Apprentissage, réseaux de neurones et modèles graphiques (RCP209)
Structured Prediction

Nicolas Thome
Prenom.Nom@cnam.fr
http://cedric.cnam.fr/vertigo/Cours/ml2/

Département Informatique
Conservatoire National des Arts et Métiers (Cnam)

Outline

1 Instanciations

2 Prédiction Structurée avec variable latentes (LSSVM)

SSVM : Instanciation

Rappel : ingrédient d'un problème d'apprentissage structuré

- Définition de l'entrée $\mathbf{x} \in \mathcal{X}$
- Définition de la sortie structurée $\mathbf{y} \in \mathcal{Y}$
- Pour la prédiction :
  - $\Psi(\mathbf{x}, \mathbf{y})$
  - Résolution du problème d'inférence : $\hat{\mathbf{y}}(\mathbf{x}, \mathbf{w}) = \underset{\mathbf{y} \in \mathcal{Y}}{arg\ max} \langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{y}) \rangle$
- Pour l'apprentissage du modèle :
  - $\Psi(\mathbf{x}, \mathbf{y})$
  - $\Delta(\mathbf{y}_i^*, \mathbf{y})$
  - Résolution du problème de "Loss-Augmented Inference" (LAI) : $\tilde{\mathbf{y}}_i = \underset{\mathbf{y} \in \mathcal{Y}}{arg\ max} \left[\Delta(\mathbf{y}_i^*, \mathbf{y}) + \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle \right]$

SSVM : Instanciation

Exemples classiques d'instantiations SSVM en vision

- Classification Multi-classes
- Ranking
- Classification hiérarchique
- Détection d'objets
- Estimation de pose
- Segmentation d'images (segmentation sémantique)
- Prédiction de séquences

SSVM : instantiation

Classification multi-classes

- $\mathcal{Y} = \{1, 2, \dots, K\}, \quad \Delta(y, y') = \begin{cases} 1 & \text{for } y \neq y' \\ 0 & \text{otherwise.} \end{cases}$
- $\varphi(x, y) = \Big([\![y = 1]\!]\Phi(x),\ [\![y = 2]\!]\Phi(x),\ \dots,\ [\![y = K]\!]\Phi(x) \Big)$
  $= \Phi(x) e_y^\top$ with e_y=y-th unit vector

Solve:

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n} \xi^i$$

subject to, for $i = 1, \dots, n,$

$$\langle w, \varphi(x^i, y^i)\rangle - \langle w, \varphi(x^i, y)\rangle \geq 1 - \xi^i \quad \text{for all } y \in \mathcal{Y} \setminus \{y^i\}.$$

Classification: MAP $\quad f(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \ \langle w, \varphi(x, y)\rangle$

Crammer-Singer Multiclass SVM

SSVM : instantiation classification multi-classes

Connections

- Cas particulier de cette instanciation avec 2 classes ?
- Lien avec la classification binaire ?

Inference et "loss-augmented" inference

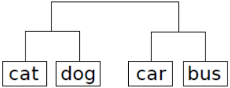
- Classification multi-classes et hiérarchique
  - $|\mathcal{Y}| = K$ où K est le nombre de classes
  - Dans ce cas, il est envisageable de calculer l'inférence où le loss-augmented inference exhaustivement

SSVM : instantiation

Classification hiérarchique

Hierarchical Multiclass Classification

Loss function can reflect hierarchy:



$$\Delta(y, y') := \frac{1}{2}(\text{distance in tree})$$

$$\Delta(\text{cat}, \text{cat}) = 0, \quad \Delta(\text{cat}, \text{dog}) = 1, \quad \Delta(\text{cat}, \text{bus}) = 2, \quad \textit{etc.}$$

Solve:

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n} \xi^i$$

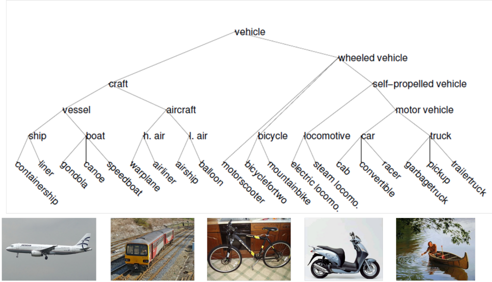
subject to, for $i = 1, \dots, n$,

$$\langle w, \varphi(x^i, y^i)\rangle - \langle w, \varphi(x^i, y)\rangle \geq \Delta(y^i, y) - \xi^i \quad \text{for all } y \in \mathcal{Y} \setminus \{y^i\}.$$

SSVM : instantiation

Classification hiérarchique

- Comment définir $\Delta(\mathbf{y}_i, \mathbf{y})$?
- Wordnet $\Rightarrow$ distance sémantique entre classes



SSVM : instantiation

Ordonnancement (Ranking)

- Entrée $\mathbf{x} \in \mathcal{X}$: ensemble d'éléments : $\mathbf{x} = (d_1, ... d_n)$, *e.g.* d_i image
  - On suppose que pour chaque $d_i \Rightarrow \phi(d_i) \in \mathbb{R}^d$
- Sortie structurée $\mathbf{y} \in \mathcal{Y}$: ordonnancement de ces éléments / requête
- **requête: classification binaire** $\Rightarrow$ $\mathbf{y} \sim$ ordre de pertinence / classe
  - Chaque elt $d_i \in \oplus$ (pertinent) ou $d_i \in \ominus$ (non pertinent)
- Sortie $\mathbf{y}$: liste $\mathbf{y} = (y_1, ... y_n)$, y_i indice de l'exemple placé au rang i
  - Taille $|\mathcal{Y}|$ de l'espace $\mathcal{Y}$?
  - **Que vaut $\mathbf{y}^*$ donné par la supervision ?** Unicité ?
- Ordonnancement $\mathbf{y}$ peut être représenté par une matrice $\mathbf{Y}$ tq :

$$y_{ij} = \begin{cases} +1 & \text{si } d_i \prec_y d_j \ (d_i \text{ est classé avant } d_j \text{ dans la liste ordonnée}) \\ -1 & \text{si } d_i \succ_y d_j \ (d_i \text{ est classé après } d_j) \end{cases}$$

