Counterfactual Risk Minimization

Adith Swaminathan
Cornell University
Ithaca, NY 14850
adith@cs.cornell.edu

Thorsten Joachims
Cornell University
Ithaca, NY 14850
tj@cs.cornell.edu

1. INTRODUCTION

We develop a learning principle and an efficient algorithm for batch learning from logged bandit feedback. Unlike in supervised learning, where the algorithm receives training examples (x_i, y_i^*) with annotated correct labels y_i^*, bandit feedback merely provides a cardinal reward $\delta_i \in \mathbb{R}$ for the prediction y_i that the logging system made for context x_i. Such bandit feedback is ubiquitous in online systems (e.g. observing a click δ_i on ad y_i for query x_i), while "correct" labels (e.g. the best ad y_i^* for query x_i) are difficult to assess.

Our work builds upon recent approaches to the off-policy evaluation problem [5], [8], [7], [9], where we re-use data collected from the interaction logs of one bandit algorithm to evaluate another system. These approaches use counterfactual reasoning [3] to derive an unbiased estimate of the system's performance. Our work centers around the insight that, to perform robust learning, it is not sufficient to have just an unbiased estimate of system performance. We must also reason about how the variances of these estimators differ across the hypothesis space, and pick the hypothesis with the tightest conservative bound on system performance.

We first derive generalization error bounds analogous to structural risk minimization [15] for a *stochastic hypothesis* family. The constructive nature of these bounds suggests a general principle – Counterfactual Risk Minimization (CRM) – for designing methods for batch learning from bandit feedback. Using the CRM principle, we derive a new learning algorithm – Policy Optimizer for Exponential Models (POEM) – for structured output prediction. We evaluate POEM on several multi-label classification problems and verify that its empirical performance supports the theory.

Existing approaches for batch learning from logged bandit feedback fall into two categories. The first approach reduces the problem to supervised learning, using techniques like cost weighted classification [16] or the Offset Tree algorithm [2] when the space of possible predictions is small. In contrast, our approach generalizes structured output prediction with exponential-sized prediction spaces.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742564.

The second approach uses propensity scoring [12], [13], [3] to derive unbiased estimators from the logged data. These estimators are used for a small set of candidate policies, and the best estimated candidate is picked via exhaustive search. In contrast, our approach can be optimized via gradient descent, over hypothesis families of infinite size. The current work focuses on the inverse propensity scoring estimator (which assumes we have a good model of the logging system [7], [9]), but the results we derive hold equally for doubly robust estimators (which are more efficient when we additionally have a good model of user feedback [5]). In the current work, we concentrate on the case where the logging system was a stationary, stochastic policy. Techniques like exploration scavenging [6] and bootstrapping [10] allow us to perform counterfactual evaluation even when the logging system was deterministic or adaptive.

A full version of this abstract can be found on arXiv [14].

2. METHOD

To formalize the problem of batch learning from bandit feedback, consider a structured prediction problem that takes as input $x \in \mathcal{X}$ and outputs a prediction $y \in \mathcal{Y}$. In multi-label document classification, x could be a news article and y a bitvector indicating the labels assigned to this article. The inputs are assumed drawn i.i.d. from a fixed but unknown distribution $\Pr(\mathcal{X})$.

Consider the hypothesis space $\mathcal{H}$ of *stochastic policies*. A hypothesis $h(\mathcal{Y} \mid x) \in \mathcal{H}$ defines a probability distribution over the output space $\mathcal{Y}$, and the hypothesis makes predictions by *sampling*, $y \sim h(\mathcal{Y} \mid x)$. Note that stochastic policy families include deterministic rules. We denote $h(\mathcal{Y} \mid x)$ by $h(x)$, and the probability assigned by $h(x)$ to y as $h(y \mid x)$.

In interactive learning systems, we only observe feedback $\delta(x, y)$ for the y presented to the user, but not for any of the other possible predictions $\mathcal{Y} \setminus y$. In this work, feedback $\delta : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ is a cardinal loss. Small values for $\delta(x, y)$ indicate user satisfaction with y for x, while large values indicate dissatisfaction. The expected loss – called risk – of a hypothesis $R(h)$ is defined as,

$$R(h) = \mathbb{E}_{x \sim \Pr(\mathcal{X})} \mathbb{E}_{y \sim h(x)} \left[\delta(x, y) \right]. \tag{1}$$

The goal of the system is to find a hypothesis $h \in \mathcal{H}$ that has minimum risk (i.e., provides maximum user satisfaction).

