Lecture 1 scribbles

Tuesday, January 8, 2019 14:37

bit.ly/IFT6132-W19 ← fill survey ASAP

structured prediction basics & setup

Learning problem:

given a training dataset $D = \left(x^{(i)}, y^{(i)}\right)_{i=1}^{n}$, $x^{(i)} \in \mathcal{X}$, $y^{(i)} \in \mathcal{Y}$

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

that has low generalization error $L(w; P) \triangleq \mathbb{E}_{(x,y)\sim P}\left[\ell\left(y, h_w(x)\right)\right]$

"test distribution"

structured error fct.

→ evaluates how bad different mistakes are

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

[See lecture 4 & 5 of my PGM → review of statistical decision theory]

regularized ERM
empirical risk minimization

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

regularizer

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

⟹ replace it with surrogate loss / contrast fct.

ML statistics

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

"M-estimator" in statistics

surrogate loss e.g. convex in w

examples:
- structured hinge loss → structured SVM
- log-loss → CRF

generative vs. discriminative learning continuum

generative modelling of (x, y) — input x, output y

$p_w(x,y)$ ⟶ conditional: $p_w(y|x)$ ⟶ $h_w^{(p_w)}(x) = \arg\max_{y \in \mathcal{Y}} p_w(y|x)$

"more discriminative"

$y \in \mathcal{Y}$

"more discriminative"

$p_w(x,y)$ — learn $\hat{w}$ by ML

$p_w(y|x)$ — learn $\hat{w}$ by MCL

$h_w : \mathcal{X} \to \mathcal{Y}$, $\ell(y, y')$ — learn $\hat{h}_w$ by using surrogate loss minimization e.g. structured SVM

← more assumptions you make / less robust for classification task