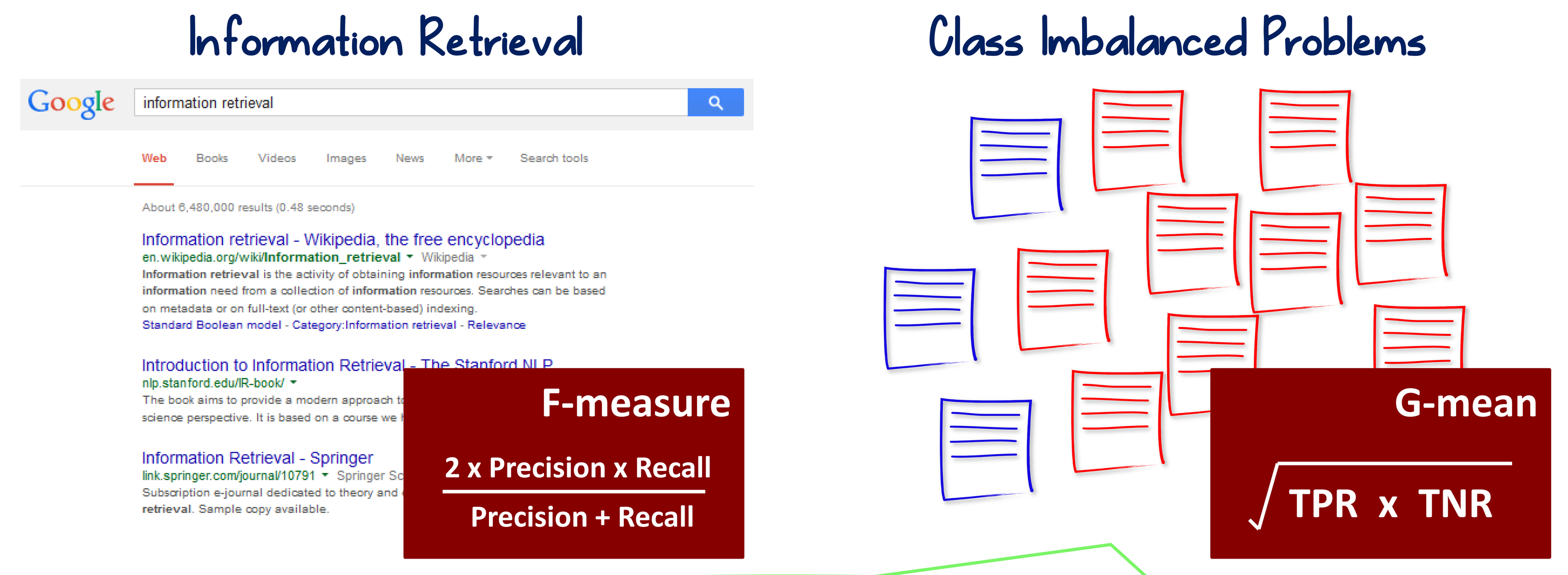
On the Statistical Consistency of Plug-in Classifiers for Non-decomposable Performance Measures

Harikrishna Narasimhan, Rohit Vaish and Shivani Agarwal

Department of Computer Science and Automation, Indian Institute of Science, Bangalore

Non-decomposable Performance Measures

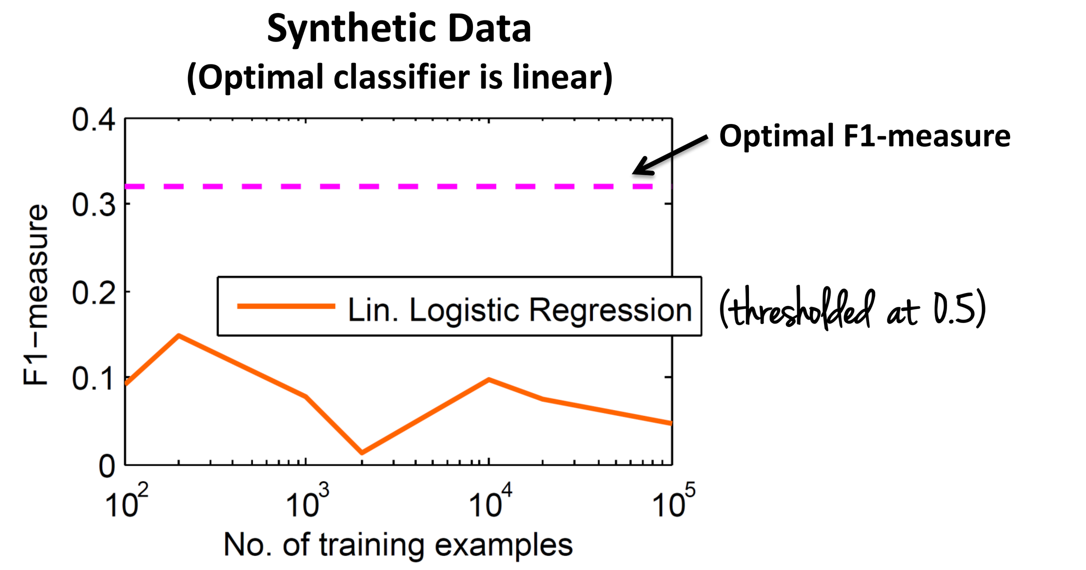


Cannot be expressed as a sum of point-wise errors!

