Surrogate Objectives for Batch Policy Optimization in One-step Decision Making

Minmin Chen[1], Ramki Gummadi[1], Chris Harris[1], Dale Schuurmans[1,2]

COST SENSITIVE CLASSIFICATION

Batch policy optimization

Assume given **complete** data

	a_1	a_2	…	a_n
x_1	r_{11}	r_{12}	$r_{1\ldots}$	r_{1n}
x_2	r_{21}	r_{22}	$r_{2\ldots}$	r_{2n}
x_3	r_{31}	r_{32}	$r_{3\ldots}$	r_{3n}
x_4	r_{41}	r_{42}	$r_{4\ldots}$	r_{4n}
x_5	r_{51}	r_{52}	$r_{5\ldots}$	r_{5n}
x_6	r_{61}	r_{62}	$r_{6\ldots}$	r_{6n}
⋮	$r_{:1}$	$r_{:2}$	$r_{:\ldots}$	$r_{:n}$
x_m	r_{m1}	r_{m2}	$r_{m\ldots}$	r_{mn}

Optimize policy $\pi : X \to \Delta^n$

$\pi(a\,|\,x) = e^{q(x)_a - F(q(x))}$

$F(q(x)) = \log \sum_a e^{q(x)_a}$

$q : X \to \Re^n$ neural network

Target objective

- expected reward: $\max \sum_i \mathbf{r}_i \cdot \boldsymbol{\pi}(x_i)$

Done, right?
Not so fast …

This objective has serious problems

- actually trying to solve: $\max \sum_i \mathbf{r}_i \cdot \mathbf{f}(q(x_i))$
- plateaus everywhere
- can have **exponentially many** local maxima
- nearly impossible to reach a global optima

Also: you already know not to train this way!

to maximize expected reward on **test** contexts

Recall: supervised classification

Special case: **supervised classification**

	a_1	a_2	…	a_n
x_1	0	1	0	0
x_2	0	0	0	1
x_3	1	0	0	0
x_4	0	0	1	0
x_5	1	0	0	0
x_6	0	1	0	0
⋮	0	0	1	0
x_m	0	0	0	1

Target objective

- expected **accuracy**: $\max \sum_i \mathbf{r}_i \cdot \boldsymbol{\pi}(x_i)$

But you have never trained with this objective
Instead, you used a **surrogate objective**

maximum likelihood
$\max \sum_i \mathbf{r}_i \cdot \log \boldsymbol{\pi}(x_i)$

What's going on?

- $\mathbf{r}_i \cdot \boldsymbol{\pi}(x_i)$ is differentiable, that's not the issue
- training with $\mathbf{r}_i \cdot \log \boldsymbol{\pi}(x_i)$ actually achieves better values of $\mathbf{r}_i \cdot \boldsymbol{\pi}(x_i)$ on the training data

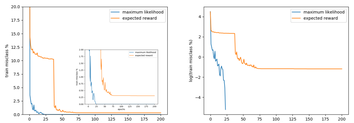
Useful properties of maximum likelihood

- $\mathbf{r}_i \cdot \log \boldsymbol{\pi}(x_i)$ is concave in $\mathbf{q}(x_i)$
- it is also calibrated w.r.t. $\mathbf{r}_i \cdot \boldsymbol{\pi}(x_i)$:

$\forall c > 0 \exists \delta > 0 \;\; \mathbf{r} \cdot \log \boldsymbol{\pi}^* - \mathbf{r} \cdot \log \boldsymbol{\pi} < \delta \Rightarrow \mathbf{r} \cdot \boldsymbol{\pi}^* - \mathbf{r} \cdot \boldsymbol{\pi} < \epsilon$