  - Taille $|\mathcal{Y}|$ de l'espace $\mathcal{Y}$?
  - **Que vaut $\mathbf{Y}^*$ donné par la supervision ?** Unicité ?
- **Modèle graphique de la sortie y pour la ranking ?**

Ordonnancement (Ranking)

Prédiction

- Ranking feature map:

$$\Psi(\mathbf{x},\mathbf{y}) = \frac{1}{N_+ \cdot N_-} \sum_{d_i \in \oplus} \sum_{d_j \in \ominus} y_{ij} \left[\phi(\boldsymbol{d_i}) - \phi(\boldsymbol{d_j})\right]$$

- Score pour chaque sortie :

$$\langle \mathbf{w}; \Psi(\mathbf{x},\mathbf{y})\rangle = \frac{1}{N_+ \cdot N_-} \sum_{d_i \in \oplus} \sum_{d_j \in \ominus} y_{ij} \langle \mathbf{w}; \left[\phi(\boldsymbol{d_i}) - \phi(\boldsymbol{d_j})\right]\rangle$$

- **Inférence ??**, *i.e.* prédiction avec le modèle ($\mathbf{w}$ fixé) :

$$\hat{\mathbf{y}}(\mathbf{x},\mathbf{w}) = \underset{\mathbf{y}\in\mathcal{Y}}{arg\ max} \langle \mathbf{w}, \Psi(\mathbf{x},\mathbf{y})\rangle$$

$\Rightarrow$ tri des exemples / $\langle \mathbf{w}; \phi(\boldsymbol{d_i})\rangle$

Ordonnancement (Ranking)

Apprentissage

- **Un seul exemple d'apprentissage x !**
- Un (très gd) ensemble de configuration $\mathbf{y} \in \mathcal{Y}$
- Score pour chaque sortie :

$$\langle \mathbf{w}; \Psi(\mathbf{x}, \mathbf{y}) \rangle = \frac{1}{N_+ \cdot N_-} \sum_{d_i \in \oplus} \sum_{d_j \in \ominus} y_{ij} \langle \mathbf{w}; [\phi(d_i) - \phi(d_j)] \rangle$$

- **Score pour la sortie $\mathbf{y}^*$ donné par la supervision ?**

$$\langle \mathbf{w}; \Psi(\mathbf{x}, \mathbf{y}^*) \rangle = \frac{1}{N_+ \cdot N_-} \sum_{d_i \in \oplus} \sum_{d_j \in \ominus} \langle \mathbf{w}; [\phi(d_i) - \phi(d_j)] \rangle$$

$$\Rightarrow \langle \mathbf{w}; \Psi(\mathbf{x}, \mathbf{y}^*) \rangle \uparrow \text{ si } \langle \mathbf{w}; [\phi(d_i) - \phi(d_j)] \rangle \uparrow$$

Ordonnancement (Ranking)

Apprentissage

- Rappel : résolution du problème de "Loss-Augmented Inference" (LAI) :

$$\tilde{\mathbf{y}}_i = \underset{\mathbf{y} \in \mathcal{Y}}{arg\ max} \left[\Delta(\mathbf{y}_i^*, \mathbf{y}) + \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle \right] \quad (1)$$

- Optimisation $\mathbf{w}$ par descente de gradient (fonction convexe) :

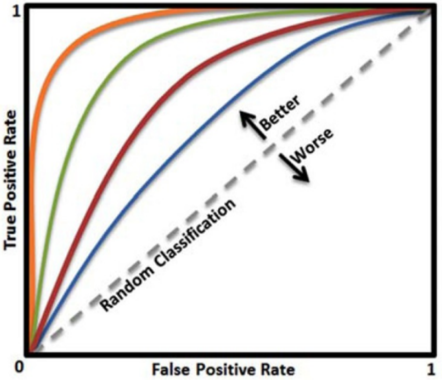
$$\frac{\partial \mathcal{P}}{\partial \mathbf{w}} = \lambda \mathbf{w} + \left(\Psi(\mathbf{x_i}, \tilde{\mathbf{y}}_\mathbf{i}) - \Psi(\mathbf{x_i}, \mathbf{y_i^*}) \right)$$

- Solution au problème LAI, Eq. (1) : dépend de la métrique $\Delta(\mathbf{y}_i^*, \mathbf{y})$!