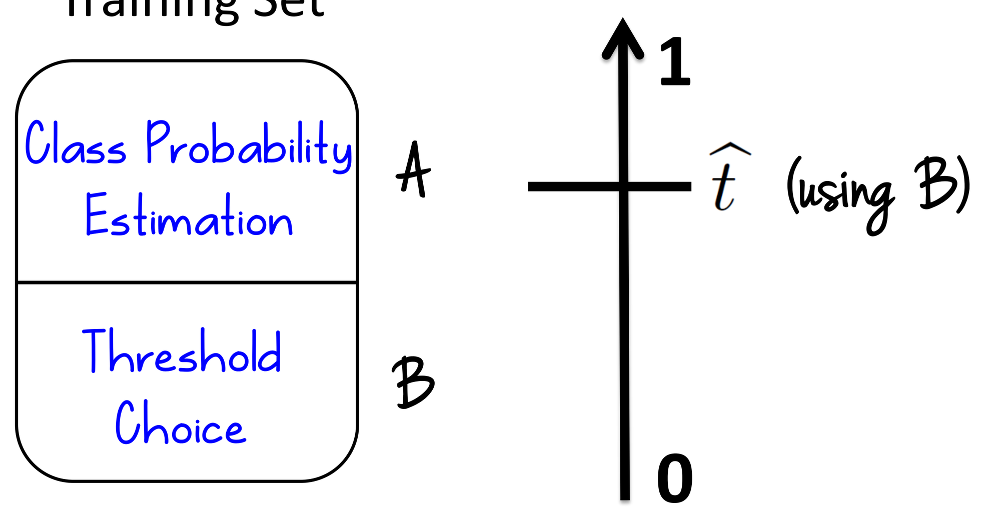
Measure	Definition	References
G-Mean (GM)	$\sqrt{\text{TPR}\cdot\text{TNR}}$	Kubat & Matwin (1997); Daskalaki et al. (2006)
H-Mean (HM)	$2/(\frac{1}{\text{TPR}}+\frac{1}{\text{TNR}})$	Kennedy et al. (2009)
Q-Mean (QM)	$1-((\text{FPR})^2+(\text{FNR})^2)/2$	Lawrence et al. (1998)
G-TP/PR	$\sqrt{\text{TPR}\cdot\text{Prec}}$	Daskalaki et al. (2006)

Are Popular Methods Statistically Consistent?

Are popular classification methods statistically consistent for non-decomposable measures, i.e., converge in the limit of infinite training data to the optimal classifier for the measures?



Simple Plug-in Method

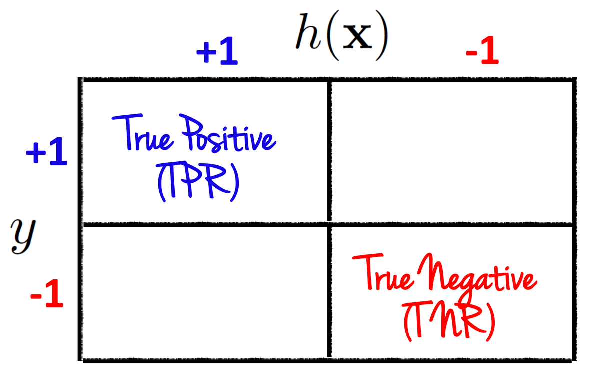


$\widehat{\eta}(\mathbf{x}) \approx \mathbf{P}(y=1\,|\,\mathbf{x})$ (using A)

$\widehat{t}$ (using B)

Our Work: **Statistical consistency of the plug-in method for a general non-decomposable performance measure**

