prédiction structurée

L lettres

y: B R A C E

motivation : OCR

x: [b][r][a][c][e]

K lettre possible

$f : \mathcal{X} \to \mathcal{Y}$ } objets structurés

$|\mathcal{Y}| = K^{L} \to$ # exponentiel de possibilités

approche multiclasse : one-vs-rest } impossible ?
régression logistique multiclasse }

approche indépendante $\to$ ignore les corrélations entre décision

$$Y = \underbrace{(y_1, \ldots, y_L)}_{\text{à corrélation ?}}$$

$p(Y|x)$

prédiction structuré
$\to$ décision jointe

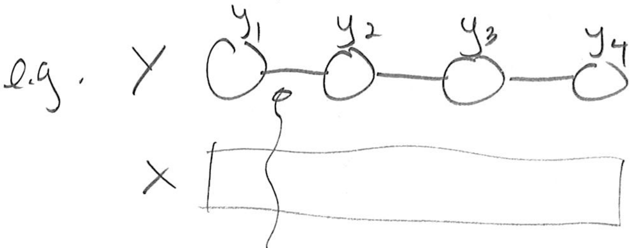
problème : $p(Y|x)$
pour x fixé ; $K^L - 1$ paramètres en général (!)
[# augmente très rapidement...]

exploiter structure pour < statistique (trop de paramètres)
relié
computationnel ($\max_{y \in \mathcal{Y}} p(y|x)$ NP-dur en général)

outil (cours MVA) : modèle graphique

suppose des indépendances conditionnelles entre les variables :

e.g.

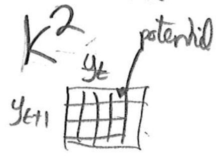


chaîne

$$p(Y|x) = \frac{1}{Z(x)} \prod_t \underbrace{\Psi_t(y_t, y_{t+1}, x)}_{\#\text{paramètres } K^2}$$

factorise avec facteur sur paire (y_t, y_{t+1})

tableau



clique (y_t, y_{t+1})

graphe encode les indépendances conditionnelles

donc passe de $\approx K^L$ à LK^2 paramètres dans cours MVA, montre graphe $\Leftrightarrow$ factorisation $\Leftrightarrow$ indépendances conditionnelles

ici $\quad y_{t+2} \perp\!\!\!\perp y_t \mid y_{t+1}$

i.e. $p(y_{t+2}, y_t \mid y_{t+1}, x) = p(y_{t+2} \mid y_{t+1}, x) \cdot p(y_t \mid y_{t+1}, x)$

<u>pourquoi factorisation aide ?</u>

$$\text{score}(x,y) \triangleq \log p(y|x) = \sum_t \underbrace{\log \psi_t(y_t, y_{t+1}, x)}_{\text{score}(y_t, y_{t+1}, x)} - \log Z(x)$$

$$\max_{y \in \mathcal{Y}} [\text{score}(x,y)] = \max_{y_1} \Big[\max_{y_2} sc(y_1,y_2) + \Big[\dots + \underbrace{\max_{y_{K-1}} \Big[\underbrace{sc(y_{K-2},y_{K-1}) + \underbrace{\max_{y_K} sc(y_{K-1},y_K)}_{m(y_{K-1})}}_{m(y_{K-2})}\Big]}_{m(y_2)}\Big]\Big]$$

$m(y_2)$ } "messages" dans graphe

temps $\rightarrow O(LK^2)$ [vs. $O(K^L)$!?]

programmation dynamique pour trouver argmax

algorithme de <u>viterbi</u>

par similarité :

[anneau $(\max, +)$ vs. $\overbrace{(+, \times)}^{\text{distributivité}}$]

$$Z(x) = \sum_y \Big[\prod_t \psi_t(y_t, y_{t+1})\Big] = \sum_{y_1} \Big[\sum_{y_2} \psi_1(y_1,y_2) \Big[\sum_{y_3} \psi_2(y_2,y_3) \Big[\dots \overbrace{\sum_{y_L} \psi_L(y_{L-1}, y_L)}^{m(y_{L-1})}\Big]\Big]\Big]$$

distributivité $\quad$ "sum-product algorithm"

$ax + ay + bx + by$

$= (a+b)(x+y)$

donc manière "tractable" de faire regression logistique structurée

CRF = Conditional $\underbrace{\text{Random Field}}$

$p(y|x)$ is $\underbrace{\text{random field}}_{\text{special case of graphical model}}$

$p(y|x) \propto \exp(\text{score}(x,y))$

$Z(x) = \sum_y \exp(\text{score}(x,y))$

$\text{score}(x,y) = y w^T x$

pour $y \in \{0,1\}$

donc regression logistique

$w_y^T x \rightarrow$ multiclasse

exemple de paramétrisation : $\text{score}(w; x,y) \triangleq \sum_t (w_{node})^T \varphi_{node}(x_t, y_t)$

$+ \sum_t \underbrace{(w_{edge})^T \varphi_{edge}(y_t, y_{t+1})}_{sc(y_t, y_{t+1}, x)}$

$= w^T \varphi(x,y)$

pour OCR :

$\varphi(y_t, y_{t+1}) = \underbrace{\begin{pmatrix} \mathbb{1}\{y_t = 'a', y_{t+1} = 'c'\} \\ \\ \\ \end{pmatrix}}_{\text{dimension } K^2}$

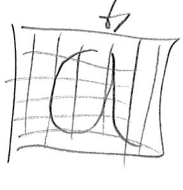
feature "indicatrice"

donc possibilité d'apprendre les probabilités $p(y_t, y_{t+1})$ marginale [corrélation adjacente de lettres]

$\varphi(x_t, y_t) = \begin{pmatrix} \vdots \\ \mathbb{1}\{y_t = 'a'\} \vec{x}_t \\ \vdots \end{pmatrix}$

Kd paramètres

$w_{node}^{'a'}$



régression logistique régularisé :

$$\min_{w} \quad \lambda\|w\|^2 + \frac{1}{n}\sum_i \Big[-\log \underbrace{p(y_i|x_i)}_{\frac{\exp(w^T\psi(x,y))}{Z(x)}}\Big] = f_\lambda(w)$$

$$\lambda\|w\|^2 + \frac{1}{n}\sum_i \Big[\log(Z(x)) - w^T\psi(x_i,y_i)\Big]$$

$$\nabla f_\lambda(w) = \lambda \vec{w} + \frac{1}{n}\sum_i \nabla_w \log(\underbrace{Z(w)}_{}) - \psi(x_i,y_i)$$

$$\log\Big[\sum_y \exp(w^T\psi(x,y))\Big]$$

$$\sum_y \underbrace{\frac{\exp(w^T\psi(x,y))}{Z(x_i)}}_{p(y|x)} \psi(x,y)$$

$$= \underbrace{\mathbb{E}_{p_w(y|x)} \psi(x_i,y)}_{\text{peut calculer avec sum-product}} \qquad - \psi(x_i,y_i)$$

[sans régularisation, $\nabla f_\lambda = 0 \Rightarrow$ $\frac{1}{n}\sum_i \Big[\mathbb{E}_{p_w(y|x_i)}\psi(x_i,y) - \psi(x_i,y_i)\Big] = 0$

"feature moment matching"

---

voir slides Ben Taskar

---

autres exemples :
- reconnaissance vocale
- traduction automatique
- coding région detection in DNA
- etc ?
- alignment de mots
- ...

autres sujets $p(y)$ ← apprendre

↑ continue

- "unsupervised learning"
- dimensionality reduction

etc ...

CS + STAT !