Ordonnancement (Ranking) : Métriques

Area Under Curve (AUC)

- Courbe ROC : Vrai Positifs (TP) *vs* Faux Positifs (FP)
  - Pratique: tri $\langle \mathbf{w}; \phi(\boldsymbol{d_i}) \rangle$, seuil variable s à chaque elt, début $s = +\infty$
  - Vrai Positifs (TP) : nombres éléments *score* $> s$ / nombre $\oplus$
  - Faux Positifs (FP) : nombres éléments *score* $> s$ / nombre $\oplus$



- **AUC : aire sous la courbe ROC**

Ordonnancement (Ranking) : AUC

Area Under Curve (AUC)

- **AUC : aire sous la courbe ROC**
- $\Delta_{AUC}(y, y^*) = 1 - AUC$, on peut montrer que :

$$\Delta_{AUC}(y, y^*) = \sum_{i \in \oplus} \sum_{j \in \ominus} \frac{(1 - y_{ij})}{2} \qquad (2)$$

- Interprétation: Δ_{AUC} = nombre de paires échangées ($i.e.$ $y_{ij} \neq y^*_{ij}$)
- Ne prend pas en compte la position dans la liste de l'échange
- Pas très adapté pour la Recherche d'Informations (RI)
  - Classe $\oplus$ (requête) et $\ominus$ pas des rôles symétriques
  - Objectif : retrouver des éléments pertinents au sommet de la liste

Ordonnancement (Ranking) : AUC

Area Under Curve (AUC) et "Loss-Augmented Inference" (LAI)

- Δ_{AUC} se décompose linéairement / paires $\oplus/\ominus$
- Résolution du pb LAI Eq. (1) avec Δ_{AUC} Eq. (2) se ramène à :

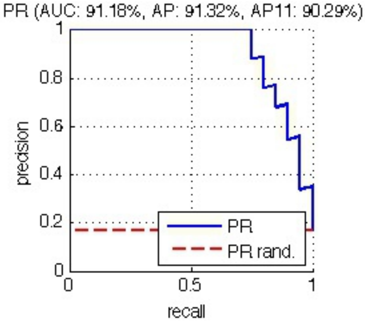
$$\tilde{\mathbf{y}} = \underset{\mathbf{y} \in \mathcal{Y}}{arg\ max} \sum_{i \in \oplus} \sum_{j \in \ominus} y_{ij} ((\langle w; \phi(x_i) \rangle - 0.25) - ((\langle w; \phi(x_j) \rangle + 0.25)$$

- $\tilde{y_{ij}}$ du même signe $((\langle w; \phi(x_i) \rangle - 0.25) - ((\langle w; \phi(x_j) \rangle + 0.25)$
  - Calcul $\oplus$ / $((\langle w; \phi(x_i) \rangle - 0.25)$
  - Calcul $\ominus$ / $((\langle w; \phi(x_j) \rangle + 0.25)$
  - Tri des score résultant sur les $N = N_+ + N_-$ elts
- LAI $\sim$ Inference

Ordonnancement (Ranking) : Métriques

Precision Recall Curve and Average Precision (AP)

- Courbe Précision-Rappel : Précision *vs* Rappel
- Pratique: tri $\langle \mathbf{w}; \phi(\boldsymbol{d_i}) \rangle$, seuil variable s à chaque elt, début $s = +\infty$
  - Précision : nombre d'éléments pertinents / nombre d'éléments total qui ont un score supérieur à s
  - Rappel : nombre d'éléments pertinents avec score supérieur à s / nombre d'éléments pertinents total (=1-FP)



- **Average Precision (AP) : aire sous la courbe Précision-Rappel**

Ordonnancement (Ranking) : AP

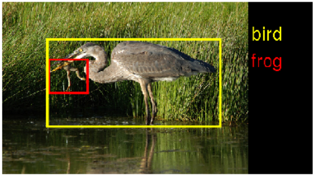
Average Precision (AP)

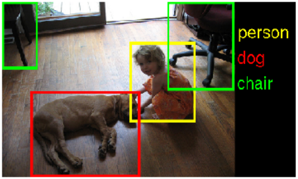
- **Average Precision (AP) : aire sous la courbe Précision-Rappel**
- $\Delta_{AP}(y, y^*) = 1 - AP$
- Δ_{AP} ne se décompose pas linéairement / paires $\oplus/\ominus$!! $\neq \Delta_{AUC}$
- Résolution du pb LAI Eq. (1) avec Δ_{AP} : plus complexe que l'inférence !
- Algorithme optimal glouton [YFRJ07] (voir TP) :
  - Tri des $\oplus$ / $\langle \mathbf{w}; \phi(\boldsymbol{d_i}) \rangle$
  - Tri des $\ominus$ / $\langle \mathbf{w}; \phi(\boldsymbol{d_j}) \rangle$
  - Déterminer l'insertion optimale de chacun des exemples $\ominus$ dans la liste des $\oplus$: pour le $j^{ème}$ $\ominus$ trié, $i^* = \underset{i}{arg\,max}\ \delta_j(i)$

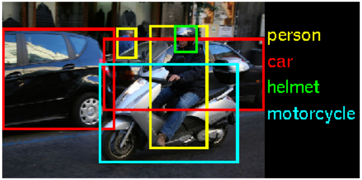
$$\delta_j(i) = \frac{1}{N_+} \sum_{k=i}^{N_+} \left(\frac{j}{k+j} - \frac{j-1}{k+j-1} \right) - \frac{2(s_k^p - s_j^n)}{N_-}$$

SSVM : instantiation

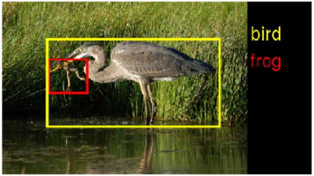
Détection d'objet : fonction de prédiction ⇒ $\Psi(x, y) = ?$

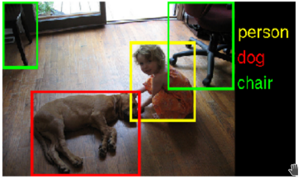






Détection d'objet : apprentissage $\Rightarrow \Delta(\mathbf{y}_i, \mathbf{y}) = ?$

