Optimizing Non-decomposable Performance Measures: A Tale of Two Classes

Harikrishna Narasimhan*, Purushottam Kar#, and Prateek Jain#

*Indian Institute of Science, INDIA

#Microsoft Research, INDIA

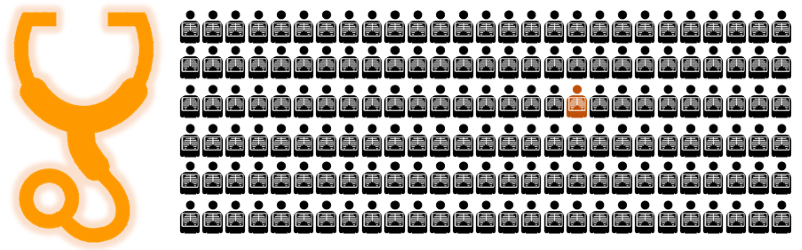
Goal: Scalable (*point-wise*) stochastic optimization methods for two broad families of non-decomposable performance measures

SPADE: Stochastic PrimAl Dual mEthod (for *concave* measures based on *dual structure*)

STAMP: STochastic Alternating Maximization Procedure (for *pseudo-linear* measures based on *level set structure*)

Non-decomposable Performance Measures

$\mathbf{x} \longrightarrow \boxed{h} \longrightarrow y \in \{-1, +1\}$

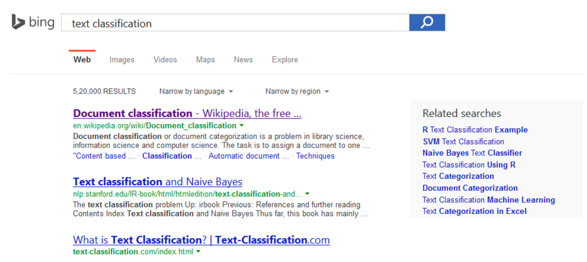
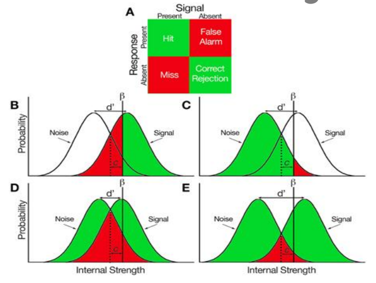


Classification Error:

$$\underbrace{\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}(y_t \neq h(x_t))}_{\text{point-wise loss}}$$

KDD Cup 08 Cancer Detection Challenge: $p_+ = 0.6\%$

Default classifier that predicts 'all positives': accuracy of 99.4%

Medical Diagnosis	Text Classification	Detection Theory
		
G-mean $\sqrt{\mathrm{TPR} \times \mathrm{TNR}}$	F-measure $\frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$	Min-max $\min\{\mathrm{TPR}, \mathrm{TNR}\}$

cannot be expressed as a sum of point-wise losses!

Stochastic Gradient Descent?

'*Point-wise*' performance measures:

$$\min_{\mathbf{w} \in \mathcal{W}} \frac{1}{T}\sum_{t=1}^{T} \ell(\mathbf{w}; \mathbf{x}_t, y_t)$$

SGD Update:

$$\mathbf{w}_{t+1} \leftarrow \Pi_{\mathcal{W}}(\mathbf{w}_t - \underbrace{\nabla_{\mathbf{w}} \ell(\mathbf{w}_t; \mathbf{x}_t, y_t)}_{\text{unbiased gradient estimate}})$$

'*Non-decomposable*' performance measures:

$$\min_{\mathbf{w} \in \mathcal{W}} \underbrace{\mathcal{L}((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T); \mathbf{w})}_{\text{function of all points!}}$$

Mini-batch SGD (Kar et al., '14)

- Large buffers, weak convergence rates
- Applies to general losses – unable to exploit problem structure completely

KEY IDEA

$\exists$ c_1 and c_2 s.t. (Koyejo et al., '14; Narasimhan et al., '14)

$$\operatorname*{argmax}_h \Psi(\mathrm{TPR}(h), \mathrm{TNR}(h)) \quad = \quad \operatorname*{argmax}_h c_1 \mathrm{TPR}(h) + c_2 \mathrm{TNR}(h)$$

Cross-validate for c_1 and c_2?

