Large Margin Methods for Structured and Interdependent Output Variables

Tsochantaridis, Joachims, Hofmann and Altun
JMRL 2005

27 octobre 2005

Plan...

Introduction

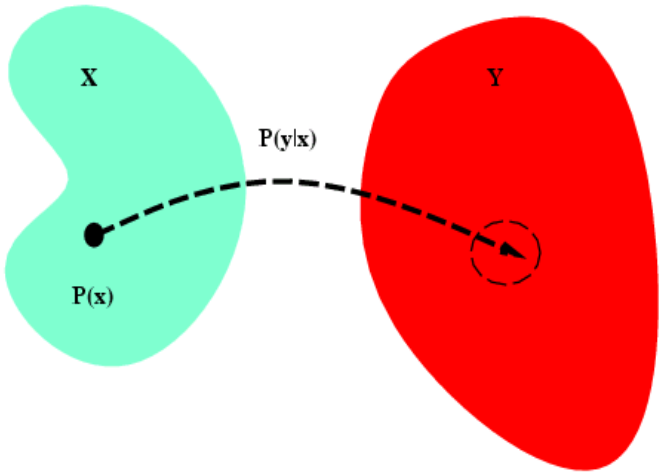
Pourquoi cet article ?

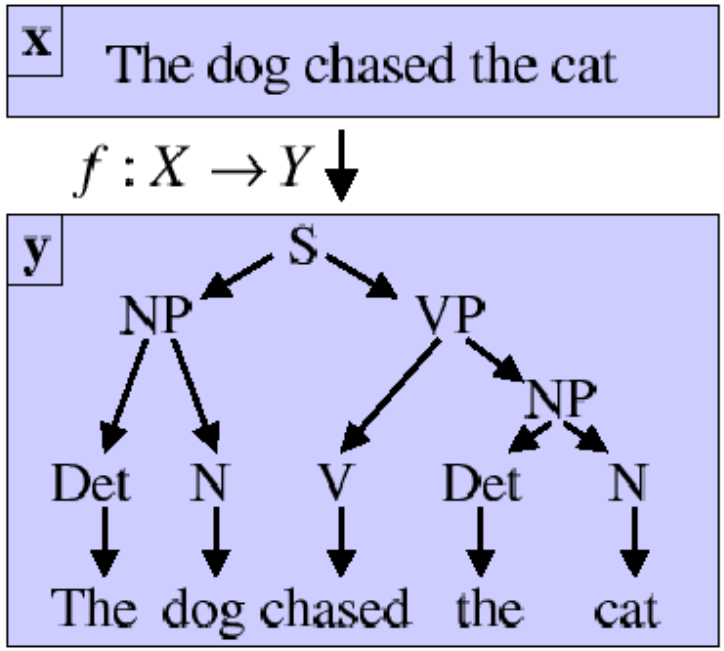
- Le best paper de ICML 2005 de Joachims en est un cas particulier

Objectif du papier

estimer la relation de dépendance entre 2 ensembles $\mathcal{X}$ et $\mathcal{Y}$, où $\mathcal{Y}$ est un ensemble discret d'objets structurés (graphes, séquences, arbres).

Exemple


X
Y
P(y|x)
P(x)


x
The dog chased the cat
$f : X \rightarrow Y$
y
S
NP
VP
NP
Det
N
V
Det
N
The
dog
chased
the
cat

Cadre

- A partir d'un ensemble d'exemples

$$(x_1, y_1) \cdots (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$$

  issus de l'échantillonage *i.i.d* d'une loi inconnue $P(X, Y)$
- On cherche une fonction $f : \mathcal{X} \longrightarrow \mathcal{Y}$
- On définit une fonction discriminante $F(\cdot, \cdot; w)$

$$F : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$$

  paramétrée par w
- Espace d'hypothèses f

$$f(x; w) = \arg \max_{y \in \mathcal{Y}} F(x, y; w)$$

Des détails sur $F(x, y; w)$

Interprétations

- F est une mesure de compatibilité entre x et y.
- $F(x, \cdot; w)$ est une fonction que l'on cherchera à maximiser de sorte que le maximum $y^\star$ corresponde à la sortie désirée pour x.

Forme

$$F(x, y; w) = \langle w, \Psi(x, y) \rangle$$

le choix de Ψ dépend du problème.

Fonction de coût et Minimisation du risque

On veut utiliser une fonction de coût plus générale qu'un coût $0-1$.

- fonction cout $\Delta : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}$
- $\Delta(y, y) = 0$ et $\Delta(y, y') > 0$ si $y \neq y'$
- Δ est borné.

