Lecture 1 - setup

Thursday, January 14, 2021 13:23

IFT6132

TODO: fill survey http://bit.ly/IFT6132-W21 by Monday evening!

structured prediction basics & setup

Learning problem:

given a training 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 low classification error / generalization error

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

P: test distribution; ℓ: structured error function

Statistical decision loss / "risk" in ML (Vapnik)

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

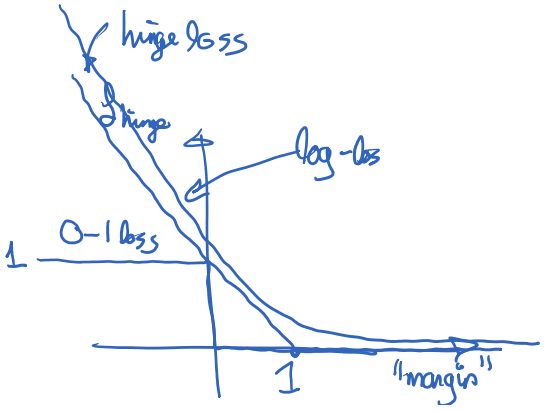
Example of learning approach:

regularized ERM ("empirical risk minimization")

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

not cts. in w; non-convex; messy [NP hard to minimize in general]

regularizer

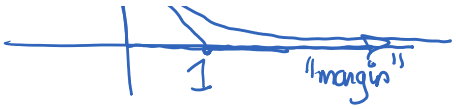


↝ replace $\hat{L}$ with a surrogate loss / contrast fct. (ML / statistics)

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

surrogate loss e.g. convex in w

↳ examples: • structured hinge loss



M-estimator in statistics

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

Generative vs. discriminative learning continuum

generative modeling (x, y) (input, output)

$p_w(x,y)$ → 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 surrogate loss minimization
	CRF		e.g. structured SVM

← more assumptions / less robust on classification tasks

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

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

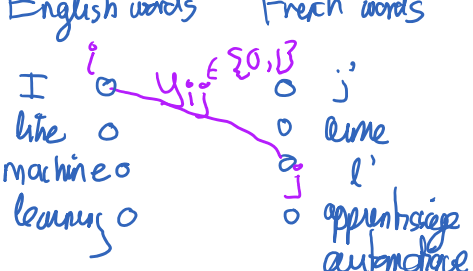
2) structured error function $\ell(y,y')$

3) sometimes constraints on pieces of y

↳ consider word alignment example

note: need "encoding fct."

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



matching constraints $\begin{cases} \sum_j y_{ij} \leq 1 \ \forall i \\ \sum_i y_{ij} \leq 1 \end{cases}$

$y = (y_{ij})_{(i,j) \in E} \in \{0,1\}^{|E|}$

possible edges
(here, all pairs)