Target vs surrogate optimization

Misclassification error on MNIST training data



COST SENSITIVE CLASSIFICATION

Definitions

- Data $\mathcal{D} = \{(x_i, \mathbf{r}_i)\}_{i=1}^T$, where $\mathbf{r}_i \in \mathbb{R}^K$ specifies reward for each action in context x_i
- *True risk* of a policy is $\mathcal{R}(\boldsymbol{\pi}) = -\mathbb{E}[\boldsymbol{\pi}(x) \cdot \mathbf{r}]$
- *Empirical risk* on data set $\mathcal{D}$ is $\hat{\mathcal{R}}(\boldsymbol{\pi}, \mathcal{D}) = -\frac{1}{T}\sum_{(x_i,\mathbf{r}_i)\in\mathcal{D}} \boldsymbol{\pi}(x_i) \cdot \mathbf{r}_i$

Note policies normally represented with composition $\boldsymbol{\pi}(x) = \mathbf{f}(q(x))$ where $\mathbf{f}(\mathbf{q}) = e^{\mathbf{q} - F(\mathbf{q})}$ with $F(\mathbf{q}) = \log(\mathbf{1} \cdot e^{\mathbf{q}})$

Theorem Even for a linear model $q(x) = W\phi(x)$, the function $\mathbf{r} \cdot \mathbf{f}(q(x))$ can have exponentially many local maxima in W

Motivation To get around this problem, need to consider *surrogate* training objectives

Calibrated convex surrogate

Definitions

- *Minimal risk* is $\mathcal{R}^*(\mathbf{r}, x) = \inf_{\boldsymbol{\pi}\in\mathcal{P}} \mathcal{R}(\boldsymbol{\pi}, \mathbf{r}, x) = \inf_{\mathbf{q}\in\mathcal{Q}} \mathcal{R}(\mathbf{f} \circ \mathbf{q}, \mathbf{r}, x)$
- Loss $L^*(\mathbf{r}, x) = \inf_{\mathbf{q}\in\mathcal{Q}} L(\mathbf{q}, \mathbf{r}, x)$ is *calibrated* w.r.t. $\mathcal{R}$ if: $\exists$ function $\delta(\epsilon, x) \geq 0$ s.t. $\forall\, \epsilon > 0,\ x \in \mathcal{X},\ \mathbf{r} \in \mathbb{R}^K,\ \mathbf{q} \in \mathcal{Q}$:

$$L(\mathbf{q}, \mathbf{r}, x) - L^*(\mathbf{r}, x) \;<\; \delta(\epsilon, x) \Rightarrow \mathcal{R}(\mathbf{f} \circ \mathbf{q}, \mathbf{r}, x) \;<\; \mathcal{R}^*(\mathbf{r}, x) + \epsilon$$

- *Smoothed risk* is

$$\mathcal{S}_\tau(\boldsymbol{\pi}, \mathbf{r}, x) = -\mathbf{r} \cdot \boldsymbol{\pi}(x) + \tau\boldsymbol{\pi}(x) \cdot \log \boldsymbol{\pi}(x)$$

- Let $\tilde{\boldsymbol{\pi}}_\tau = \arg\min_{\boldsymbol{\pi}\in\mathcal{P}} \mathcal{S}_\tau(\boldsymbol{\pi})$

Proposition $\tau < \epsilon / \log K$ implies $\mathcal{R}(\tilde{\boldsymbol{\pi}}_\tau) < \mathcal{R}^* + \epsilon$

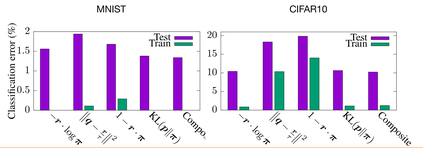
Proposition Local smoothed risk is equivalent to

$$\mathcal{S}_\tau(\boldsymbol{\pi}, \mathbf{r}, x) = -\tau F(\tfrac{\mathbf{r}}{\tau}) + \tau D_F\left(\tfrac{\mathbf{r}}{\tau} \| \mathbf{q}(x)\right)$$