Risque

$$R_P^{\Delta}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \Delta(y, f(x)) dP(x, y)$$

Risque empirique

$$R_S^{\Delta}(f) = \frac{1}{n} \sum_{i=1}^{n} \Delta(y_i, f(x_i))$$

Maximisation de la marge : cas séparable

On cherche à apprendre w par un paradigme de maximisation de marge

si il existe $f(\cdot; w)$ aboutissant à un risque empirique nul alors :

$$\forall i : \max_{y \in \mathcal{Y} \backslash y_i} \langle w, \Psi(x_i, y) \rangle \leq \langle w, \Psi(x_i, y_i) \rangle$$

$$\forall i, \forall y \in \mathcal{Y} \backslash y_i \quad \langle w, \Psi(x_i, y) \rangle \leq \langle w, \Psi(x_i, y_i) \rangle$$

on aboutit à $n|\mathcal{Y}| - n$ contraintes.

Maximisation de la marge : cas séparable (2)

Parmi tous les w qui satisfont les contraintes, on cherche celui qui maximise la différence entre les scores $\langle w, \Psi(x_i, y_i)\rangle$ et $\langle w, \Psi(x_i, y)\rangle$

Le pb QP associé

$$\begin{array}{ll} \min\limits_{w} & \frac{1}{2}\|w\|^2 \\ s.t & \forall i, \forall y \in \mathcal{Y}\backslash y_i \quad \langle w, \Psi(x_i, y_i) - \Psi(x_i, y)\rangle \geq 1 \end{array}$$

On note $\delta\Psi_i(y) = \Psi(x_i, y_i) - \Psi(x_i, y)$

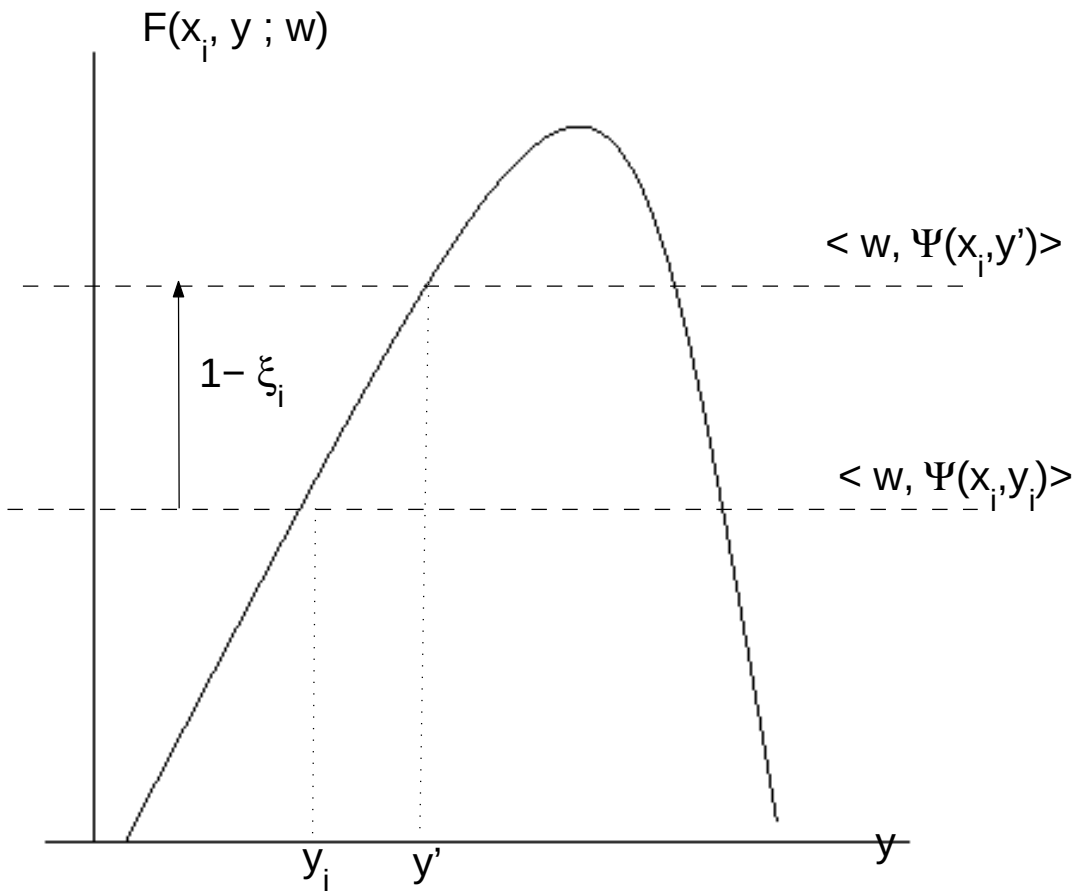
Maximisation de la marge : soft margin

Une variable de relachement unique pour chaque exemple i.

