Learning Probability Distributions

Learning Probability Distributions

Ronald Cruz and Yves Greatti

May 1st, 2018

Setting

- $\mathcal{X}$: input space, $\mathcal{Y}$: output space
- Q: source distribution, P: target distribution
- $\hat{Q}$, $\hat{P}$: empirical distributions
- f_Q, f_P: labeling functions from $\mathcal{X} \to \mathcal{Y}$
- (Q, f_Q): source domain, (P, f_P): target domain
- Learner receives:
  1. $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ i.i.d from Q
  2. Unlabeled sample $\mathcal{T} = \{x'_1, \ldots, x'_n\}$ i.i.d from P
  3. Possibly: labeled sample $\mathcal{T}' = \{(x''_1, y''_1), \ldots, (x''_s, y''_s)\}$

Goal: Learn the target labeling function f_P

Setting

Definition (Expected Loss over Distribution)

Given two functions $f, g : \mathcal{X} \to \mathcal{Y}$, a loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ and a distribution D over $\mathcal{X}$. The expected loss of f and g with respect to L is:

$$\mathcal{L}_D(f, g) = \mathbb{E}_{x \sim D}[L(f, g)]$$

Objective: Find $h \in H$ that minimizes

$$\mathcal{L}_P(h, f_P) = \mathbb{E}_{x \sim P}[L(h(x), f_P(x)]$$

Discrepancy

Definition (Discrepancy)

Given a hypothesis set H, the discrepancy between two distributions P and Q over $\mathcal{X}$ is defined by:

$$\mathrm{disc}(P, Q) = \max_{h, h' \in H} |\mathcal{L}_P(h', h) - \mathcal{L}_Q(h', h)|$$

Smaller empirical discrepancy, $\mathrm{disc}(\hat{P}, \hat{Q})$, guarantees a closeness of pointwise losses.

DM Algorithm

Given a PSD kernel K, the hypothesis returned by this algorithm solves the following optimization problem

$$\min_{h \in \mathbb{H}} \lambda \|h\|_K^2 + \mathcal{L}_{\mathsf{q}_{\min}}(h, f_Q)$$

where $\mathsf{q}_{\min} = \operatorname{argmin}_{\operatorname{supp}(\mathsf{q}) \subseteq \operatorname{supp} \hat{Q}} \operatorname{disc}(\mathsf{q}, \hat{P})$.

$\mathsf{q}_{\min}$ can be thought of as a constant reweighing function from $\mathcal{S}_{\mathcal{X}} = \{x_1, \ldots, x_m\} \to [0, 1]$.

Learning Guarantee

Theorem

Let q *be an arbitrary distribution over* $\mathcal{S}_{\mathcal{X}}$ *and let* h^* *and* h_{q} *be the hypothesis minimizing* $\lambda\|h\|_K^2 + \mathcal{L}_{\hat{P}}(h, f_P)$ *and* $\lambda\|h\|_K^2 + \mathcal{L}_{\mathsf{q}}(h, f_Q)$ *respectively. Then, the following inequality holds:*

$$\lambda\|h^* - h_{\mathsf{q}}\|_K^2 \leq \mu\eta_H(f_P, f_Q) + disc(\hat{P}, \mathsf{q})$$

where $\eta_H(f_P, f_Q) =$

$$\min_{h \in H}\left(\max_{x \in \mathrm{supp}(\hat{P})} |f_P(x) - h(x)| + \max_{x \in \mathrm{supp}(\hat{Q})} |f_Q(x) - h(x)|\right)$$

GDM Algorithm Idea

- Consider ideal scenario with access to target labels.

$$\min_{h \in H} F(h) = \lambda \|h\|_K^2 + \mathcal{L}_{\hat{P}}(h, f_P)$$

- Reweighing scheme: For any $h \in H$, $\mathsf{Q}_h : \mathcal{S}_{\mathcal{X}} \to \mathbb{R}$ such that $|\mathcal{L}_{\mathsf{Q}_h}(h, f_Q) - \mathcal{L}_{\hat{P}}(h, f_P)|$ is minimized.
- f_P is unknown: relax search to a nonempty convex surrogate hypothesis set $H'' \subseteq H$ that may contain f_P.

$$Q_h = \operatorname*{argmin}_{\mathsf{q} \in \mathcal{F}(\mathcal{S}_{\mathcal{X}}, \mathbb{R})} \max_{h'' \in H''} |\mathcal{L}_{\mathsf{q}}(h, f_Q) - \mathcal{L}_{\hat{P}}(h, h'')|$$

where $\mathcal{F}(\mathcal{S}_{\mathcal{X}}, \mathbb{R})$ is the set of all real valued functions defined over $\mathcal{S}_{\mathcal{X}}$.