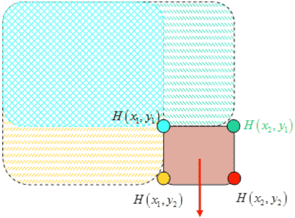




SSVM : instantiation pour la Détection d'objet

Inference et "loss-augmented" inference

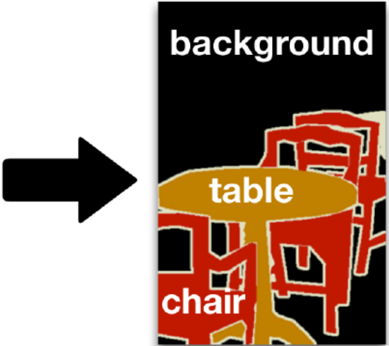
- Détection d'objet : $|\mathcal{Y}|$: nombre de régions possibles
  - Approche classique : fenêtre glissante $\Rightarrow$ très grand nb de régions candidates
  - Car on parcours à différentes positions, échelles, ratio (voire rotation)
- Pour accélérer le calcul : "histogrammes intégrales"
  - Permet de calculer l'hostogramme (BoW) en tps contsant



- Mais nécessite tjs l'évaluation de tous les produits scalaires
- autres méthodes : couper l'espace de recherche
  - stratégie branch and bound [BL08]

SSVM : instantiation

Segmentation sémantique d'images : prédiction $\Rightarrow \Psi(x, y) = ?$



SSVM : instantiation

Segmentation sémantique d'images : $\Delta(\mathbf{y}_i, \mathbf{y})$ = ?

- Données d'apprentissage : $\mathcal{X}$ image, $\mathcal{Y}$ masque de segmentation
- Loss utilisé : Hamming loss $\Delta(\mathbf{y}_i, \mathbf{y}) = \frac{1}{|V|} I(\mathbf{y}_i \neq \mathbf{y})$
    - Compte le ratio de pixels mal étiquetés

$t = 1$: $\hat{y} =$  $\phi(y^n) - \phi(\hat{y})$: black $+$, white $+$, green $-$, blue $-$, gray $-$

$t = 2$: $\hat{y} =$  $\phi(y^n) - \phi(\hat{y})$: black $+$, white $+$, green $=$, blue $=$, gray $-$

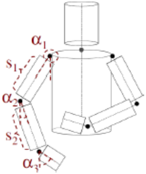
$t = 3$: $\hat{y} =$  $\phi(y^n) - \phi(\hat{y})$: black $=$, white $=$, green $-$, blue $-$, gray $-$

$t = 4$: $\hat{y} =$  $\phi(y^n) - \phi(\hat{y})$: black $=$, white $=$, green $-$, blue $=$, gray $=$

SSVM : instantiation

Estimation de pose : $\Delta(\mathbf{y}_i, \mathbf{y})$?

input: image



body model

output: model fit

- input space $\mathcal{X} = \{images\}$
- output space $\mathcal{Y} = \{positions/angle\ of\ K\ \ body\ parts\} \triangleq \mathbb{R}^{3K}$.
- prediction function: $f : \mathcal{X} \to \mathcal{Y}$

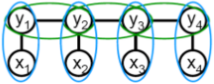
$$f(x) := \operatorname*{argmin}_{y \in \mathcal{Y}} E(x, y)$$

- energy $E(x, y) = \sum_i w_i^\top \varphi_{\mathit{fit}}(x_i, y_i) + \sum_{i,j} w_{ij}^\top \varphi_{\mathit{pose}}(y_i, y_j)$

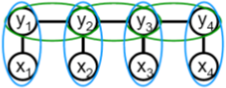
[Ferrari, Marin-Jimenez, Zisserman: "Progressive Search Space Reduction for Human Pose Estimation", CVPR 2008.]

SSVM : instantiation

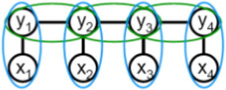
Prédiction de séquence



- Emissions (blue)
  - $f_e(x_i, y_i) = \langle w_e, \varphi_e(x_i, y_i)\rangle$
  - Can simply use the multi-class joint feature map for φ_e
- Transitions (green)
  - $f_t(y_i, y_{i+1}) = \langle w_t, \varphi_t(y_i, y_{i+1})\rangle$
  - $\varphi_t(y_i, y_{i+1}) = \varphi_y(y_i) \otimes \varphi_y(y_{i+1})$ or $\begin{cases} [1\ 0]^T & \text{if } y_i = y_{i+1} \\ [0\ 1]^T & \text{if } y_i \neq y_{i+1} \end{cases}$



$$
\begin{aligned}
w &= \begin{pmatrix} w_e \\ w_t \end{pmatrix} \\
\varphi(x,y) &= \begin{pmatrix} \sum_i \varphi_e(x_i, y_i) \\ \sum_i \varphi_t(y_i, y_{i+1}) \end{pmatrix} \\
f(x,y) &= \langle w, \varphi(x,y)\rangle
\end{aligned}
$$



$$
\begin{aligned}
p(x,y) &\propto \prod_i e^{f_e(x_i,y_i)} \prod_i e^{f_t(y_i,y_{i+1})} \quad \text{for an HMM} \\
f(x,y) &= \sum_i f_e(x_i, y_i) + \sum_i f_t(y_i, y_{i+1}) \\
&= \langle w_e, \sum_i \varphi_e(x_i, y_i)\rangle + \langle w_t, \sum_i \varphi_t(y_i, y_{i+1})\rangle
\end{aligned}
$$

