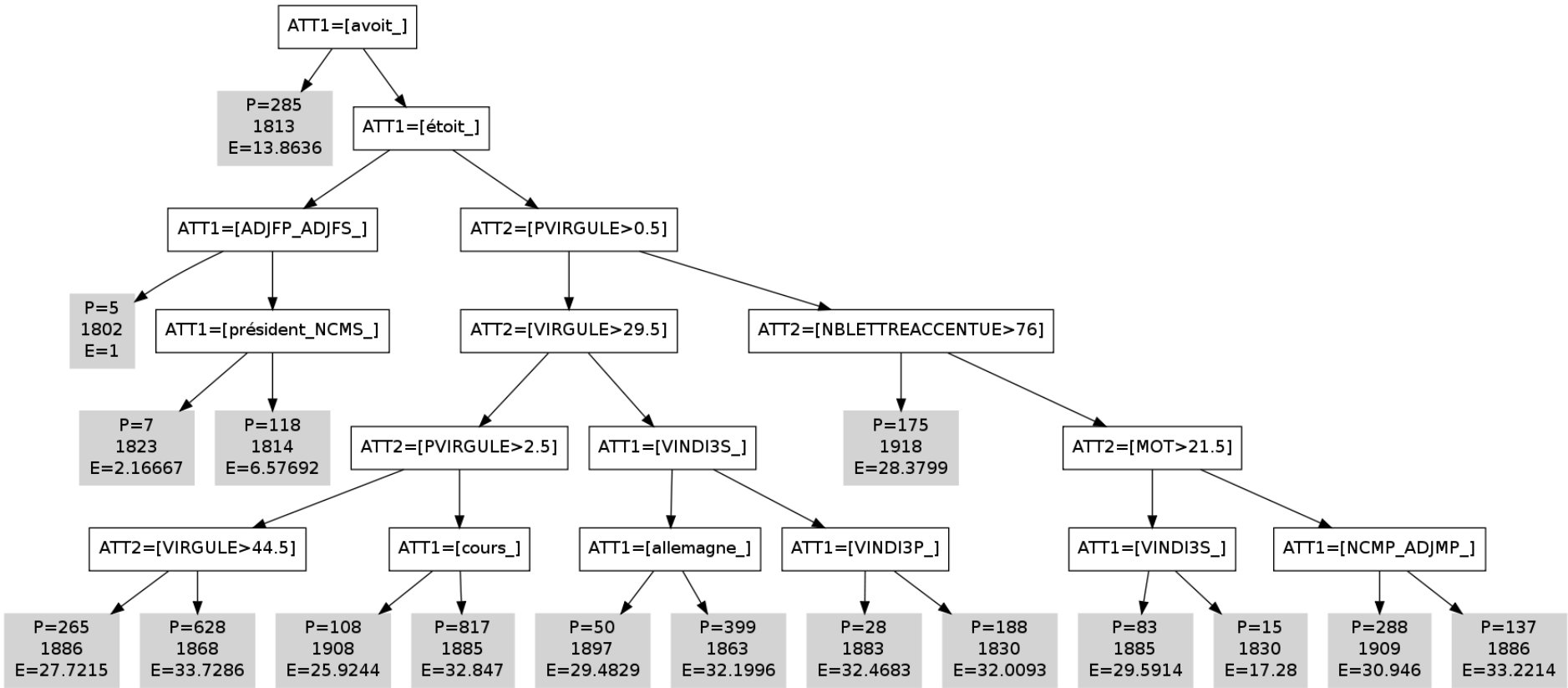
Outline

Arbre de décision : critère variance

- minimiser la somme des variances autour de l'année médiane dans chaque nœud
- prise en compte de l'erreur relative
- identifier des périodes temporelles plutôt que des années
- performance tâche 1 $S \approx 0.17$
- mesure d'évaluation plutôt que variance

pas d'indices performants pour décider si les documents sont antérieurs ou postérieurs à une période

