Lecture 1 - setup

Tuesday, January 10, 2023 2:25 PM

IFT6132

TODO: fill survey http://bit.ly/IFT6132-W23 by tomorrow!

structured prediction basics & setup

learning problem:

given a training data dataset $D = (x^{(i)}, y^{(i)})_{i=1}^n$

$x^{(i)} \in \mathcal{X}$ $y^{(i)} \in \mathcal{Y}$

goal: learn a prediction mapping $h_w : \mathcal{X} \to \mathcal{Y}$ (w: parameter)

that has a low classification error or generalization error

$$L(w; P) \triangleq \mathbb{E}_{(x,y)\sim P}[\ell(y, h_w(x))]$$

P: test distribution

ℓ: structured error function

L: statistical decision loss ("risk" in ML (Vapnik risk))

[see lecture 5 of my PGM → review statistical decision theory]

example of learning approach:

regularized ERM ("empirical risk minimization")

$$\hat{L}(w) = \frac{1}{n}\sum_{i=1}^n \ell(y^{(i)}, h_w(x^{(i)})) + R(w)$$

R: regularizer

not cts. in w

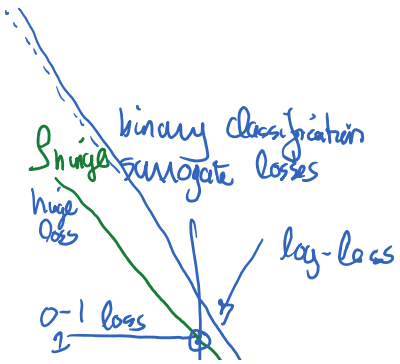
messy, non-convex, NP hard to minimize in general]

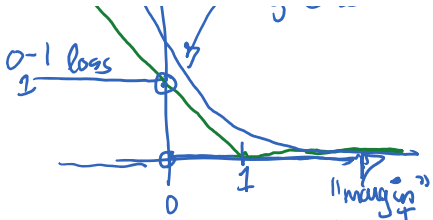
↝ replace $\hat{L}$ with a surrogate loss (context: est. ← statistics, ML)

$$\hat{\mathcal{L}}(w) = \frac{1}{n}\sum_{i=1}^n \mathcal{L}(x^{(i)}, y^{(i)}, w) + R(w)$$

$\mathcal{L}$: surrogate loss e.g. convex in w

$\hat{\mathcal{L}}$: M-estimator





M-estimator in statistics

examples:
- structured hinge loss → structured SVM
- log-loss → CRF aka. "cross-entropy"

generative vs. discriminative learning continuum

(x, y) (input, output)

generative modeling

$p_w(x, y)$ → conditional $p(y|x)$ → prediction function

"more discriminative"

model:

$p_w(x, y)$ | $p_w(y|x)$ | $h_w : \mathcal{X} \to \mathcal{Y}$

learning: learn $\hat{w}$ by ML | learn $\hat{w}$ by MCL | use $\ell(y, y')$ to construct surrogate loss

learn $\hat{w}$ by empirical surrogate loss minimization

e.g. structured SVM

← more assumptions / less robust classification tasks

some important aspects of structured prediction (vs. binary classification)

1) $\mathcal{Y}$ output is usually exponentially big (in natural size of input)

2) structured error function $\ell(y, y')$ [$\ell(x, y, y')$]

e.g. Hamming loss $\ell(y, y') = \sum_p \mathbb{1}\{y_p \neq y'_p\}$

3) sometimes constraints on pieces of y

↳ eg. consider word alignment example:

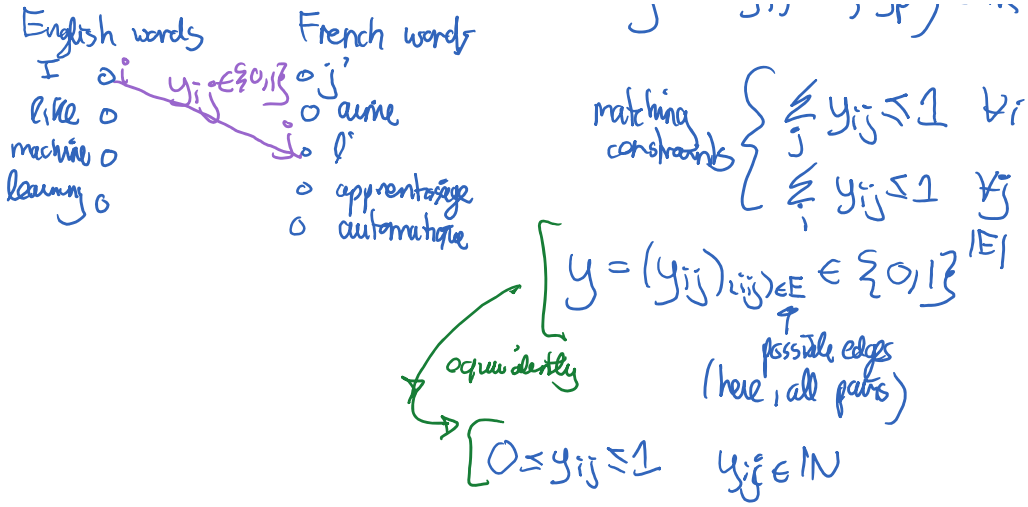
English words | French words

I | ai

$y_i \in \{0, 1\}$ "j"

note: "need encoding fct."

$y = (y_1, \dots, y_p) \in \mathbb{R}^p$


English words
French words
I
like
machine
learning
j'
aime
l'
apprentissage
automatique
$y_{ij} \in \{0,1\}$
matching constraints
$\sum_j y_{ij} \le 1 \quad \forall i$
$\sum_i y_{ij} \le 1 \quad \forall j$
$y = (y_{ij})_{(i,j) \in E} \in \{0,1\}^{|E|}$
possible edges (here, all pairs)
equivalently
$0 \le y_{ij} \le 1 \quad y_{ij} \in \mathbb{N}$