Theorem The surrogate objective $L(\mathbf{q}, \mathbf{r}, x) = \tau D_F\left(\mathbf{q}(x) + \frac{\nu}{\tau}\left\|\tfrac{\mathbf{r}}{\tau}\right) + \frac{\tau}{4}\left\|\mathbf{q}(x) - \frac{\mathbf{r}-\nu}{\tau}\right\|^2$. is strongly convex in $\mathbf{q}$ and *calibrated* w.r.t. $\mathcal{S}_\tau(\mathbf{f} \circ \mathbf{q}, \mathbf{r} - \nu, x)$ with $\delta(\epsilon, x) = \epsilon$

Experimental evaluation

Comparing objectives



BATCH CONTEXTUAL BANDITS

Coping with missing data

	a_1	a_2	…	a_n
x_1	0	$\frac{r_2}{\beta_1}$	0	0
x_2	0	0	0	$\frac{r_2}{\beta_2}$
x_3	$\frac{r_3}{\beta_3}$	0	0	0
x_4	0	0	$\frac{r_4}{\beta_4}$	0
x_5	$\frac{r_5}{\beta_5}$	0	0	0
x_6	0	$\frac{r_6}{\beta_6}$	0	0
⋮	0	0	$\frac{r_:}{\beta_:}$	0
x_m	0	0	0	$\frac{r_m}{\beta_m}$

Optimize policy $\pi : X \to \Delta^n$

Example
importance corrected expected reward

$$\max \sum_i \frac{\pi(a_i\,|\,x_i)}{\beta_i} r_i$$

where β are proposal probabilities from behavior strategy

We already know this is a poor objective but what about missing data inference?

Equivalent to $\max \hat{\mathbf{r}} \cdot \boldsymbol{\pi}$ using $\hat{\mathbf{r}}_i = \mathbf{1}_{a_i} \frac{r_i}{\beta_i}$

That is

- exaggerate observed values by $1/\beta_i$
- fill in all unobserved values with 0

This is a pretty lame inference principle
But … its unbiased!

$$\mathbb{E}[\hat{\mathbf{r}}\,|\,x] = \sum_a \beta_a \mathbf{1}_a \tfrac{r_a}{\beta_a} = \sum_a \mathbf{1}_a r_a = \mathbf{r}$$

Missing data inference

	a_1	a_2	…	a_n
x_1	τq_{11}	$\tau q_{12} + \lambda(r_{12} - \tau q_{12})$		τq_{1n}
x_2	τq_{21}	τq_{22}		$\tau q_{2n} + \lambda(r_{2n} - \tau q_{2n})$
x_3	$\tau q_{31} + \lambda(r_{31} - \tau q_{31})$	τq_{32}		τq_{3n}
x_4	τq_{41}	τq_{42}	$\tau q_{4\ldots} + \lambda(r_{4\ldots} - \tau q_{4\ldots})$	τq_{4n}
x_5	$\tau q_{51} + \lambda(r_{51} - \tau q_{51})$	τq_{52}		τq_{5n}
x_6	τq_{61}	$\tau q_{62} + \lambda(r_{62} - \tau q_{62})$		τq_{6n}
⋮	$\tau q_{:1}$	$\tau q_{:2}$	$\tau e_{:\ldots} + \lambda(r_{:\ldots} - \tau e_{:\ldots})$	$\tau q_{:n}$
x_m	τq_{m1}	τq_{m2}		$\tau q_{mn} + \lambda(r_{mn} - \tau q_{mn})$

Improvement
"doubly robust estimation"

- instead of filling in with 0s
- fill in with guesses from a model $\mathbf{q}(x)$

$$\hat{\mathbf{r}} = \tau\mathbf{q} + \lambda \mathbf{1}_a (r - \tau q_a)$$

Also unbiased

- as long as $\lambda = 1/\beta_i$

but still alters observed data

Where should the model come from?