Arbre de décision : critère variance



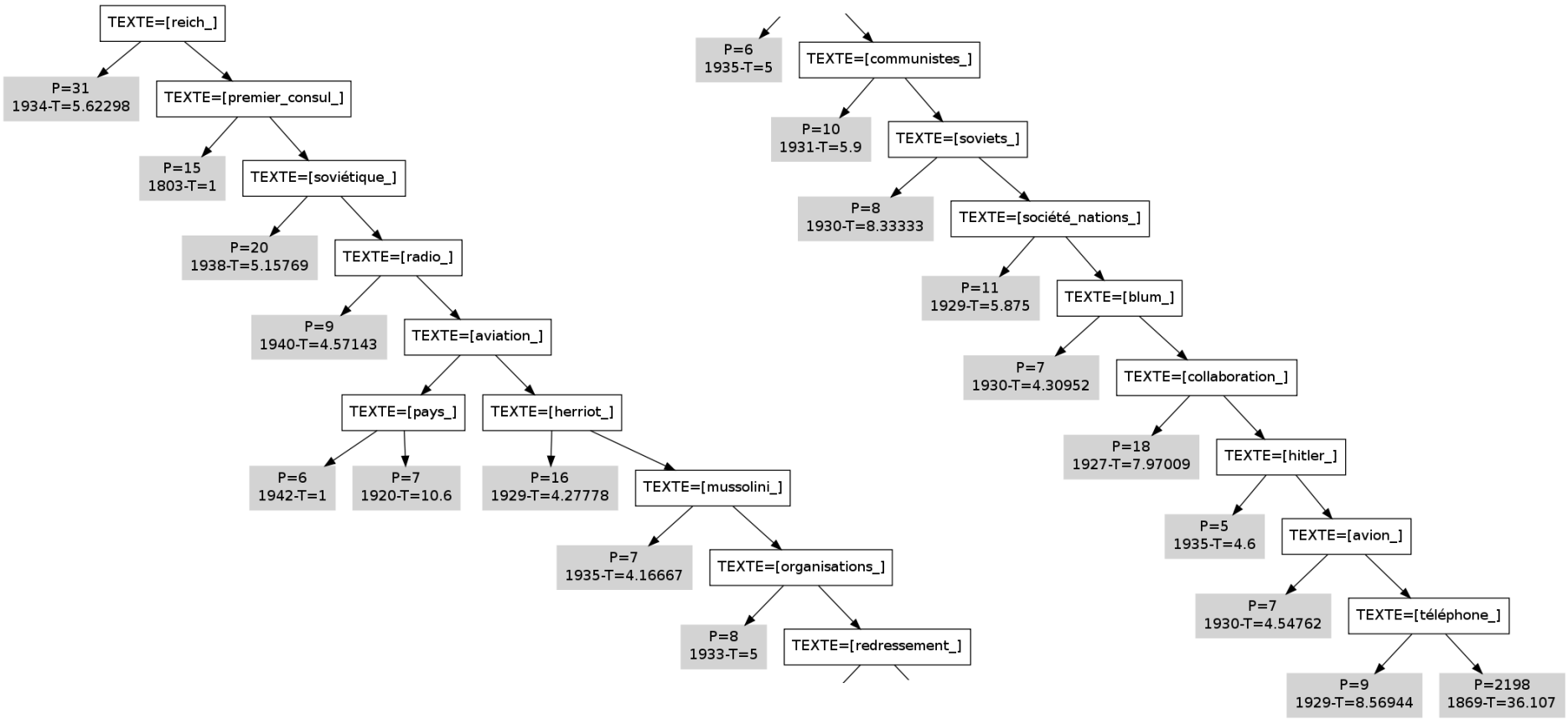
Outline

Arbre « peigne »

- approche précédente a du sens
- plutôt que discriminer antérieur/postérieur :
  ↪discriminer l'appartenance à une période ou pas
- minimiser la variance seulement dans le nœud gauche
- l'arbre trouve des indices caractéristiques de périodes temporelles

