Nonparametric Estimation of the Precision-Recall Curve

Nicolas Vayatis (ENS Cachan, France)

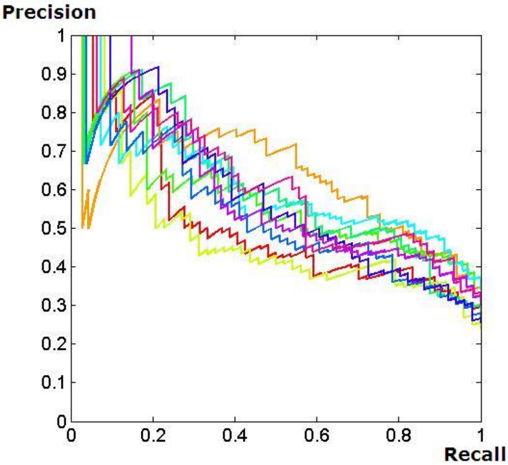
ICML - June 2009

Joint work with Stéphan Clémençon (Telecom ParisTech)

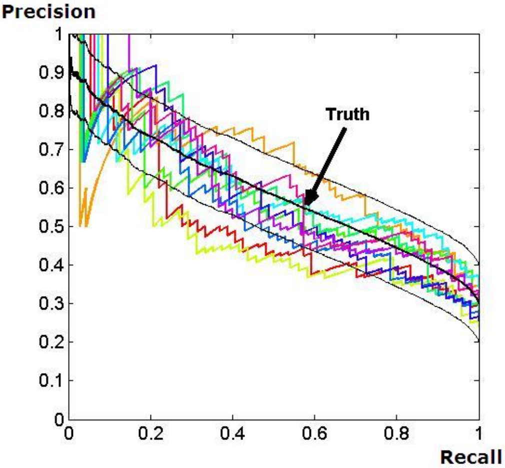
Context

- **Problem:** Bipartite ranking
- **Assume:** we have designed a scoring rule for ranking new data
- **Issue:** Performance assessment
- **Choice of a performance measure:** Precision and Recall

Variability of a ranking performance measure



Confidence bands for Precision-Recall curves?



Previous work

- Some work on estimation of the ROC curve:
  - [Hsieh and Turnbull, AOS 1996]
  - [Macskassy and Provost, ECAI 2004], and [M., P., and Rosset, ICML 2005]
  - [Bertail, Clémençon, and Vayatis, NIPS 2008]
  - [Horvath, Horvath, and Zhou, JSPI 2008]

Previous work

- Some work on estimation of the ROC curve:
  - [Hsieh and Turnbull, AOS 1996]
  - [Macskassy and Provost, ECAI 2004], and [M., P., and Rosset, ICML 2005]
  - [Bertail, Clémençon, and Vayatis, NIPS 2008]
  - [Horvath, Horvath, and Zhou, JSPI 2008]
- None on PR curves!

Motivations for using Precision-Recall

- Visual display of performance at various levels
- Justification: the optimal curve is above all the others

Motivations for using Precision-Recall

- Visual display of performance at various levels
- Justification: the optimal curve is above all the others
- ROC vs. Precision-Recall?

Motivations for using Precision-Recall

- Visual display of performance at various levels
- Justification: the optimal curve is above all the others
- ROC vs. Precision-Recall?
- ROC curves are independent of $p = \mathbb{P}\{Y = +1\}$
- PR curves best for highly skewed distributions (p small)

  (see Davis & Goadrich, ICML 2006)

Probabilistic model

- (Z, Y) random pair with unknown distribution P
- $Z \in \mathbb{R}$ pointwise score evaluation
- $Y \in \{-1, +1\}$ binary label/class
- Conditional distributions:

$$F_+(z) = \mathbb{P}\{Z \leq z \mid Y = +1\} \quad \text{and} \quad F_-(z) = \mathbb{P}\{Z \leq z \mid Y = -1\}$$

- Proportion: $p = \mathbb{P}\{Y = +1\}$
- Marginal distribution of Z:

$$F = pF_+ + (1 - p)F_-$$

True Precision-Recall curve

- Precision: $\mathbb{P}\{Z \geq t \mid Y = +1\}$
- Recall: $\mathbb{P}\{Y = +1 \mid Z \geq t\}$
- Definition of the PR curve:

$$\mathrm{PR} \ : \ t \in \mathbb{R} \mapsto (\mathbb{P}\{Z \geq t \mid Y = +1\}, \ \mathbb{P}\{Y = +1 \mid Z \geq t\}) \ ,$$

or

$$\mathrm{PR} \ : \ t \in \mathbb{R} \mapsto \left(1 - F_+(t), \ p\left(\frac{1 - F_+(t)}{1 - F(t)}\right)\right) \ .$$

Properties of the PR curve

- **Identical populations.** If $F_+ = F_-$ then $\mathrm{PR}(t) = (1 - F_+(t),\ p)$
- **Limits.**
  - $\lim_{t \to -\infty} \mathrm{PR}(t) = (1, p)$
  - $\lim_{t \to +\infty} \mathrm{PR}(t) = \left(0, \dfrac{p\ell}{p\ell + 1 - p}\right)$, where $\ell = \lim_{t \to +\infty} \dfrac{dF_+}{dF_-}(t)$
- **Monotonicity.**

  PR curve is decreasing if likelihood ratio dF_+/dF_- is monotone.