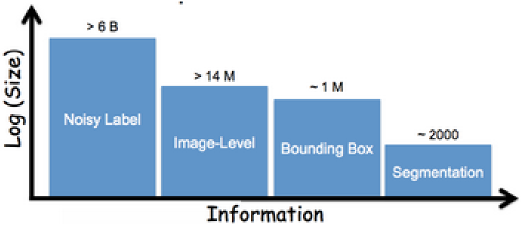
Outline

1. Instanciations
2. Prédiction Structurée avec variable latentes (LSSVM)

Apprentissage faiblement supervisé

Motivation

- Big data : beaucoup de données
- MAIS peu d'annotations en proportion
  - : surtout des annotations fines/précises
- $\Rightarrow$ Apprendre avec des annotations faibles



Apprentissage faiblement supervisé & structuré

Étude d'un formalisme

- Latent Structural SVM [YJ09]
- la relation entrée-sortie ne peut pas être complètement caractériser par l'ensemble d'apprentissage $\mathcal{S} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$, mais dépend d'un ensemble de variables cachées $h \in \mathcal{H}$
- évaluer un Latent Structural SVM revient à apprendre une prédiction de la forme :

$$(\hat{\mathbf{y}}(\mathbf{x}_i; \mathbf{w}), \hat{\mathbf{h}}(\mathbf{x}_i; \mathbf{w})) = \underset{(\mathbf{y}, \mathbf{h}) \in \mathcal{Y} \times \mathcal{H}}{\operatorname{argmax}} \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}, \mathbf{h}) \rangle$$

- $\Psi(\mathbf{x}, \mathbf{y}, \mathbf{h})$ décrit la relation entre l'entrée $\mathbf{x}$, la sortie structurée $\mathbf{y}$ et les variables cachées $\mathbf{h}$

Latent Structural SVM

Extension de la fonction de coût

- mesure la différence entre 2 paires

$$\Delta((\mathbf{y}_i, \mathbf{h}^*(\mathbf{x}_i, \mathbf{w})), (\hat{\mathbf{y}}(\mathbf{x}_i; \mathbf{w}), \hat{\mathbf{h}}(\mathbf{x}_i, \mathbf{w})))$$

avec $(\hat{\mathbf{y}}(\mathbf{x}_i; \mathbf{w}), \hat{\mathbf{h}}(\mathbf{x}_i; \mathbf{w})) = \underset{(\mathbf{y},\mathbf{h}) \in \mathcal{Y} \times \mathcal{H}}{\operatorname{argmax}} \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}, \mathbf{h}) \rangle$

et $\mathbf{h}^*(\mathbf{x}_i; \mathbf{w}) = \underset{\mathbf{h} \in \mathcal{H}}{\operatorname{argmax}} \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}) \rangle$

Latent Structural SVM

majoration de la fonction de coût (cas général) :

$$
\begin{aligned}
&\Delta((\mathbf{y}_i,\mathbf{h}^*(\mathbf{x}_i,\mathbf{w})),(\hat{\mathbf{y}}(\mathbf{x}_i;\mathbf{w}),\hat{\mathbf{h}}(\mathbf{x}_i,\mathbf{w})))\\
&\leq \Delta((\mathbf{y}_i,\mathbf{h}^*(\mathbf{x}_i,\mathbf{w})),(\hat{\mathbf{y}}(\mathbf{x}_i;\mathbf{w}),\hat{\mathbf{h}}(\mathbf{x}_i,\mathbf{w})))\\
&\quad + \underbrace{\langle \mathbf{w},\Psi(\mathbf{x}_i,\hat{\mathbf{y}}(\mathbf{x}_i;\mathbf{w}),\hat{\mathbf{h}}(\mathbf{x}_i,\mathbf{w}))\rangle - \langle \mathbf{w},\Psi(\mathbf{x}_i,\mathbf{y}_i,\mathbf{h}^*(\mathbf{x}_i,\mathbf{w}))\rangle}_{\geq 0}\\
&\leq \left(\max_{(\mathbf{y},\mathbf{h})\in\mathcal{Y}\times\mathcal{H}} \langle \mathbf{w},\Psi(\mathbf{x}_i,\mathbf{y},\mathbf{h})\rangle\right) + \Delta((\mathbf{y}_i,\mathbf{h}^*(\mathbf{x}_i,\mathbf{w})),(\hat{\mathbf{y}}(\mathbf{x}_i;\mathbf{w}),\hat{\mathbf{h}}(\mathbf{x}_i,\mathbf{w})))\\
&\quad - \left(\max_{\mathbf{h}\in\mathcal{H}} \langle \mathbf{w},\Psi(\mathbf{x}_i,\mathbf{y}_i,\mathbf{h})\rangle\right)
\end{aligned}
$$

Latent Structural SVM

- **hypothèse :** la fonction de coût ne dépend pas de pas de la variable cachée $\mathbf{h}^*(\mathbf{x}_i, \mathbf{w})$

$$\Delta((\mathbf{y}_i, \mathbf{h}^*(\mathbf{x}_i, \mathbf{w})), (\hat{\mathbf{y}}(\mathbf{x}_i; \mathbf{w}), \hat{\mathbf{h}}(\mathbf{x}_i, \mathbf{w}))) = \Delta(\mathbf{y}_i, \hat{\mathbf{y}}(\mathbf{x}_i; \mathbf{w}), \hat{\mathbf{h}}(\mathbf{x}_i, \mathbf{w}))$$

- nouvelle majoration :

$$\Delta(\mathbf{y}_i, \hat{\mathbf{y}}(\mathbf{x}_i; \mathbf{w}), \hat{\mathbf{h}}(\mathbf{x}_i, \mathbf{w})) \leq \left(\max_{(\mathbf{y},\mathbf{h}) \in \mathcal{Y} \times \mathcal{H}} \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}, \mathbf{h}) \rangle \right)$$
$$+ \Delta(\mathbf{y}_i, \hat{\mathbf{y}}(\mathbf{x}_i; \mathbf{w}), \hat{\mathbf{h}}(\mathbf{x}_i, \mathbf{w})) - \left(\max_{\mathbf{h} \in \mathcal{H}} \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}) \rangle \right)$$