To achieve this goal, we wish to use past interaction logs of the system for batch learning. Assume that the logging system acted according to a *stationary* policy $h_0(x)$ with full support over $\mathcal{Y}$. The data collected from this system is

$$\mathcal{D} = \{(x_1, y_1, \delta_1, p_1), \ldots, (x_n, y_n, \delta_n, p_n)\}, \tag{2}$$

where $y_i \sim h_0(x_i)$, $\delta_i \equiv \delta(x_i, y_i)$, and $p_i \equiv h_0(y_i \mid x_i)$ is the propensity of the logging system to predict y_i for input x_i.

The Inverse Propensity Scoring approach [5], [7] uses $\mathcal{D}$ to Monte Carlo approximate $R(h)$ via importance sampling,

$$\hat{R}(h) = 1/n \sum_{i=1}^{n} \delta_i h(y_i \mid x_i)/p_i. \quad (3)$$

At first thought, one may think that directly estimating $\hat{R}(h)$ over $h \in \mathcal{H}$ and picking the empirical minimizer is a valid learning strategy. Unfortunately, there are several potential pitfalls.

First, this estimator has unbounded variance, since $p_i \simeq 0$ in $\mathcal{D}$ can cause $\mathbb{E}_{\mathcal{D}}\left[\hat{R}(h)\right]$ to be arbitrarily far away from the true risk $R(h)$. This problem can be fixed by "clipping" the importance sampling weights [4].

Second, importance sampling typically estimates $\hat{R}(h)$ of different hypotheses $h \in \mathcal{H}$ with vastly different variances. Consider two hypotheses h_1 and h_2, where h_1 is similar to h_0, but where h_2 samples predictions that were not well explored by h_0. Importance sampling gives us low-variance estimates for $\hat{R}(h_1)$, but highly variable estimates for $\hat{R}(h_2)$.

For a bounded $\delta(\cdot,\cdot) \in [\nabla, \triangle]$, define ${u_h}^i$, $\boldsymbol{Var}_h(u)$, $\hat{R}^M(h)$,

$$\begin{aligned} {u_h}^i &\equiv (\delta_i - \textstyle\bigtriangleup) \min\{M, h(y_i \mid x_i)/p_i\}, \\ \boldsymbol{Var}_h(u) &\equiv \frac{1}{n(n-1)} \sum_{i,j=1}^{n} \frac{({u_h}^i - {u_h}^j)^2}{2}, \\ \hat{R}^M(h) &\equiv \sum_{i=1}^{n} {u_h}^i/n, \end{aligned} \quad (4)$$

where $M > 0$ is a hyper-parameter chosen to trade-off bias and variance in the importance sampling estimate. Using an empirical Bernstein argument [11], we prove that w.h.p.,

$$\forall h \in \mathcal{H} \ : \ R(h) \leq \hat{R}^M(h) + \mathcal{O}\left(\sqrt{\boldsymbol{Var}_h(u)/n}\right). \quad (5)$$

Full details of the proof (which requires us to define a notion of capacity for stochastic hypothesis classes), is given in [14].

This bound motivates the *Counterfactual Risk Minimization* principle: jointly optimize the estimate $\hat{R}^M(h)$ as well as its empirical standard deviation, where the latter serves as a *data-dependent regularizer*.

$$\hat{h}^{CRM} = \operatorname*{argmin}_{h \in \mathcal{H}} \left\{ \hat{R}^M(h) + \lambda \sqrt{\frac{\boldsymbol{Var}_h(u)}{n}} \right\}. \quad (6)$$

$M > 0$ and $\lambda \geq 0$ are regularization hyper-parameters that can be selected via validation.

Using the CRM principle, we now derive a learning algorithm called POEM (Policy Optimizer for Exponential Models) for structured output prediction using stochastic hypotheses of the form

$$h_w(y \mid x) = \exp(w \cdot \phi(x, y))/\mathbb{Z}(x), \quad (7)$$

where $\mathbb{Z}(x) = \sum_{y' \in \mathcal{Y}} \exp(w \cdot \phi(x, y'))$ is the partition function. These models are the stochastic (soft-max) extension of (hard-max) Structured SVMs, where w is a $d-$dimensional weight vector and $\phi(x, y)$ is a $d-$dimensional joint feature map. For example, in multi-label document classification, for a news article x and a possible assignment of labels y represented as a bitvector, $\phi(x, y)$ could simply be a concatenation of the bag-of-words features of the document ($\overline{x}$), one copy for each of the assigned labels in y, $\overline{x} \otimes y$.

The CRM principle gives rise to the following training objective for this hypothesis space.

$$w^* = \operatorname*{argmin}_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} {u_w}^i + \lambda \sqrt{\frac{\boldsymbol{Var}_w(u)}{n}} + \mu \|w\|^2. \quad (8)$$

While the objective in Equation (8) is not convex in w (even for $\lambda = 0$), we find that gradient descent is sufficient to find a good local optimum that generalizes well. For efficient large-scale training, we derived a stochastic gradient algorithm that uses repeated linear variance majorization (see [14]).