. . . mais la construction s'interrompt rapidement : seules certaines périodes sont facilement caractérisables

Arbre « peigne »



Outline

Boosting

- manque évident de caractéristiques fortement discriminantes :
  $\hookrightarrow$ approche de classification moins rigide
- combinaison de classifieurs faibles
  $\hookrightarrow$ Boosting
- AdaBoost.MH : approche brutale $S \approx 0.23$

Outline

Réduction multi-label binaire

- problème multi-classes $\rightarrow$ multi-label/binaire
  $\hookrightarrow$année $\rightarrow$ensemble des positionnements temporels/chaque année possible
- 1840 $\Rightarrow$ $POST_{1800}, POST_{1801}, \ldots, POST_{1839}$
- boosting multi-classes/multi-labels : AdaBoost.MH
- si un label est retrouvé, on vote pour l'ensemble des années postérieures sinon antérieures
- on choisi l'année qui rassemble le plus de vote
- amélioration très significative des résultats : $S \approx 0.33$ sur tâche 1

cette réduction binaire est toujours rigide : 1839 est tout autant antérieur à 1840 que 1800
approches à base de modèle peu performantes en l'état

Vote des classifieurs faibles en fonction de la présence ou l'absence du descripteur sélectionné

Tour	descripteur	présence	absence
1	étoit	[1813, 1944]	[1879, 1944]
2	VINDI3S	[1934, 1944]	[1802, 1937] 1942
5	avoit	[1802, 1944]	
8	reich	1944	[1802, 1943]
10	monsieur1 DETMS	[1826, 1944]	[1802, 1825]
12	#lettraccent>65.5	[1802, 1832]	[1833, 1944]
17	télégraphie	1930 [1932, 1944]	[1802, 1931]
18	allemagne	[1802, 1811] [1932, 1944]	[1813, 1931]
19	cit MOT	[1802, 1944]	1944
21	lit PREP DETMS	[1835, 1944]	[1802, 1834]
24	milieux	[1942, 1943]	1944 [1802, 1941]
25	président	[1935, 1944]	[1802, 1934]
28	société des nations	[1935, 1944]	[1802, 1934]

Outline

Vision de la tâche

Éléments à prendre en compte

- classification supervisée multilabel
  - structure des labels (proximité 1D)
- traitement sur des données bruitées par OCR
- variabilité intra-classe ?
  - 2 articles de la même année ne traitent pas du même sujet
  - bruit