GDM Algorithm

Proposition

The following identity holds for any $h \in H$:

$$\mathcal{L}_{\mathsf{Q}_h}(h, f_Q) = \frac{1}{2}\left(\max_{h'' \in H''} \mathcal{L}_{\hat{P}}(h, h'') + \min_{h'' \in H''} \mathcal{L}_{\hat{P}}(h, h'')\right)$$

This leads to the following convex optimization problem (assuming L jointly convex)

$$\min_{h \in H} \lambda \|h\|_K^2 + \frac{1}{2}\left(\max_{h'' \in H''} \mathcal{L}_{\hat{P}}(h, h'') + \min_{h'' \in H''} \mathcal{L}_{\hat{P}}(h, h'')\right)$$

Generalized Discrepancy

Let $\mathcal{A}(H)$ be the set of functions U: h $\rightarrow U_h$ such that for all $h \in H$, h $\rightarrow \mathcal{L}_{U_h}(h, f_Q)$ is a convex function. $\mathcal{A}(H)$ includes the function Q: h $\rightarrow Q_h$.

Definition (Generalized Discrepancy)

For any $U \in \mathcal{A}(H)$ the generalized discrepancy between two distributions $\hat{P}$ and U is defined as

$$\mathrm{DISC}(\hat{P}, U) = \max_{h \in H, h'' \in H''} |\mathcal{L}_{\hat{P}}(h, h'') - \mathcal{L}_{U_h}(h, f_Q)|$$

Generalization Bound

Theorem

Let U an arbitrary element of $\mathcal{A}(H)$, and h^* and h_U the minimizers of $\lambda\|h\|_K^2 + \mathcal{L}_{\hat{P}}(h, f_P)$ and $\lambda\|h\|_K^2 + \mathcal{L}_{U_h}(h, f_Q)$ respectively. Then for any convex set $H'' \subseteq H$:

$$\lambda\|h^* - h_u\|_k^2 \leq \mu d_\infty^{\hat{P}}(f_P, H'') + DISC(\hat{P}, U)$$

where $d_\infty^{\hat{P}}(f_P, H'') =$

$$\min_{h_0 \in H''} \max_{x \in \mathrm{supp}(\hat{P})} |h_0(x) - f_P(x)|$$

Local Discrepancy

Definition (Local Discrepancy)

For a convex set $H'' \subseteq H$, we can define the local discrepancy

$$\mathrm{disc}_{H''}(\hat{P}, \mathrm{q}) = \max_{h \in H, h'' \in H''} |\mathcal{L}_{\hat{P}}(h, h'') - \mathcal{L}_{\mathrm{q}}(h, h'')|$$

This is a finer measure than standard discrepancy where the max is defined over all pairs of hypothesis *both* in $H \supseteq H''$.

Relating DM and GDM Algorithm

Theorem

Let L be an L_p loss and h_0^* the minimizer of $\eta_H(f_P, f_Q)$. Define $r \geq 0$ by $r = \max_{x \in \operatorname{supp} \hat{Q}} |f_Q(x) - h_0^*(x)|$. Let q be a distribution over $\mathcal{S}_{\mathcal{X}}$ and $H'' = \{h'' \in H \mid \mathcal{L}_{\mathrm{q}}(h'', f_Q) \leq r^p\}$. Then, $h_0^* \in H''$ and the following inequality holds:

$$\mu d_\infty^{\hat{P}}(f_P, H'') + DISC(\hat{P}, \mathrm{q}) \leq \mu \eta_H(f_P, f_Q) + disc_{H''}(\hat{P}, q)$$

Generalization Bound

Theorem

Let h^* and h_{Q} be a minimizer of $\lambda\|h\|_K^2 + \mathcal{L}_{\hat{P}}(h, f_P)$ and $\lambda\|h\|_K^2 + \mathcal{L}_{\mathrm{Q}_h}(h, f_Q)$ respectively. Then, for all $x \in \mathcal{X}, y \in \mathcal{Y}$, the following holds for any convex set $H'' \subseteq H$:

$$|L(h_{\mathrm{Q}}(x), y) - L(h^*(x), y)| \leq \mu R \sqrt{\frac{\mu d_\infty^{\hat{P}}(f_P, H'') + DISC(\hat{P}, Q)}{\lambda}}$$

where $R^2 = \sup_{x \in \mathcal{X}} K(x, x)$. If further L is an L_p loss for some $p \geq 1$ and H'' is defined by the previous theorem, then the following holds for all $x \in \mathcal{X}, y \in \mathcal{Y}$:

$$|L(h_Q(x), y) - L(h^*(x), y)| \leq \mu R \sqrt{\frac{\mu \eta_H(f_P, f_Q) + disc_{H''}(\hat{P}, q_{\min})}{\lambda}}$$

Summary and Further Remarks

- Extension to hypothesis-dependent reweighing
- Convex optimization problem
- Formulation of exact solution as SDP and approximation as QP
- Empirical improvements over DM

Training GANs with optimism Improved Training of Wasserstein GANs

Why study generative modeling (Ian Goodfellow NIPS 2016 Tutorial)

1. Test our ability to represent and manipulate high-dimension probabilities
2. GAN can be incorporated into reinforcement learning by simulating possible futures
3. They can be trained with small set of samples and provide predictions with missing data (semi-supervised learning)
4. "Finding Nash equilibria in high-dimensional, continuous, non-convex games is an important open research problem"