Reparameterization of the PR curve

- Conditional quantile function:

$$x \in [0,1] \mapsto (F_+)^{-1}(1-x)$$

- False positive rate at level x:

$$\alpha(x) = 1 - F_- \circ (F_+)^{-1}(1-x)$$

- PR curve as the plot of PR function:

$$\mathrm{PR} \ : \ x \in [0,1] \mapsto \frac{px}{px + (1-p)\alpha(x)} \ .$$

Empirical PR function

- Data: $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ i.i.d.
- Number of positives:

$$n_+ = \sum_{i=1}^{n} \mathbb{I}\{Y_i = +1\}$$

- Empirical false positive rate at x:

$$\widehat{\alpha}(x) = 1 - \widehat{F}_- \circ (\widehat{F}_+)^{-1}(1 - x)$$

- Empirical PR function:

$$\widehat{\mathrm{PR}}(x) = \frac{n_+ x}{n_+ x + (n - n_+)\widehat{\alpha}(x)} \; .$$

The PR fluctuation process

- Set $\widehat{\mathrm{PR}}$ to be the empirical PR function based on i.i.d. data
- Normalized PR fluctuation process:

$$R_n(x) = \sqrt{n}\left(\widehat{\mathrm{PR}}(x) - \mathrm{PR}(x)\right)$$

- Set $\epsilon > 0$ and consider $x \in [\epsilon, 1 - \epsilon]$

Technical assumptions

- Conditional distributions F_+ and F_- are equivalent and continuous
- For all $x \in (\epsilon, 1-\epsilon)$:
$$F'_+(F_+^{-1}(x)) > 0$$
- Tangent of $x \mapsto \alpha(x)$ is bounded, i.e.
$$\sup_{x \in [\epsilon, 1-\epsilon]} \frac{F'_- \circ F_+^{-1}(x)}{F'_+ \circ F_+^{-1}(x)} < \infty$$
- There exists $\gamma > 0$ such that:
$$\sup_{x \in (\epsilon, 1-\epsilon)} \frac{d}{dx} \log(F'_+ \circ F_+^{-1}(x)) \le \gamma < \infty \ .$$

Strong approximation result

Theorem 1

Under the previous assumptions, we have, almost surely, as $n \to \infty$:

(i) $$\sup_{x \in [\epsilon, 1-\epsilon]} |\widehat{\mathrm{PR}}(x) - \mathrm{PR}(x)| \to 0 \ ,$$

(ii) uniformly over $[\epsilon, 1-\epsilon]$: $R_n(x) = Z^{(n)}(x) + o\left(\frac{L(n,\gamma)}{\sqrt{n}}\right)$,

where

- $\{Z^{(n)}\}$ is a sequence of random processes with gaussian marginals and involves F_+, F_- and their derivatives
- $L(n,\gamma) = (\log\log n)^{\rho_1(\gamma)} (\log n)^{\rho_2(\gamma)}$

and $$\begin{cases} \rho_1(\gamma) = 0, \quad \rho_2(\gamma) = 1, & \text{if } \gamma < 1 \\ \rho_1(\gamma) = 0, \quad \rho_2(\gamma) = 2, & \text{if } \gamma = 1 \\ \rho_1(\gamma) = \gamma, \quad \rho_2(\gamma) = \gamma - 1 + \varepsilon, \ \varepsilon > 0, & \text{if } \gamma > 1. \end{cases}$$

Expression of the strong approximation

- Set $\{B_1^{(n)}\}$ and $\{B_2^{(n)}\}$ two independent sequences of brownian bridges on $[0,1]$
- Set W a gaussian r.v. independent from $\{B_1^{(n)}\}$, $\{B_2^{(n)}\}$
- Formula for $Z^{(n)}$:

$$Z^{(n)}(x) = \frac{\mathrm{PR}(x)^2}{x}\left(\alpha(x)\left(\sqrt{\frac{1-p}{p^3}}\right)W + \frac{1-p}{p^{3/2}}\left(\frac{F'_- \circ F_+^{-1}(x)}{F'_+ \circ F_+^{-1}(x)}\right)B_1^{(n)}(x) + \left(\frac{\sqrt{1-p}}{p}\right)B_2^{(n)}(\alpha(x))\right)$$

for some W, $\{B_1^{(n)}\}$ and $\{B_2^{(n)}\}$.