$\Rightarrow$ approche robuste
$\Rightarrow$ approche simple et adaptée

À propos d'apprentissage

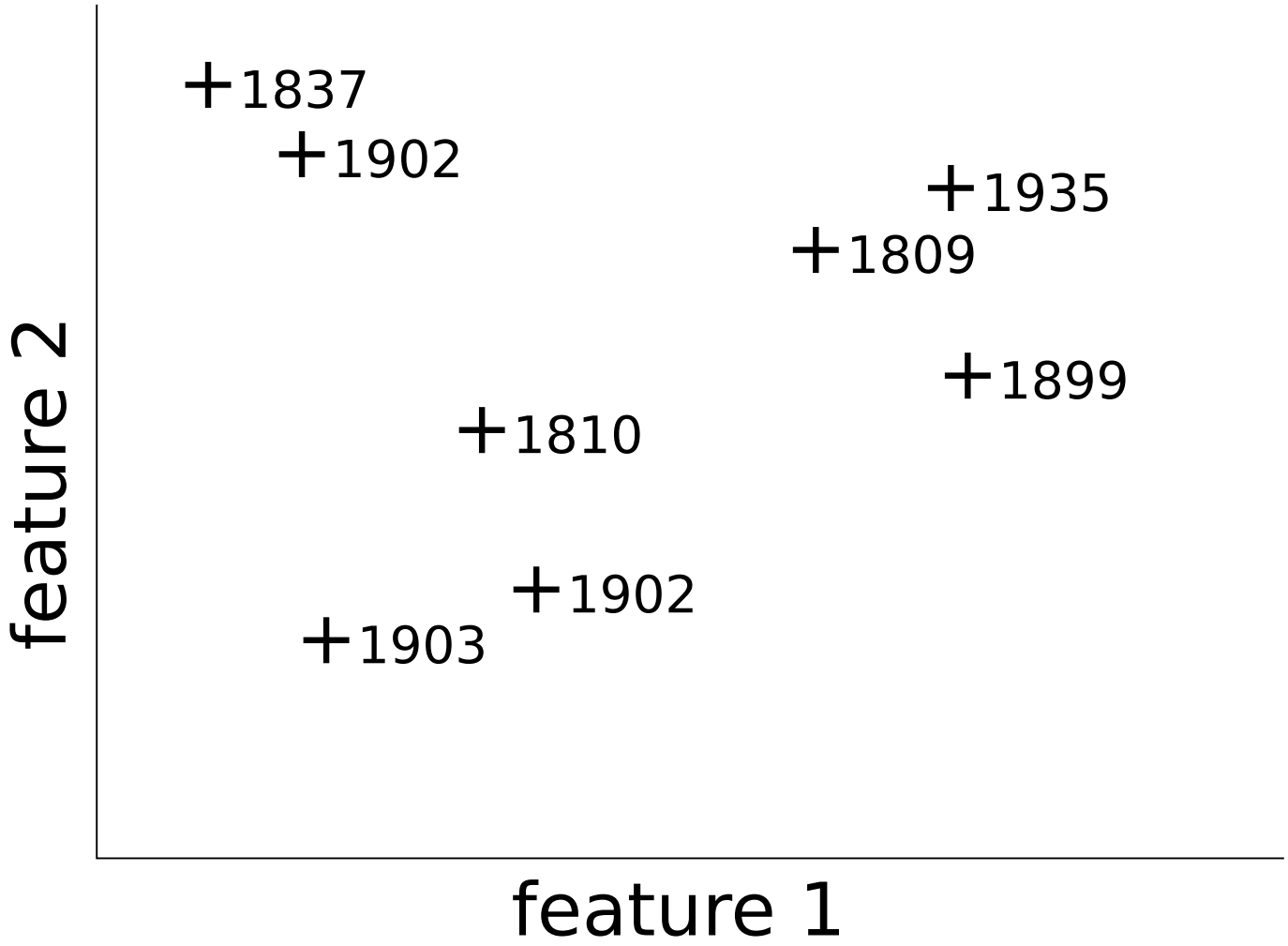
Espace de représentation vs. classifieur

- représentation sac-de-mot $\Rightarrow$ classes disjointes dans l'espace de représentation
- certains classifieurs ne sont pas du tout adaptés, d'autres sont inutilement complexes