- Expensive in online settings
- Not incremental

***Adaptive Linearization of* ψ:** Tune c_1 and c_2 on the fly

Multiclass settings: exponential time (Narasimhan et al., '15)

Methods	Online / Any-time?	Point-wise update?
Batch: SVMPerf (Joachims, '05), DTA (Ye et al., '12)	✗	✗
Cross-validation : —Plug-in (Koyejo et al., '14, Narasimhan et al., '14) —Cost-based classification (Parambath et al., '14)	✗	✗
Mini-batch SGD: (Kar et al., '14)	✓	✗
This paper	✓	✓

Concave Measures

$$\Psi(P, N) = \text{Concave function of } P \text{ and } N$$

G-mean (Daskalaki et al., '06)	Min-TPR/TNR (Vincent, '94)	H-mean (Kennedy et al., '10)	Q-mean (Liu & Chawla, '11)
$\sqrt{PN}$	$\min\{P, N\}$	$\frac{2PN}{P+N}$	$1 - \sqrt{\frac{(1-P)^2 + (1-N)^2}{2}}$

Fenchel Duality

$$\sup_{\mathbf{w} \in \mathcal{W}} \mathcal{P}(\mathbf{w}) = \boxed{\Psi(P(\mathbf{w}), N(\mathbf{w}))}$$

$$\equiv$$

$$\sup_{\mathbf{w} \in \mathcal{W}} \inf_{\alpha, \beta \in \mathbb{R}} \{\alpha P(\mathbf{w}) + \beta N(\mathbf{w}) - \Psi^*(\alpha, \beta)\}$$

For any $\alpha, \beta \in \mathbb{R}$ (dual variables),

$$\Psi^*(\alpha, \beta) = \inf_{u, v \in \mathbb{R}} \{\alpha u + \beta v - \Psi(u, v)\}$$

Linear in P and N for fixed dual variables

SPADE

Maintain: $\mathbf{w}_t$, $\begin{bmatrix}\alpha_t \\ \beta_t\end{bmatrix}$ relative weights on positives and negatives

(Concave rewards r^+ and r^- used in place of true positive and true negative indicators)

x_1	x_2	x_3	…	…	x_t	…	…
y_1	y_2	y_3	…	…	y_t	…	…

$$\overline{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$$

→ ***Primal ascent update:***

$$\mathbf{w}_{t+1} \leftarrow \Pi_{\mathcal{W}}(\mathbf{w}_t + \eta_t \nabla_{\mathbf{w}}[\alpha_t r^+(\mathbf{w}_t; \mathbf{x}_t, y_t) + \beta_t r^-(\mathbf{w}_t; \mathbf{x}_t, y_t)])$$

→ ***Dual descent update:***

$$\begin{bmatrix}\alpha_{t+1} \\ \beta_{t+1}\end{bmatrix} \leftarrow \Pi_{\mathcal{A}}\left(\begin{bmatrix}\alpha_t \\ \beta_t\end{bmatrix} - \eta_t' \nabla_{\alpha,\beta}[\alpha_t r^+(\mathbf{w}_t; \mathbf{x}_t, y_t) + \beta_t r^-(\mathbf{w}_t; \mathbf{x}_t, y_t)]\right)$$

Convergence Guarantees

$$\textbf{Theorem: } \text{For Lipschitz } \Psi, \text{ after } T \text{ points, w.h.p. } \sup_{\mathbf{w} \in \mathcal{W}} \mathcal{P}(\mathbf{w}) - \mathcal{P}(\overline{\mathbf{w}}) \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$$

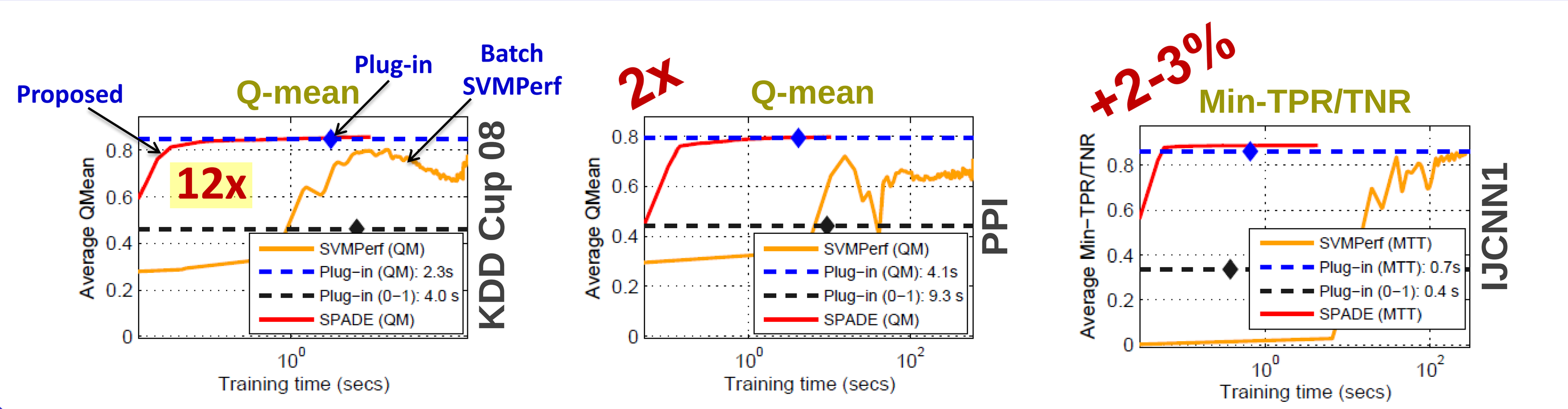
(Holds for: Min-TPR/TNR, H-mean, Q-mean)