$$\begin{aligned} \min_{w,\xi} \quad & \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\ s.t \quad & \forall i, \forall y \in \mathcal{Y}\backslash y_i \quad \langle w, \delta\Psi_i(y)\rangle \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned}$$

$$\begin{aligned} \min_{w,\xi} \quad & \frac{1}{2}\|w\|^2 + \frac{C}{2n}\sum_{i=1}^{n}\xi_i^2 \\ s.t \quad & \forall i, \forall y \in \mathcal{Y}\backslash y_i \quad \langle w, \delta\Psi_i(y)\rangle \geq 1 - \xi_i \end{aligned}$$

Interpretation



Prise en compte de la fonction coût $\Delta(y_i, y)$

Idée : pondérer les variables de relachement ou modifier la marge en fonction de $\Delta(y_i, y)$

slack rescaling

$$\forall i, \forall y \in \mathcal{Y} \backslash y_i \quad \langle w, \delta\Psi_i(y) \rangle \geq 1 - \frac{\xi_i}{\Delta(y_i, y)}$$

margin rescaling

$$\forall i, \forall y \in \mathcal{Y} \backslash y_i \quad \langle w, \delta\Psi_i(y) \rangle \geq \Delta(y_i, y) - \xi_i$$

Propriétés

$$\frac{1}{n} \sum_i \xi_i^\star \geq R_S^\Delta(w)$$

Le problème dual

cas du slack rescaling

$$\max_{\alpha} \quad -\frac{1}{2}\sum_{i,y\neq y_i}\sum_{j,\bar{y}\neq y_i}\alpha_{(i,y)}\alpha_{(j,\bar{y})}\langle\delta\Psi_i(y),\delta\Psi_j(\bar{y})\rangle + \sum_{i,y\neq y_i}\alpha_{(i,y)}$$

$$st \quad \forall i \quad \sum_{y\neq y_i}\frac{\alpha_{(i,y)}}{\Delta(y_i,y)} \leq \frac{C}{n} \quad \alpha_{(i,y)} \geq 0$$

- $$w = \sum_i \sum_{y\neq y_i} \alpha_{(iy)}\delta\Psi_i(y)$$
- $n|\mathcal{Y}| - n$ contraintes
- Contraintes par blocs

Algorithme de résolution

1. Entrée : $(x_1, y_1), \cdots, (x_n, y_n)$, C, ϵ
2. $S_i = \emptyset$, pour tout i
3. répeter
4. pour $i = 1$ à n
5. préparer $H(y) = (1 - \langle w, \Psi(x_i, y) \rangle)\Delta(y_i, y)$
6. $\hat{y} = \arg\max_{y \in \mathcal{Y}} H(y)$
7. calculer $\xi_i = max(0, \max_{y \in S_i} H(y))$
8. si $H(\hat{y}) > \xi_i + \epsilon$
9. $S_i \leftarrow S_i \cup \hat{y}$
10. Résoudre le QP avec les contraintes $S = \cup S_i$
11. finsi
12. finpour
13. jusqu'à ce qu'aucun ensemble S_i ne change durant le pour

Commentaires

- Preuve de convergence
- Cardinal de S borné
- Difficulté :

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} H(y)$$

Application au Winner Take All

- $\mathcal{Y} = \{y_1, \cdots, y_K\}$, $w = [v_1', \cdots v_K']'$, $v_k \in \mathbb{R}^D$, $\Phi(x) \in \mathbb{R}^D$
- règle du WTA : $f(x) = \arg\max_y F(x, y; w)$ avec $F(x, y_k; w) = \langle v_k, \Phi(x) \rangle$

Modèle

- représentation d'un $y \in \mathcal{Y}$ à l'aide de vecteur de $\mathbb{R}^K$

$$\Lambda^c(y) \equiv [\delta(y_1, y), \cdots, \delta(y_K, y)]'$$

- Représentation jointe $\Psi(x, y)$

$$\Psi(x, y) \equiv \Phi(x) \otimes \Lambda^c(y)$$

$$F(x, y; w) = \langle w, \Psi(x, y) \rangle$$

Extensions à des cas plus complexes

Modélisation

- Si $\Psi = \Phi \otimes \Lambda$ alors

$$\begin{aligned} \langle \Psi(x,y), \Psi(x',y') \rangle &= \langle \Phi(x), \Phi(x') \rangle \cdot \langle \Lambda(y), \Lambda(y') \rangle \\ &= K_\Phi(x,x') \cdot K_\Lambda(y,y') \end{aligned}$$

- $F(x,y;w) = \sum_{r=1}^{R} \lambda_r(y) \langle v_r, \Phi(x), \quad \Lambda(y) \in \mathbb{R}^R$

- Classification avec taxonomie
- Apprentissage tenant compte de la similarité des classes

Conclusions

- Maximisation de marge pour estimation de dépendances sur données de sortie structurées.
- Modélisation du problème à l'aide d'une fonction discriminante linéaire dans un espace de caractéristiques $\Psi(x, y)$
- Méthodes à noyaux pour l'apprentissage de dépendances complexes.