- could use a separate critic
- train via least squares, then optimize $\boldsymbol{\pi}$
- works okay, but not great

Note

- there is only one action value function for single-step decision making, $r(x, a)$
- actor-critic approaches trivialized

Unified approach

Unified approach

- actor and critic are same model
- $\boldsymbol{\pi} = e^{\mathbf{q} - F(\mathbf{q})}$ where $F(\mathbf{q}) = \log \mathbf{1} \cdot e^{\mathbf{q}}$
- use logits $\tau\mathbf{q}(x)$ to predict rewards $q(x, a) \approx \frac{r(x,a)}{\tau}$

Can combine with previous objectives

- $\mathbf{KL}(\boldsymbol{\pi}\|\hat{\mathbf{p}})$ where $\hat{\mathbf{p}} = e^{\hat{\mathbf{r}}/\tau - \hat{F}(\hat{\mathbf{r}}/\tau)}$
- $\mathbf{KL}(\hat{\mathbf{p}}\|\boldsymbol{\pi})$
- $\mathbf{KL}(\boldsymbol{\pi}\|\hat{\mathbf{p}}) \leq \mathbf{KL}(\hat{\mathbf{p}}\|\boldsymbol{\pi}) + \frac{\tau}{4}\|\hat{\mathbf{r}}/\tau - \mathbf{q}\|^2$

these are somewhat sensitive to ranking, unlike least squares

Empirical Bayes estimation

- optimize hyperparameters $\mathbf{q}$ (neural network)
- integrate out parameters ξ

Example
marginal likelihood

$$\begin{aligned} &-\log p(r_0\,|\,a_0, \mathbf{q}) \\ &= -\log \textstyle\int p(r_0\,|\,a_0, \xi) p(\xi\,|\,\mathbf{q})\, d\xi \\ &= \frac{1}{2\sigma^2}(\phi(a_0) \cdot q - r_0)^2 + \frac{1}{2}\log\sigma^2 + c \end{aligned}$$

- essentially least squares regression

Can alternatively use surrogates
$\min \mathbf{KL}(\text{prior}\|\text{posterior})$
$\min \mathbf{KL}(\text{posterior}\|\text{prior}) \approx \min I(\xi; r_0)$

BATCH CONTEXTUAL BANDITS

Reward estimation For x, a, r_a, parameters $\lambda(x, a)$, τ, estimate

$$\text{full } \hat{r}(x) = \tau\mathbf{q}(x) + \mathbf{1}_a \lambda(x, a)(r_a - \tau\mathbf{q}(x)_a)$$

Surrogate objective

Definition Optimal imputed local risk and suboptimality gap

$$\mathcal{S}_\tau^*(\hat{\mathbf{r}}, x) = \inf_{\mathbf{q}\in\mathcal{Q}} \mathcal{S}_\tau(\mathbf{f} \circ \mathbf{q}, \hat{\mathbf{r}}, x),\ \mathcal{G}_\tau(\boldsymbol{\pi}, \hat{\mathbf{r}}, x) = \mathcal{S}_\tau(\boldsymbol{\pi}, \hat{\mathbf{r}}, x) - \mathcal{S}_\tau^*(\hat{\mathbf{r}}, x)$$

Proposition $\forall\ \mathbf{q},\ \tau > 0,\ (x, a, r_a) :\ \tau D_F\left(\frac{\hat{r}(x)}{\tau}\middle\|\mathbf{q}(x)\right) = \mathcal{G}_\tau(\mathbf{f} \circ \mathbf{q}, \hat{r}, x)$

Theorem $\forall \mathbf{q},\ \tau > 0,\ (x, a, r_a),\ v$:

$$L(\mathbf{q}, \hat{r}, x) \geq \tau D_F\left(\frac{\hat{r}(x)}{\tau}\middle\|\mathbf{q}(x) + \frac{v}{\tau}\right) = \mathcal{G}_\tau(\mathbf{f} \circ \mathbf{q}, \hat{r}, x) \geq 0$$

L calibrated w.r.t. $\mathcal{S}_\tau(\mathbf{f} \circ \mathbf{q}, \hat{\mathbf{r}} - v, x)$