For non-Lipschitz ψ, apply SPADE to Lipschitz approximation. E.g. for G-mean,

$$\textbf{Theorem: } \text{After } T \text{ points, w.h.p. } \sup_{\mathbf{w} \in \mathcal{W}} \mathcal{P}_{\text{G-mean}}(\mathbf{w}) - \mathcal{P}_{\text{G-mean}}(\overline{\mathbf{w}}) \leq \tilde{\mathcal{O}}\left(\frac{1}{T^{1/4}}\right)$$

Proof idea: Analysis of primal and dual updates, tied together by Fenchel duality

Experiments



Pseudo-linear Measures

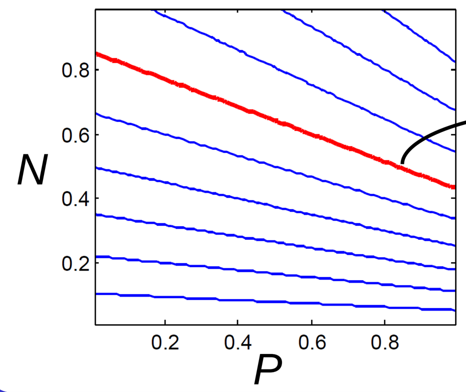
$$\Psi(P, N) = \frac{a_0 + a_1 P + a_2 N}{b_0 + b_1 P + b_2 N}$$

F-measure (Manning et al., '08)	Jaccard Coefficient (Koyejo et al., '14)
$\frac{2P}{1 + \theta + P - \theta N}$	$\frac{P}{1 + \theta - \theta N}$

Gower-Legendre measure (Sokolova & Lapalme, '09)

(where θ is the ratio of proportions of positives to negatives)

Linear Level Sets



Example: F-measure $\mathcal{P}$

Level sets (linear in P and N)

$$\left(1 - \frac{v}{2}\right)P(\mathbf{w}) + \frac{v}{2}N(\mathbf{w}) = v$$

Adaptive linearization: Use level set function as a *proxy*?

(also exploited in the method of Parambath et al., '14)

STAMP

Batch (full training set)

(A) $\mathbf{w}_{j+1} = \operatorname*{argmax}_{\mathbf{w} \in \mathcal{W}} \left(1 - \frac{v_j}{2}\right)P(\mathbf{w}) + \frac{v_j}{2}N(\mathbf{w})$

(B) $v_{j+1} = \mathcal{P}(\mathbf{w}_{j+1})$

$\mathcal{P}$: $v_j \to v_{j+1} = \mathcal{P}(\mathbf{w}_{j+1})$

Historical note: Update of Dinkelbach ('67) & Jagannathan ('66) over parameterized spaces

Stochastic (streaming) '*double*' epoch size after each iteration

A	B	A	B	A	B	……

Convergence Guarantees

Batch:

$$\textbf{Theorem: } \text{After } j \text{ iterations, } \sup_{\mathbf{w} \in \mathcal{W}} \mathcal{P}(\mathbf{w}) - \mathcal{P}(\mathbf{w}_j) \leq \mathcal{O}(2^{-j})$$

Stochastic:

$$\textbf{Theorem: } \text{After } T \text{ points, w.h.p. } \sup_{\mathbf{w} \in \mathcal{W}} \mathcal{P}(\mathbf{w}) - \mathcal{P}(\overline{\mathbf{w}}) \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$$

(also holds for other performance measures for appropriate epoch scaling)

Proof idea: Batch alternating maximization procedure with noisy updates

Experiments