Problème d'optimisation du Latent Structural SVM

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\left(\max_{(\mathbf{y},\mathbf{h})\in\mathcal{Y}\times\mathcal{H}}\langle \mathbf{w}, \Psi(\mathbf{x}_i,\mathbf{y},\mathbf{h})\rangle + \Delta(\mathbf{y}_i, \hat{\mathbf{y}}(\mathbf{x}_i;\mathbf{w}), \hat{\mathbf{h}}(\mathbf{x}_i,\mathbf{w}))\right)$$
$$-\frac{C}{n}\sum_{i=1}^{n}\left(\max_{\mathbf{h}\in\mathcal{H}}\langle \mathbf{w}, \Psi(\mathbf{x}_i,\mathbf{y}_i,\mathbf{h})\rangle\right)$$

Problème d'optimisation du Latent Structural SVM

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\left(\max_{(\mathbf{y},\mathbf{h})\in\mathcal{Y}\times\mathcal{H}}\langle \mathbf{w}, \Psi(\mathbf{x}_i,\mathbf{y},\mathbf{h})\rangle + \Delta(\mathbf{y}_i,\hat{\mathbf{y}}(\mathbf{x}_i;\mathbf{w}),\hat{\mathbf{h}}(\mathbf{x}_i,\mathbf{w}))\right)$$
$$-\frac{C}{n}\sum_{i=1}^{n}\left(\max_{\mathbf{h}\in\mathcal{H}}\langle \mathbf{w}, \Psi(\mathbf{x}_i,\mathbf{y}_i,\mathbf{h})\rangle\right)$$

- différence de 2 fonctions convexes : $f(\mathbf{w}) - g(\mathbf{w})$

$$f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\left(\max_{(\mathbf{y},\mathbf{h})\in\mathcal{Y}\times\mathcal{H}}\langle \mathbf{w}, \Psi(\mathbf{x}_i,\mathbf{y},\mathbf{h})\rangle + \Delta(\mathbf{y}_i,\hat{\mathbf{y}}(\mathbf{x}_i;\mathbf{w}),\hat{\mathbf{h}}(\mathbf{x}_i,\mathbf{w}))\right)$$

$$g(\mathbf{w}) = \frac{C}{n}\sum_{i=1}^{n}\left(\max_{\mathbf{h}\in\mathcal{H}}\langle \mathbf{w}, \Psi(\mathbf{x}_i,\mathbf{y}_i,\mathbf{h})\rangle\right)$$

Problème d'optimisation du Latent Structural SVM

Algorithm 1 Concave-Convex Procedure (CCCP)

1: Set $t = 0$ and initialize $\boldsymbol{w}_0$
2: **repeat**
3: Find hyperplane $\boldsymbol{v}_t$ such that $-g(\boldsymbol{w}) \leq -g(\boldsymbol{w}_t) + (\boldsymbol{w} - \boldsymbol{w}_t) \cdot \boldsymbol{v}_t$ for all $\boldsymbol{w}$
4: Solve $\boldsymbol{w}_{t+1} = \text{argmin}_{\boldsymbol{w}}\, f(\boldsymbol{w}) + \boldsymbol{w} \cdot \boldsymbol{v}_t$
5: Set $t = t + 1$
6: **until** $[f(\boldsymbol{w}_t) - g(\boldsymbol{w}_t)] - [f(\boldsymbol{w}_{t-1}) - g(\boldsymbol{w}_{t-1})] < \epsilon$

- Rédoudre le problème écrit comme une différence de fonctions convexes : Concave-Convex Procedure [YR03] (CCCP)
  - $\oplus$ Garanties de convergence (vers un minimum local)
  - $\oplus$ Critère formel d'arrêt de l'algorithme

Problème d'optimisation du Latent Structural SVM

Algorithm 1 Concave-Convex Procedure (CCCP)

1: Set $t = 0$ and initialize $\boldsymbol{w}_0$
2: **repeat**
3: Find hyperplane $\boldsymbol{v}_t$ such that $-g(\boldsymbol{w}) \leq -g(\boldsymbol{w}_t) + (\boldsymbol{w} - \boldsymbol{w}_t) \cdot \boldsymbol{v}_t$ for all $\boldsymbol{w}$
4: Solve $\boldsymbol{w}_{t+1} = \text{argmin}_{\boldsymbol{w}}\, f(\boldsymbol{w}) + \boldsymbol{w} \cdot \boldsymbol{v}_t$
5: Set $t = t + 1$
6: **until** $[f(\boldsymbol{w}_t) - g(\boldsymbol{w}_t)] - [f(\boldsymbol{w}_{t-1}) - g(\boldsymbol{w}_{t-1})] < \epsilon$

$$3 \quad \rightarrow \quad \forall i, \quad \mathbf{h}^*(\mathbf{x}_i;\mathbf{w}) = \underset{\mathbf{h}\in\mathcal{H}}{\text{argmax}} \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h})\rangle$$

$$\mathbf{v}_t = \sum_{i=1}^{n} \Psi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}^*(\mathbf{x}_i;\mathbf{w}))$$

Problème d'optimisation du Latent Structural SVM

Algorithm 1 Concave-Convex Procedure (CCCP)

