Confusion-Based Online Learning and a Passive–Aggressive Scheme

— NIPS 2012, Poster W77 —

Liva Ralaivola

LIF, CNRS, Aix-Marseille Université, France

Minding the Confusion

Multiclass learning: find a predictor $h \in \mathcal{Y}^{\mathcal{X}}$

- $\mathcal{X} = \mathrm{R}^d$, $\mathcal{Y} = \{1, \ldots, Q\}$
- $\mathbf{Z} = \{(X_i, Y_i)\}_{i=1}^n$, IID sample drawn according to D

Binary classification, $Q = 2$, e.g. $\mathcal{Y} = \{\text{toxic}, \neg\text{toxic}\}$

If $\mathbb{P}(Y = \text{toxic}) = 0.01$ and $\mathbb{P}(Y = \neg\text{toxic}) = 0.99$, then the predictor h_{majority} that always outputs $\neg$toxic

- has error rate $\mathbb{P}(h(X) \neq Y) = 0.01$
- makes 100% error on *positive* data

Finer grain measure of performance than the error rate

	$\widehat{Y} = +1$	$\widehat{Y} = -1$
$Y = +1$	True Positive (TP)	False Negative (FN)
$Y = -1$	False Positive (FP)	True Negative (TN)

Cf.: precision, recall, F1 score, ROC, ...

Minding the Confusion

$Q > 2$, (0-diagonal) confusion matrix wrt h

$$\mathcal{C}(h) = \begin{bmatrix} \mathbf{0} & \cdots & \cdots & \mathbb{P}(h(X) = Q | Y = 1) \\ \mathbb{P}(h(X) = 1 | Y = 2) & \cdots & \cdots & \mathbb{P}(h(X) = Q | Y = 2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{P}(h(X) = 1 | Y = Q) & \cdots & \cdots & \mathbf{0} \end{bmatrix}$$

Size of the confusion matrix as error measure

$$\|\mathcal{C}(h)\|$$

Connection between error rate and size of the confusion

$$\mathbb{P}(h(X) \neq Y) = \left\| \boldsymbol{\pi}^\top \mathcal{C}(h) \right\|_1 \leq \sqrt{Q} \|\mathcal{C}(h)\|$$

Therefore, small $\|\mathcal{C}(h)\|$ induces small $\mathbb{P}(h(X) \neq Y)$

COPA a PA approach toward small $\|\mathcal{C}(h)\|$

COPA – COnfusion Passive Aggressive learning

$$\min_{\mathcal{W}} \frac{1}{2}\sum_{q=1}^{Q}\|\mathbf{w}_q - \mathbf{w}_q^t\|^2 + \frac{C}{n_{y_t}^2}\sum_{q\neq y_t}\left|\langle \mathbf{w}_q, \mathbf{x}_t\rangle + \frac{1}{Q-1}\right|_+^2$$

$$\text{s.t. } \sum_q \mathbf{w}_q = \mathbf{0}$$

Features

- Based on a consistent multiclass SVM
- Update made on *all* offending classifiers
- *'Optimization-free'* updates ($\mathcal{O}(Q \ln Q)$)
- Statistical soundness (cf. matrix martingales)

Bibliographical refs

[Crammer et al., 2006], [Le et al., 2004], [Matsushima et al, 2010], [Tropp, 2011]