Optimization Given $\mathcal{D} = \{(x_i, a_i, r_i, \beta_i)\}$, context, act, reward, prob

$$\min_{q\in\mathcal{Q}} \hat{L}(\mathbf{q}, \mathcal{D}) \quad \text{where} \quad \hat{L}(\mathbf{q}, \mathcal{D}) = \tfrac{1}{T}\textstyle\sum_{(x_i,a_i,r_i,\beta_i)\in\mathcal{D}} L(\mathbf{q}, \hat{\mathbf{r}}, x_i)$$

Analysis

Definition Expected smoothed risk quantities we seek to control:

$$\mathcal{S}_\tau(\boldsymbol{\pi}) = \mathbb{E}[\mathcal{S}_\tau(\boldsymbol{\pi}, \mathbf{r}, x)], \mathcal{S}_\tau^* = \inf_{\mathbf{q}\in\mathcal{Q}} \mathcal{S}_\tau(\mathbf{f} \circ \mathbf{q}), \mathcal{G}_\tau(\boldsymbol{\pi}) = \mathcal{S}_\tau(\boldsymbol{\pi}) - \mathcal{S}_\tau^*$$

Theorem For any $\mathbf{q}$, $\hat{\mathbf{r}}$ such that $\mathbb{E}[\hat{\mathbf{r}}|x] = \mathbb{E}[\mathbf{r}|x]$, and baseline v:

$$\mathbb{E}[L(\mathbf{q}, \hat{\mathbf{r}}, x)] \geq \mathbb{E}\left[\tau D_F\left(\frac{\hat{r}(x)}{\tau}\middle\|\mathbf{q}(x) + \frac{v}{\tau}\right)\right] \;\geq\; \mathcal{G}_\tau(\mathbf{f} \circ \mathbf{q}) \;\geq\; 0.$$

Lemma $\forall\ \tau, \delta > 0\ \exists$ constant C s.t. w.p. at least $1 - \delta$:

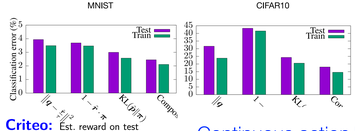
$$\mathbb{E}\left[D_F\left(\frac{\hat{r}(x)}{\tau}\middle\|\mathbf{q}(x)\right)\right] \leq \hat{D}_F(\mathbf{q}, \mathcal{D}) + \frac{C}{\sqrt{T}} \quad \forall \mathbf{q} \in \mathcal{H}.$$

Theorem $\forall\ v,\ \tau, \delta > 0,\ \exists\ C$ s.t. w.p. at least $1 - \delta$:

$$\text{if } \hat{L}(\mathbf{q}, \mathcal{D}) < \tfrac{\tau C}{\sqrt{T}} \text{ for } \mathbf{q} \in \mathcal{H} \text{ then } \mathcal{G}_\tau(\mathbf{f} \circ \mathbf{q}) \leq \tfrac{2\tau C}{\sqrt{T}}$$

Experimental evaluation

Comparing objectives

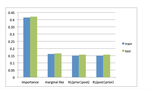


Criteo: Est. reward on test

Objectives	$\hat{\mathcal{R}}(\tau) \times 10^4$
Random	43.68 ± 2.11
Behavior	53.55
DRO $\hat{\mathcal{R}}(\boldsymbol{\pi})$	53.07 ± 2.27
POEM	51.89 ± 1.73
$\hat{\mathcal{R}}(\boldsymbol{\pi})$	51.72 ± 1.42
$\|\mathbf{q} - \frac{\mathbf{r}-\nu}{\tau}\|^2$	52.00 ± 1.28
$D_F(\mathbf{p}\|\boldsymbol{\pi})$	52.30 ± 0.83
Composite	55.09 ± 2.86

Continuous action MNIST

Sum of squared test error on continuous action MNIST ($a \in \Re^{10}$)



[1]Google Brain, [2]University of Alberta

Mailto: schuurmans@google.com