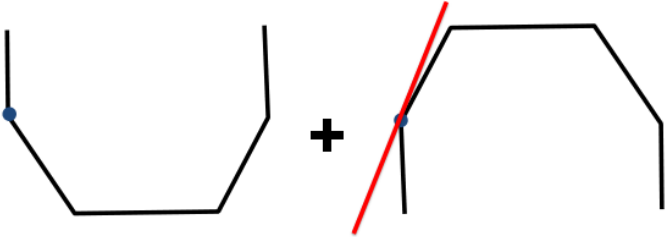
1: Set $t = 0$ and initialize $\boldsymbol{w}_0$
2: **repeat**
3: Find hyperplane $\boldsymbol{v}_t$ such that $-g(\boldsymbol{w}) \leq -g(\boldsymbol{w}_t) + (\boldsymbol{w} - \boldsymbol{w}_t) \cdot \boldsymbol{v}_t$ for all $\boldsymbol{w}$
4: Solve $\boldsymbol{w}_{t+1} = \text{argmin}_{\boldsymbol{w}} f(\boldsymbol{w}) + \boldsymbol{w} \cdot \boldsymbol{v}_t$
5: Set $t = t + 1$
6: **until** $[f(\boldsymbol{w}_t) - g(\boldsymbol{w}_t)] - [f(\boldsymbol{w}_{t-1}) - g(\boldsymbol{w}_{t-1})] < \epsilon$

3 : construit un hyperplan qui majore la fonction concave g
$\Rightarrow$ l'optimisation de la ligne 4 est convexe

Problème d'optimisation du Latent Structural SVM

Algorithm 1 Concave-Convex Procedure (CCCP)

1: Set $t = 0$ and initialize $\boldsymbol{w}_0$
2: **repeat**
3: Find hyperplane $\boldsymbol{v}_t$ such that $-g(\boldsymbol{w}) \leq -g(\boldsymbol{w}_t) + (\boldsymbol{w} - \boldsymbol{w}_t) \cdot \boldsymbol{v}_t$ for all $\boldsymbol{w}$
4: Solve $\boldsymbol{w}_{t+1} = \text{argmin}_{\boldsymbol{w}}\, f(\boldsymbol{w}) + \boldsymbol{w} \cdot \boldsymbol{v}_t$
5: Set $t = t + 1$
6: **until** $[f(\boldsymbol{w}_t) - g(\boldsymbol{w}_t)] - [f(\boldsymbol{w}_{t-1}) - g(\boldsymbol{w}_{t-1})] < \epsilon$



Problème d'optimisation du Latent Structural SVM

Algorithm 1 Concave-Convex Procedure (CCCP)

1: Set $t = 0$ and initialize $\boldsymbol{w}_0$
2: **repeat**
3: Find hyperplane $\boldsymbol{v}_t$ such that $-g(\boldsymbol{w}) \leq -g(\boldsymbol{w}_t) + (\boldsymbol{w} - \boldsymbol{w}_t) \cdot \boldsymbol{v}_t$ for all $\boldsymbol{w}$
4: Solve $\boldsymbol{w}_{t+1} = \operatorname{argmin}_{\boldsymbol{w}} f(\boldsymbol{w}) + \boldsymbol{w} \cdot \boldsymbol{v}_t$
5: Set $t = t + 1$
6: **until** $[f(\boldsymbol{w}_t) - g(\boldsymbol{w}_t)] - [f(\boldsymbol{w}_{t-1}) - g(\boldsymbol{w}_{t-1})] < \epsilon$

4 : résoud le problème d'optimisation standard du Structural SVM

Problème d'optimisation du Latent Structural SVM

Algorithm 1 Concave-Convex Procedure (CCCP)

1: Set $t = 0$ and initialize $\boldsymbol{w}_0$
2: **repeat**
3: Find hyperplane $\boldsymbol{v}_t$ such that $-g(\boldsymbol{w}) \leq -g(\boldsymbol{w}_t) + (\boldsymbol{w} - \boldsymbol{w}_t) \cdot \boldsymbol{v}_t$ for all $\boldsymbol{w}$
4: Solve $\boldsymbol{w}_{t+1} = \text{argmin}_{\boldsymbol{w}}\, f(\boldsymbol{w}) + \boldsymbol{w} \cdot \boldsymbol{v}_t$
5: Set $t = t + 1$
6: **until** $[f(\boldsymbol{w}_t) - g(\boldsymbol{w}_t)] - [f(\boldsymbol{w}_{t-1}) - g(\boldsymbol{w}_{t-1})] < \epsilon$

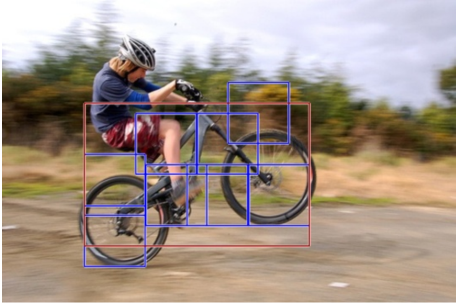
$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\left(\max_{(\mathbf{y},\mathbf{h})\in\mathcal{Y}\times\mathcal{H}} \langle \mathbf{w}, \Psi(\mathbf{x}_i,\mathbf{y},\mathbf{h})\rangle + \Delta(\mathbf{y}_i, \hat{\mathbf{y}}(\mathbf{x}_i;\mathbf{w}), \hat{\mathbf{h}}(\mathbf{x}_i,\mathbf{w}))\right)$$

$$-\frac{C}{n}\sum_{i=1}^{n}\left(\max_{\mathbf{h}\in\mathcal{H}} \langle \mathbf{w}, \Psi(\mathbf{x}_i,\mathbf{y}_i,\mathbf{h}^*(\mathbf{x}_i;\mathbf{w}))\rangle\right)$$

LSSVM : applications en vision

Deformable Part Model (DPM) [FGMR10]