Problem Setup



$$\mathcal{P}[h] = \Psi(\text{TPR}[h],\, \text{TNR}[h],\, p)$$

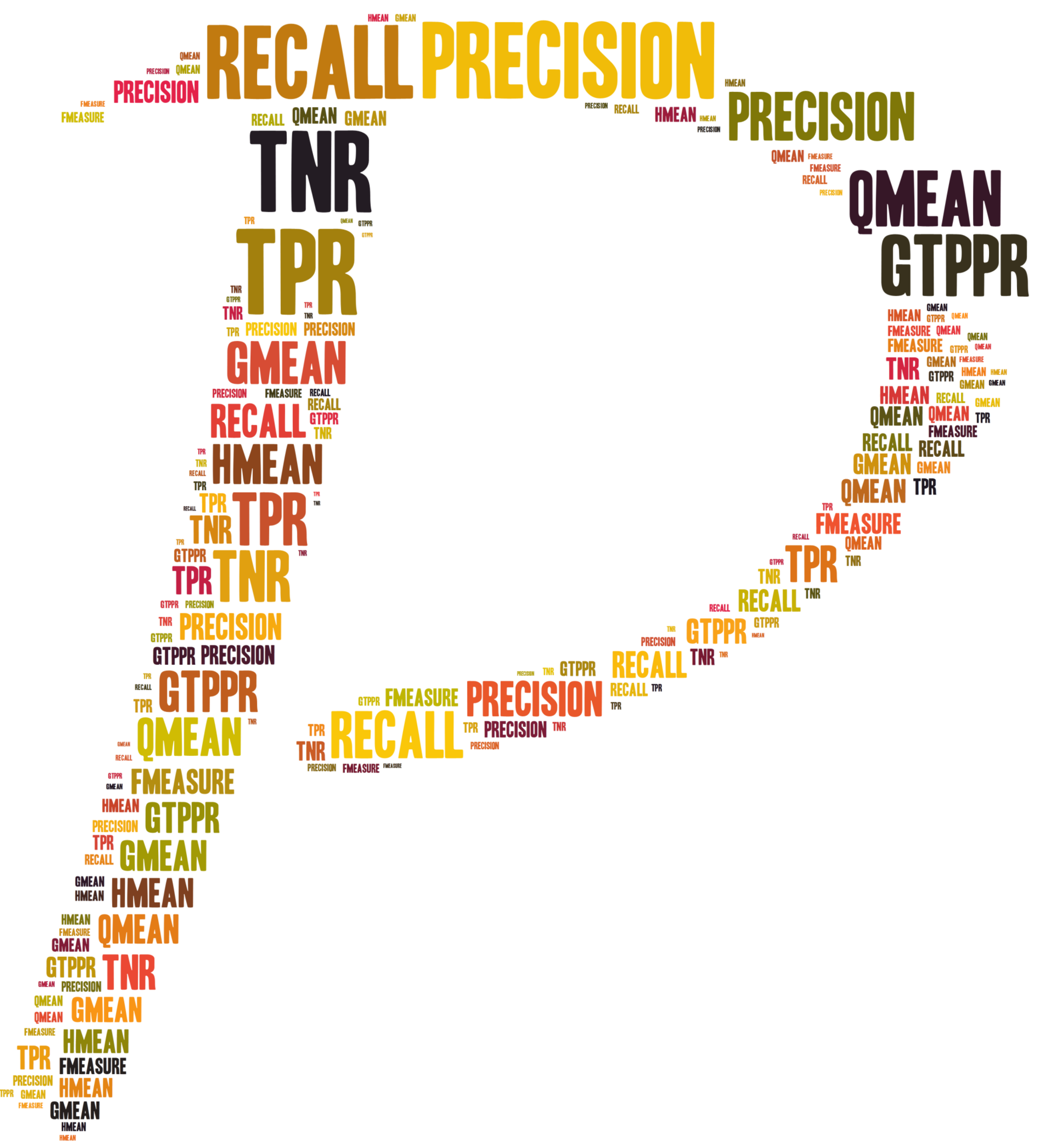
proportion of positives

Distribution 'D' over instances and labels

Statistical Consistency

An algorithm that learns $\widehat{h}$ using training sample $S \sim D^n$ is $\mathcal{P}$-consistent if

$$\mathcal{P}[\widehat{h}] \xrightarrow{P} \sup_{h:X\to\{\pm 1\}} \mathcal{P}[h] \quad (\text{as } n\to\infty).$$



Previous Work

Classification error: Well understood! Results crucially make use of decomposability of performance measure (E.g. Clemencon et al., 2008).

F-measure: Existing results assume an idealized setting where the true class probability is available to a learning algorithm (Ye et al., 2012) .

General non-decomposable performance measure?

Main Result

Given $\widehat{\eta}$ satisfies $\mathbf{E}_{\mathbf{x}}\big[|\widehat{\eta}(\mathbf{x})-\eta(\mathbf{x})|\big] \xrightarrow{P} 0$, under mild assumptions on D and $\mathcal{P}$, the plug-in method is $\mathcal{P}$-consistent if the optimal classifier for $\mathcal{P}$ is of the thresholded form $\text{sign}\circ(\eta - t^*)$ for some $t^* \in [0,1]$.

Quick Observation

Optimal classifier need not be of a thresholded form!