Is this helpful?

- **Want:** Confidence bands on the true PR
- **Got:** explicit gaussian limit distribution for R_n...

Is this helpful?

- **Want:** Confidence bands on the true PR
- **Got:** explicit gaussian limit distribution for R_n...
- ... but they depend on the unknown distribution!

Is this helpful?

- **Want:** Confidence bands on the true PR
- **Got:** explicit gaussian limit distribution for R_n...
- ... but they depend on the unknown distribution!
- **Idea:** Use bootstrap

Is this helpful?

- **Want:** Confidence bands on the true PR
- **Got:** explicit gaussian limit distribution for R_n...
- ... but they depend on the unknown distribution!
- **Idea:** Use bootstrap
- **Drawback:** Naive bootstrap for quantile estimation has a very slow rate of convergence

Bootstrap notations

- Set PR^* = empirical PR curve obtained on a bootstrap sample
- Bootstrapped PR fluctuation process:

$$R_n^* = \left\{ \sqrt{n}(\mathrm{PR}^*(x) - \widehat{\mathrm{PR}}(x)) \right\}_{x \in [\epsilon, 1-\epsilon]}$$

Repairing naive bootstrap

- Resampling from smoothed distributions: $\widehat{F}_{+/-} \rightarrow \widetilde{F}_{+/-}$

  $\longrightarrow$ use kernel smoothing

  $\longrightarrow$ e.g. gaussian kernel with bandwidth $h = h_n$

Repairing naive bootstrap

- Resampling from smoothed distributions: $\widehat{F}_{+/-} \rightarrow \widetilde{F}_{+/-}$

  $\rightarrow$ use kernel smoothing

  $\rightarrow$ e.g. gaussian kernel with bandwidth $h = h_n$

- Practical procedure:
  - Draw with replacement $(Z_1', Y_1^*), \ldots, (Z_n', Y_n^*)$ from $(Z_1, Y_1), \ldots, (Z_n, Y_n)$
  - Add an independent gaussian perturbation $\epsilon_j \sim \mathcal{N}(0, h^2)$ to each Z_j':

$$Z_j^* = Z_j' + \epsilon_j$$

  - Get bootstrap n-sample: $(Z_1^*, Y_1^*), \ldots, (Z_n^*, Y_n^*)$

Importance bootstrap confidence bands

- Importance sampling: use mixture parameter $\widetilde{p} \simeq 1/2$

  $\rightarrow$ use the importance function correction in the estimation

- Importance function:

$$\gamma_n = \left(\frac{n_+^*}{n\widetilde{p}}\right)^{n_+^*} \left(\frac{n - n_+^*}{n(1-\widetilde{p})}\right)^{n-n_+^*}$$

- Notations: $\mathbb{E}^*[.]$ expected value over bootstrap n-sample distribution
- Find $r(\delta)$ such that:

$$\mathbb{E}^* \left[\gamma_n \cdot \mathbb{I} \left\{ \sup_{x \in [\epsilon, 1-\epsilon]} |R_n^*(x)| \leq r(\delta) \right\} \right] = 1 - \delta$$

Bootstrap validity

Set:

- $H_{n,\epsilon}(r) = \mathbb{P}\left\{ \sup_{x\in[\epsilon,1-\epsilon]} |R_n(x)| \leq r \right\}$
- $H_{n,\epsilon}^{boot}(r) = \mathbb{E}^*\left[\gamma_n \cdot \mathbb{I}\left\{ \sup_{x\in[\epsilon,1-\epsilon]} |R_n^*(x)| \leq r \right\} \right]$

Theorem 2

Same assumptions as before. Take also: $h_n \simeq (n \log^3 n)^{-1/5}$. Then, we have as $n \to \infty$:

$$\sup_{r\in\mathbb{R}_+} |H_{n,\epsilon}(r) - H_{n,\epsilon}^{boot}(r)| = o_{\mathbb{P}}\left(n^{-2/5}\right) .$$

Last but not least - PR estimation and beyond!

- Promote statistical approach to machine learning concepts
- Statistical theory may be helpful
- PR curve learning still at an early stage!

Last but not least - PR estimation and beyond!

- Promote statistical approach to machine learning concepts
- Statistical theory may be helpful
- PR curve learning still at an early stage!
- Statistical theory for ROC curve learning - check our papers!
  - COLT'05, ALT'08, NIPS'08 (x 3), AISTAT'09
  - JMLR 2007, AOS 2008, IEEE IT (to app.)
  - ... and more to come!
- R package for ROC curve learning soon available!