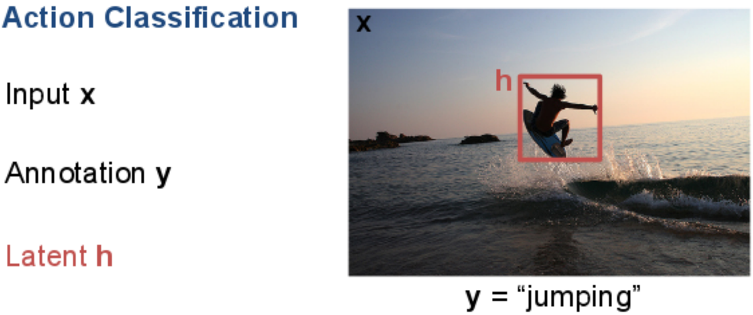
- DPM : appliqué à la détection d'objets :
  - Problème de classification binaire $\mathcal{Y} = \pm 1$
  - En entrée : une Bounding Box de l'objet
  - Variable latente : position de (sous)-parties de l'objet



LSSVM : applications en vision

Localisation faiblement supervisée [KPK10, RLYFF12, BNVG13]

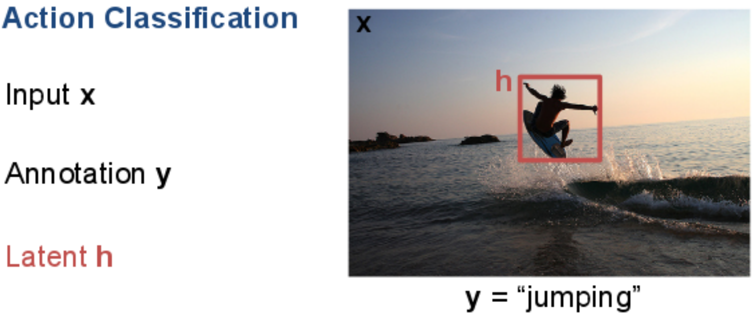
- x : image
- h, position de la région
- y : classe, *i.e.* "jumping", etc
- Feature map : cf classification multi-classes



LSSVM : applications en vision

Localisation faiblement supervisée

- x : image
- h, position de la région
- y : image pertinent/non pertinente, *i.e.* $\mathcal{Y} = \pm 1$ (pour "jumping", etc)
- Feature map : cf ranking



Localisation faiblement supervisée : extensions récentes

Extensions du LSSVM

- Optimisation : problème non convexe ⇒ difficuluté pour atteindre un minimum local pertient :
  - Idée du Curriculum learning : apprendre d'abord les paramètres du modèles avec des exemples faciles.
    Challenge : définition d'un exemple facile.
  - Variante : exploration incrementale de l'espace latent : [RLYFF12, BNVG13]
- Modèle : Comment aggréger les scores des variables latentes
  - LSSVM : max, *i.e.* sélection "dure" de la meilleure variable latente
  - HCRF sum [QWM+07]
  - Modéliser l'ambiguité entre variables latentes dans modèle LSSVM: M3E [MKP+12, MKP+12]
  - Modèle unifiés entre HCFR (sum) et LSSVM (max) : ϵ-extension [SHPU12], marginal SVM [PLI14]

References I

Matthew B. Blaschko and Christoph H. Lampert, *Learning to localize objects with structured output regression*, Proceedings of the 10th European Conference on Computer Vision: Part I (Berlin, Heidelberg), ECCV '08, Springer-Verlag, 2008, pp. 2–15.

H. Bilen, V.P. Namboodiri, and L.J. Van Gool, *Object classification with latent window parameters*, International Journal of Computer Vision, 2013.

Thibaut Durand, Nicolas Thome, and Matthieu Cord, *MANTRA: Minimum Maximum Latent Structural SVM for Image Classification and Ranking*, ICCV, 2015.

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, *Object detection with discriminatively trained part based models*, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010), no. 9, 1627–1645.

P. Kumar, B. Packer, and D. Koller, *Self-paced learning for latent variable models*, Advances in Neural Information Processing Systems (NIPS 2010), 2010.

Kevin Miller, M. Pawan Kumar, Benjamin Packer, Danny Goodman, and Daphne Koller, *Max-margin min-entropy models*, Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS, 2012.

Wei Ping, Qiang Liu, and Alex Ihler, *Marginal structured svm with hidden variables*, Proceedings of the 31st International Conference on Machine Learning (ICML-14) (Tony Jebara and Eric P. Xing, eds.), JMLR Workshop and Conference Proceedings, 2014, pp. 190–198.

Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Michael Collins, and Trevor Darrell, *Hidden conditional random fields*, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007), no. 10, 1848–1852.

Olga Russakovsky, Yuanqing Lin, Kai Yu, and Li Fei-Fei, *Object-centric spatial pooling for image classification*, ECCV, 2012.

References II

Alexander G. Schwing, Tamir Hazan, Marc Pollefeys, and Raquel Urtasun, *Efficient structured prediction with latent variables for general graphical models*, Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012, 2012.

Jia Xu, Alexander G. Schwing, and Raquel Urtasun, *Tell me what you see and i will show you where it is*, CVPR, 2014.

Yisong Yue, Thomas Finley, Filip Radlinski, and Thorsten Joachims, *A support vector method for optimizing average precision*, Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (New York, NY, USA), SIGIR '07, ACM, 2007, pp. 271–278.

Chun-Nam Yu and T. Joachims, *Learning structural svms with latent variables*, International Conference on Machine Learning (ICML), 2009.

Alan L. Yuille and Anand Rangarajan, *The concave-convex procedure*, Neural Computation (2003).