Counterfactual risk minimization: Learning from logged bandit feedback

Adith Swaminathan and Thorsten Joachims
Department of Computer Science, Cornell University

Aim: Offline learning for interactive systems
Can we re-use the interaction logs of deployed online systems (e.g. search engines, recommendation systems) to train better models offline?

Training using interaction logs is counter-factual [2].
- Logs are biased (actions favored by deployed system will be over-represented),
- and incomplete (no feedback for other plausible actions).

Our contribution
A learning principle — Counterfactual Risk Minimization — and an efficient algorithm — Policy Optimizer for Exponential Models — for this learning setting [1]. Our solution is to
- predict by sampling and log propensities,
- use counterfactual risk estimators to fix bias,
- regularize the variance.
- and optimize a conservative bound using majorization minimization.

POEM
POEM is a simple, fast, stochastic optimizer for structured output prediction available at http://www.cs.cornell.edu/~adith/poem

It is as fast and expressive as Conditional Random Fields (CRFs), and trains using logged bandit feedback, without any supervised labels.

Counterfactual risk minimization
Learning from logged data without exploration is not possible. Suppose the deployed system sampled $y \sim h_0(y \mid x)$.

$\mathbb{E}_{h \sim \text{poem}(h)} [\delta(x, y)] = \mathbb{E}_{h \sim \text{poem}(h)} [\delta(x, y)] h(y \mid x) / h(y \mid x)$.

With $D = \{(x_i, y_i, \delta_i, p_i)\}_{i=1}^n$, $p_i = h_i(y_i \mid x_i)$,

$R(h) = \frac{1}{n} \sum_{i=1}^n \delta_i h(y_i \mid x_i)p_i$.

This unbiased estimator has issues:
- Unbounded variance (think $p_i \approx 0$).
- Degenerate minimizer (think $\delta_i \approx 0$).
- Importance sampling introduces variance.

Different effective sample sizes for different $h$!

Inverse propensity scoring,
$L_{\text{ips}}(h) = \min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n (\delta_i - \delta_{\text{true}}) \min \left\{ M, \frac{h(y_i \mid x_i)}{p_i} \right\}$.

Variance regularization
With high probability in $D \sim h_0$, $\forall h \in \mathcal{H}$,

$R(h) \leq R_M(h) + O\left(\frac{\text{Var}(u_w)}{n} + \frac{M \cdot \text{O}(h)}{n}\right)$.

Counterfactual risk minimization

$\mathcal{R}_{\text{CRM}} = \min_{h \in \mathcal{H}} R_M(h) + \lambda \frac{\text{Var}(u_w)}{n}$.

Deriving POEM from CRM

For CRFs, $h_w(y \mid x) \propto e^{w(x,y)}$

$w_{\text{CRM}} = \min_{w} - \frac{1}{n} \sum_{i=1}^n u_i + \lambda \frac{\text{Var}(u_w)}{n} + \mu \|w\|^2$.

To optimize at scale, we Taylor-approximate.

- Adagrad with $\nabla w_e \left\{ 1 + \lambda \sqrt{n} (A_w + 2B_w w_e) \right\}$
- After epoch, $w_{e+1} \leftarrow w$, compute $A_{w_{e+1}}, B_{w_{e+1}}$.

Experiments
Supervised $\rightarrow$ Bandit Multi-Label classification with $\delta \equiv$ Hamming loss on four datasets.

POEM is computationally efficient versus batch L-BFGS and compares favorably with CRF of scikit-learn.

<table>
<thead>
<tr>
<th>Avg. Time (s)</th>
<th>Scene</th>
<th>Yeast</th>
<th>TMC</th>
<th>LYRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>POEM(L-BFGS)</td>
<td>75.20</td>
<td>94.16</td>
<td>949.95</td>
<td>561.12</td>
</tr>
<tr>
<td>POEM</td>
<td>4.71</td>
<td>5.02</td>
<td>276.13</td>
<td>120.09</td>
</tr>
<tr>
<td>CRF</td>
<td>4.86</td>
<td>3.28</td>
<td>99.18</td>
<td>62.93</td>
</tr>
</tbody>
</table>

POEM is statistically significantly better ($p = 0.05$) than IPS and $h_0$ (CRF trained on 5% of train set) on all datasets.

References

Acknowledgment
This research was funded through NSF Award IIS-1247637, IIS-1217686, FTCH Cornell-Technion Research Fund, and a gift from Bloomberg.