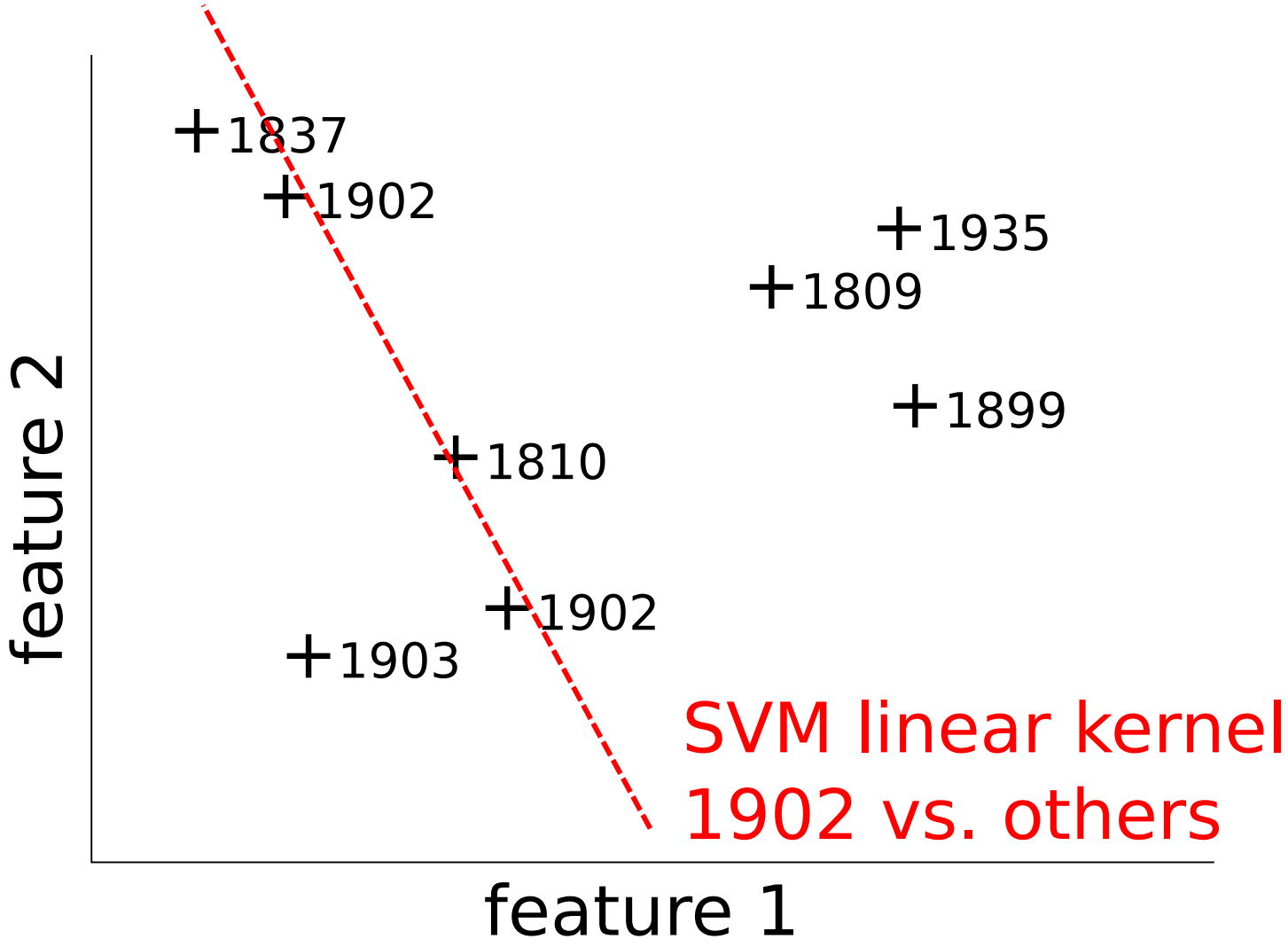
$\Rightarrow$ approche robuste : k-plus-proches voisins