	$\mathbf{P}(x)$	$\eta(x)=\mathbf{P}(y=1\vert x)$
x_1	0.25	$1/2+\epsilon$
x_2	0.5	$1/2$
x_3	0.25	$1/2-\epsilon$

Example distribution over 3 instances

Optimal Classifier for G-mean: x_1 -1, x_2 +1, x_3 -1

Monotonic Measures & Continuous Distributions

For continuous distributions (CDF of $\eta(\mathbf{x})$ is continuous) and monotonic measures, the optimal classifier is *always of the thresholded form.*

Application to Popular Performance Measures

Measure	Definition	General D	Continuous D
AM (1-BER)	$(\text{TPR}+\text{TNR})/2$	✓	✓
F-measure	$2/(\frac{1}{\text{Prec}}+\frac{1}{\text{TPR}})$	✓	✓
G-TP/PR	$\sqrt{\text{TPR}\cdot\text{Prec}}$	✓	✓
G-Mean (GM)	$\sqrt{\text{TPR}\cdot\text{TNR}}$	✗	✓
H-Mean (HM)	$2/(\frac{1}{\text{TPR}}+\frac{1}{\text{TNR}})$	✗	✓
Q-Mean (QM)	$1-\frac{\text{FPR}^2+\text{FNR}^2}{2}$	✗	✓

Proof Idea for Main Result

→ $\mathbf{E}_{\mathbf{x}}\big[|\widehat{\eta}(\mathbf{x})-\eta(\mathbf{x})|\big] \xrightarrow{P} 0$ **implies for any fixed 'c'**

$$\text{TPR}[\text{sign}\circ(\widehat{\eta}-c)] \xrightarrow{P} \text{TPR}[\text{sign}\circ(\eta-c)] \quad (\text{as } n\to\infty)$$

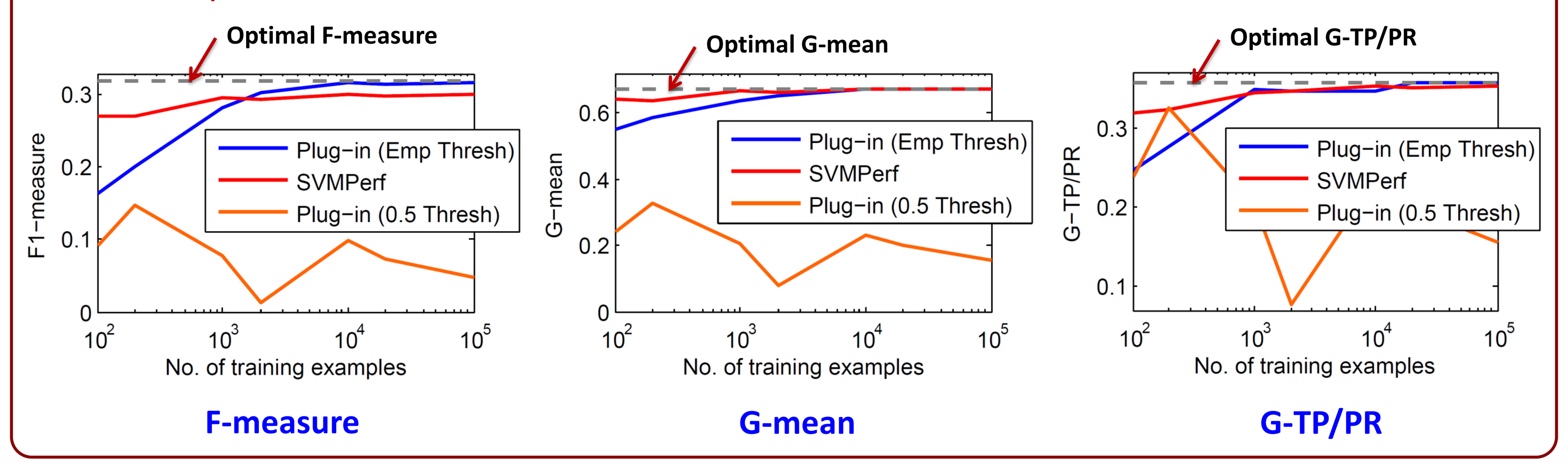
$$\text{TNR}[\text{sign}\circ(\widehat{\eta}-c)] \xrightarrow{P} \text{TNR}[\text{sign}\circ(\eta-c)] \quad (\text{as } n\to\infty)$$

→ **Uniform convergence generalization bound for** $\mathcal{P}$

Experiments

Synthetic data

- Gaussian class conditionals, equal covariance, p = 0.1
- Optimal classifier is linear



Reference

N. Ye, K.M.A. Chai, W.S. Lee, and H.L. Chieu. Optimizing F-measures: A tale of two approaches. In ICML 2012.