3. EXPERIMENTS

Consider multi-label classification with input $x \in \mathbb{R}^p$ and prediction $y \in \{0, 1\}^q$ using the feature map $\phi(x, y) = x \otimes y$. We conducted experiments on several multi-label datasets from the LibSVM repository, summarized in Table 1.

Table 1: Corpus statistics for multi-label datasets. Top-level categories used for LYRL.

Name	p(# features)	q(# labels)	n_{train}	n_{test}
Scene	294	6	1211	1196
Yeast	103	14	1500	917
TMC	30438	22	21519	7077
LYRL	47236	4	23149	781265
Media	120	101	30993	12914

For all datasets, we kept aside 25% of n_{train} as validation set (for hyper-parameter selection) and treat the rest as the training set. We employ the Supervised $\mapsto$ Bandit conversion experiment methodology [1]. A supervised dataset $\mathcal{D} = \{(x_1, y_1^*) \ \ldots \ (x_n, y_n^*)\}$ is used to simulate a bandit feedback dataset from a logging policy h_0 by sampling $y_i \sim h_0(x_i)$ and collecting feedback $\Delta(y_i^*, y_i)$. For h_0, in principle, we could use any arbitrary stochastic policy with full support on $\mathcal{Y}$. We choose a CRF (Conditional Random Field) trained on 10% of the training set using default hyper-parameters as h_0. $\Delta(y^*(x), y)$ here is the Hamming loss between the supervised label y^* vs. the sampled label y for input x. Note that using the doubly robust policy optimization scheme [5] will involve an Offset Tree reduction [2] that is exponential in q and is intractable for even moderate number of labels. After learning weight vectors (e.g. w_{crm}) on $\mathcal{D}$, we report the expected loss per test instance $\hat{R}(w) = \frac{1}{n_{test}} \sum_i \mathbb{E}_{y \sim h_w(x_i)} \Delta(y_i^*, y)$.

The expected Hamming loss of h_0 is the baseline to beat. Lower loss is better. The objective in Equation (8) with $\lambda = 0$ is indicative of the state-of-the-art for counterfactual learning (optimized via L-BFGS and reported as IPS), and Equation (8) optimized using SGD (see [14]) is reported as POEM. We also report results for a supervised CRF trained on the entire training set as a skyline in Table 2, despite its unfair advantage of having access to the supervised labels.

Table 2: Test set Hamming loss for multi-label classification datasets.

	Scene	Yeast	TMC	LYRL	Media
h_0	1.472	5.189	3.292	1.400	8.639
IPS	1.103	4.620	2.133	1.082	12.125
POEM	**1.054**	**3.957**	**2.032**	**1.018**	**3.641**
CRF	0.631	2.795	1.194	0.223	3.100

Across different datasets, the data-dependent variance regularizer of POEM consistently finds a hypothesis that generalizes better. For an extensive empirical analysis, see [14].

This research was funded in part through NSF Awards IIS-1247637 and IIS-1217686, the JTCII Cornell-Technion Research Fund, and a gift from Bloomberg.

4. REFERENCES

[1] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1638–1646, 2014.

[2] A. Beygelzimer and J. Langford. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 129–138, 2009.

[3] L. Bottou, J. Peters, J. Q. Candela, D. X. Charles, M. Chickering, E. Portugaly, D. Ray, P. Y. Simard, and E. Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

[4] E. L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

[5] J. Langford, L. Li, and M. DudÃ■k. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1097–1104, 2011.

[6] J. Langford, A. Strehl, and J. Wortman. Exploration scavenging. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 528–535, 2008.

[7] L. Li, S. Chen, J. Kleban, and A. Gupta. Counterfactual estimation and optimization of click metrics for search engines. *CoRR*, abs/1403.1891, 2014.

[8] L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 297–306, 2011.

[9] L. Li, R. Munos, and C. Szepesvári. On minimax optimal offline policy evaluation. *CoRR*, abs/1409.3653, 2014.

[10] J. Mary, P. Preux, and O. Nicol. Improving offline evaluation of contextual bandit algorithms via bootstrapping techniques. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 172–180, 2014.

[11] A. Maurer and M. Pontil. Empirical bernstein bounds and sample-variance penalization. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

[12] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[13] A. L. Strehl, J. Langford, L. Li, and S. Kakade. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada.*, pages 2217–2225, 2010.

[14] A. Swaminathan and T. Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. *CoRR*, abs/1502.02362, 2015.

[15] V. Vapnik. *Statistical Learning Theory*. Wiley, Chichester, GB, 1998.

[16] B. Zadrozny, J. Langford, and N. Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM '03, pages 435–, 2003.