  - mesure de similarité
  - procédure de vote des voisins

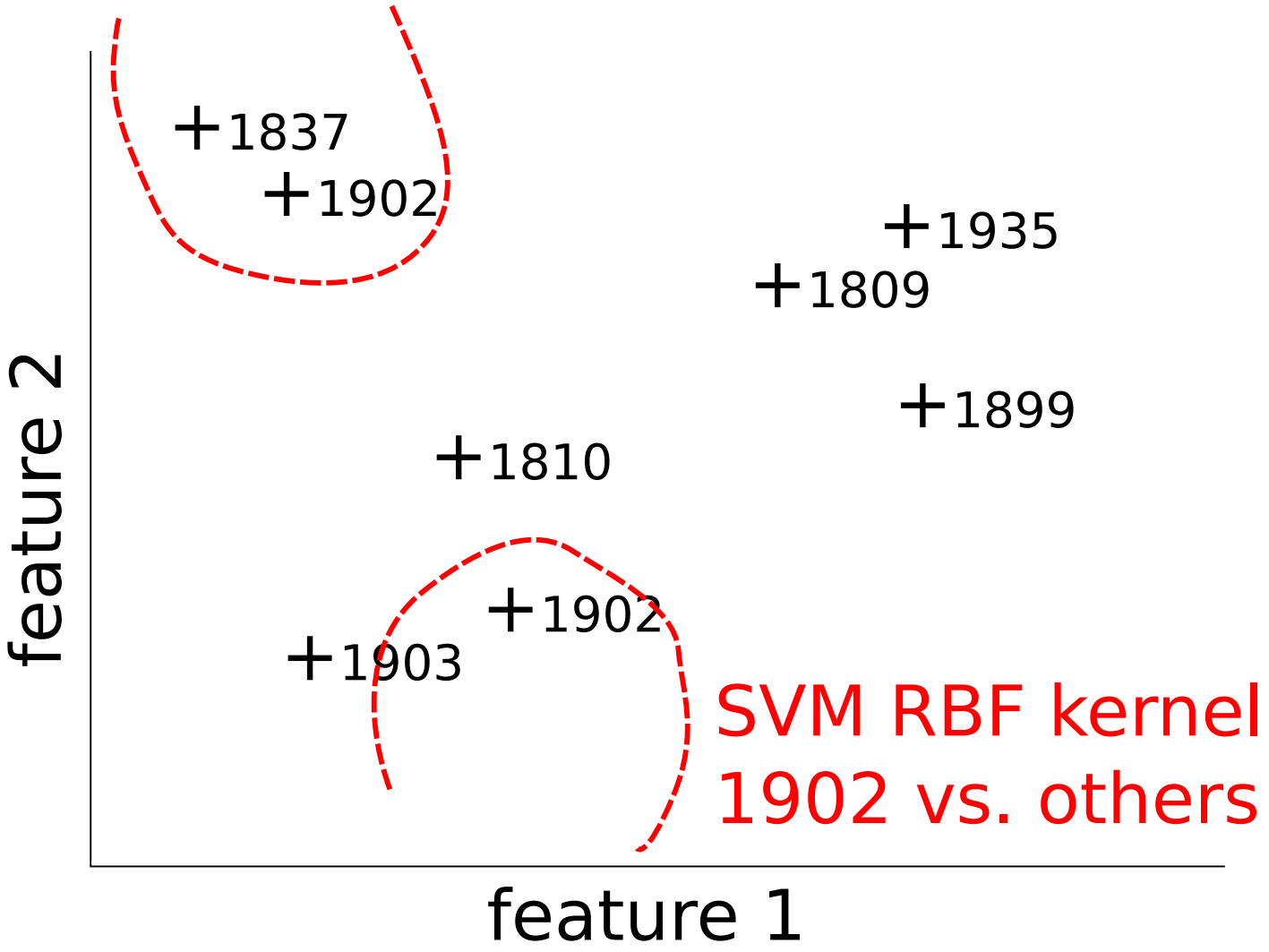
À propos d'apprentissage



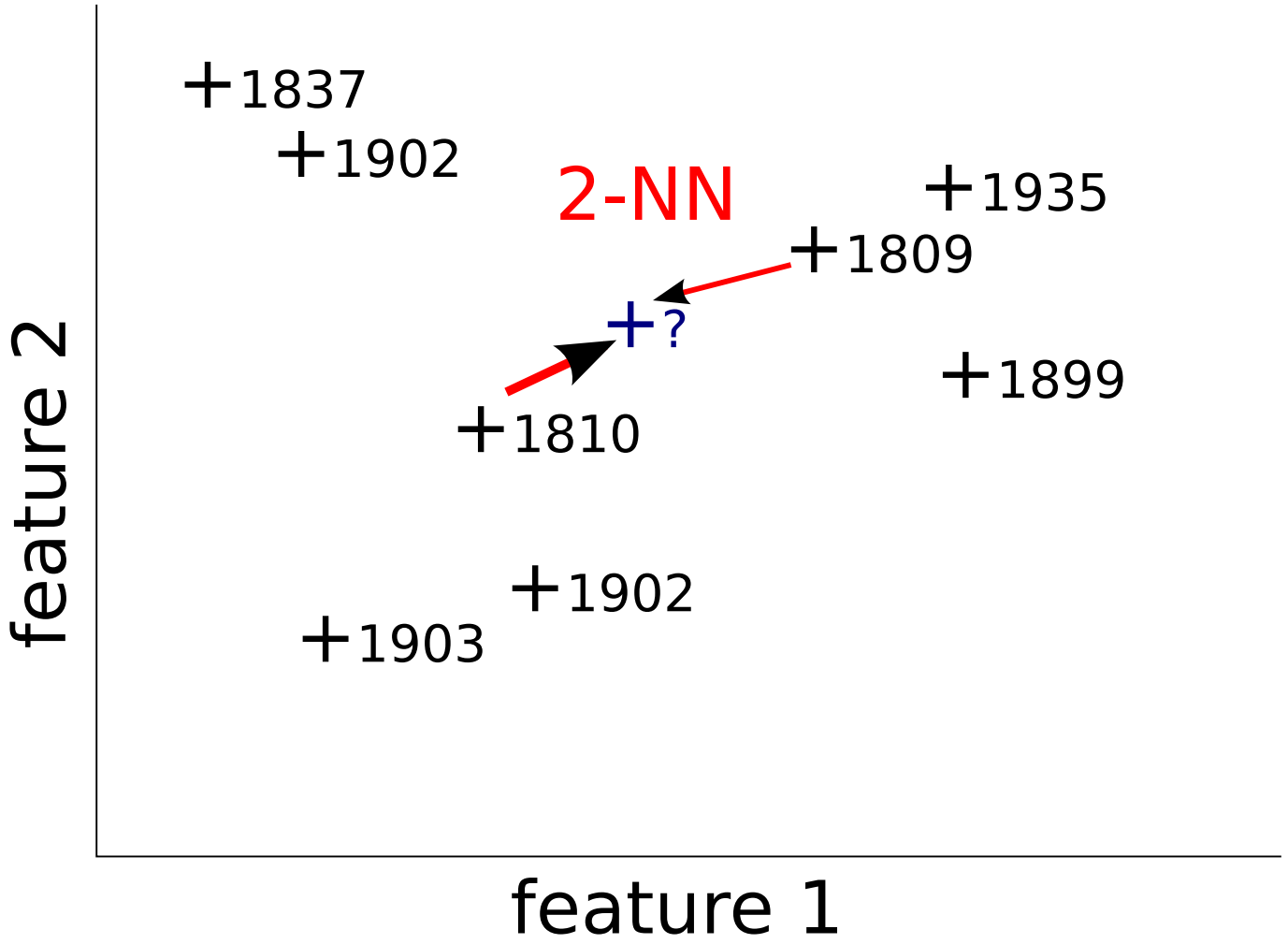
À propos d'apprentissage



À propos d'apprentissage



À propos d'apprentissage



K-plus proches voisins

Mesure de similarité

- proximité entre un article inconnu et les articles connus
- mesure standard en RI : Okapi-BM25 [Robertson 98]
  - $TF_{BM25}(t,d) = \frac{tf*(k_1+1)}{tf+k_1*(1-b+b*dl/dl_{avg})}$
  - $IDF_{BM25}(t) = \log \frac{N-df+0.5}{df+0.5}$
- autre mesure non-soumise : Hiemstra [Hiemstra 99]
  - modèle de langue pour RI

K-plus proches voisins

En pratique

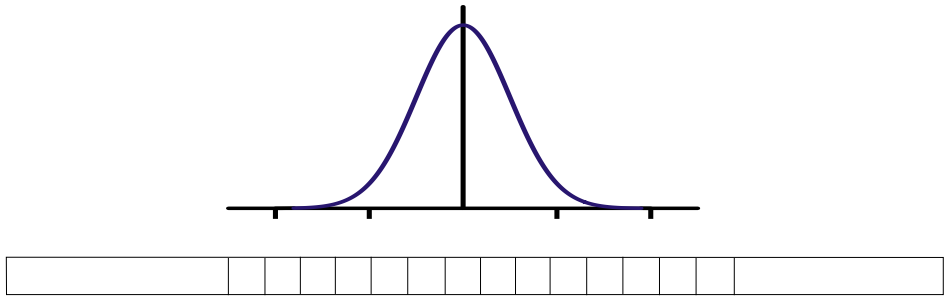
- aucun prétraitement sur l'OCR
- étiquetage (TreeTagger), on garde les lemmes des mots pleins
- discrimination des termes augmentée

$$sim(d_1, d_?) = \sum_t TF_{BM25}(t, d_?) * TF_{BM25}(t, d_1) * IDF_{BM25}(t)^3$$

K-plus proches voisins

Procédure de vote

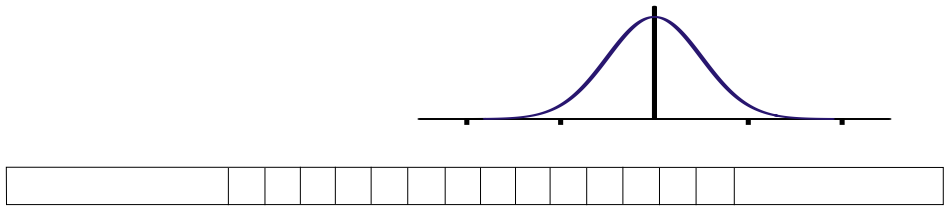
- 50 plus proches voisins $\Rightarrow$ 50 dates
  - vote pondéré par la proximité
- propagation aux années proches
  - utilisation de la même gaussienne que celle utilisée pour le score : l'année n reçoit $1 * sim(d_1, d_?)$, les années $n-1$ et $n+1$ reçoivent $0.969 * sim(d_1, d_?)$
- l'année proposée est celle qui a reçu le "plus grand poids" de vote



K-plus proches voisins

Procédure de vote

- 50 plus proches voisins $\Rightarrow$ 50 dates
  - vote pondéré par la proximité
- propagation aux années proches
  - utilisation de la même gaussienne que celle utilisée pour le score : l'année n reçoit $1 * sim(d_1, d_?)$, les années $n-1$ et $n+1$ reçoivent $0.969 * sim(d_1, d_?)$
- l'année proposée est celle qui a reçu le "plus grand poids" de vote



Résultats

Résultats quantitatifs

- track 1 (500 mots) : $S = 0.472$
- track 2 (300 mots) : $S = 0.430$
- conforme aux résultats par *leave-one-out* obtenus lors de la phase de développement

Autres considérations

- coût calculatoire faible : pas de modèle, pas d'entraînement
- calcul des similarités rapide : vecteur creux, fichiers inversés
- ajout de nouveaux exemples facile

Outline

Vision de la tâche

Tâche classique de recherche d'information

- similarité entre requête (résumé) et documents (articles)

Même problème, même solution

- recherche du 1 plus-proche voisin
- étiquetage avec TreeTagger, on ne garde que les mots pleins
- similarité calculée avec Okapi-BM25
- pas d'adjudication en cas de d'articles assignés à plusieurs résumés

Résultats

Résultats quantitatifs

- track 1 : 99.5%
- track 2 : 99%

Conclusions

À propos des tâches

- très différentes par leur niveau de difficulté
- très similaires (selon nous) par besoin de calculer des distances entre documents

Bilan

- bons résultats/classements dans les deux tâches
- emploi de méthodes standard de RI
- délai trop court pour mettre en œuvre des techniques innovantes