Generative Adversarial Networks

1. Two networks G:generator and D:discriminator
2. The Generator using as input random noise $z \in \mathbb{P}_\theta$, learns an approximation of $\mathbb{P}_r$ and tries to fool the discriminator
3. The Discriminator takes a sample either from $\mathbb{P}_r$ or from G and classifies it as real or fake
4. Discriminator and Generator are trained in turn
5. G and D play the following two-player minimax game $\min_G \max_D \mathbb{E}_{x \sim \mathbb{P}_r}[\log(D(x))] + \mathbb{E}_{x \sim \mathbb{P}_g}[\log(1 - D(x))]$
6. Optimal value for D is $\frac{p_r}{p_r + p_g}$
7. Global optimal $p_r = p_g$ and $D^* = \frac{1}{2}$
8. Cost function is $L(G, D^*) = 2D_{JS}(p_r \| p_g) - 2\log(2)$

WGAN

1. GAN training unstable (Tips and tricks to make GANs work: https://github.com/soumith/ganhacks)
2. Most of datasets concentrate in lower dimensional manifolds. If the discriminator is perfect, the loss function L goes to zero (dilemma!)
3. Wasserstein Distance to replace KL or JS divergences, it has better convergence bounds
4. The WGAN objective function is constructed using Kantorovitch-Rubinstein duality to obtain:
$$\min_G \max_{D \in \mathcal{D}} \mathbb{E}_{x \sim \mathbb{P}_r}[D(x)] - \mathbb{E}_{x \sim \mathbb{P}_g}[D(x)]$$
5. Discriminator is trained to learn a K-Lipschitz continuous function
6. To enforce the Lipschitz constraint on the critique, weights of the neural network are clipped on a compact set [-c,c]

Optimistic Mirror Descent

Solving the previous minmax problem, i.e. $min_\theta \max_w f(\theta, w)$ is equivalent to find the saddle points of f. Saddle point problems are usually solved by gradient based optimization method:

$$w_{t+1} = w_t + \eta.\nabla_{w,t}$$
$$\theta_{t+1} = \theta_t - \eta.\nabla_{\theta,t}$$

If $L(\theta, w)$ is convex in θ and concave in w, (θ, w) lie in some bounded convex set, FTL shows that on average there is convergence to an ϵ-equilibrium with $\epsilon = \mathcal{O}(\frac{1}{\sqrt{T}})$ and $\overline{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$ and $\overline{\theta} = \frac{1}{T}\sum_{t=1}^{T}\theta_t$.

Optimistic Mirror Descent

Intuition: from FTRL formulation:

$$w_{t+1} = \operatorname*{argmax}_{w} \eta \sum_{s=1}^{t} \langle w, \nabla_{w,s} \rangle - \|w\|_2^2$$
$$\theta_{t+1} = \operatorname*{argmin}_{\theta} \eta \sum_{s=1}^{t} \langle \theta, \nabla_{\theta,s} \rangle + \|\theta\|_2^2$$

If learner knows in advance gradient at next iteration: constant regret. OMD adds a predictor M_{t+1} which could be either last iteration's gradient, or an average of a window of last gradient or discounted average of past gradients.

Optimistic Mirror Descent

Objective functions:

$$w_{t+1} = \underset{w}{\operatorname{argmax}}\, \eta\Big(\sum_{s=1}^{t} \langle w, \nabla_{w,s} \rangle + \langle w, M_{w,t+1} \rangle\Big) - \|w\|_2^2$$

$$\theta_{t+1} = \underset{\theta}{\operatorname{argmin}}\, \eta\Big(\sum_{s=1}^{t} \langle \theta, \nabla_{\theta,s} \rangle + \langle \theta, M_{\theta,t+1} \rangle\Big) + \|\theta\|_2^2$$

Update rules:

$$w_{t+1} = w_t + \eta.(\nabla_{w,t} + M_{w,t+1} - M_{w,t})$$
$$\theta_{t+1} = \theta_t - \eta.(\nabla_{\theta,t} + M_{\theta,t+1} - M_{\theta,t})$$

Optimistic Mirror Descent

Theorem (Last iterate convergence)

1. $\gamma = \max\left(\|(AA^T)^\dagger\|, \|(A^TA)^\dagger\|\right)$
2. $\lambda = \|A\| \leq 1$
3. $\eta < \frac{1}{3\gamma^2}$
4. $\Delta_t = \|A^T x_t\|_2^2 + \|Ay_t\|_2^2$
5. *Initialization:* $x_0 \in \mathcal{R}(A)$, $y_0 \in \mathcal{R}(A^T)$

The OMD update rules satisfy:

$$\Delta_1 = \Delta_0 \geq \frac{1}{1+\eta^2}\Delta_2$$

$$\forall t \geq 3 : \Delta_t \leq (1 - (\frac{\eta}{\gamma})^2)\Delta_{t-1} + 16\eta^3\Delta_0$$

Last iterate convergence

Proof: using induction then

$$\Delta_t \le (1 - (\frac{\eta}{\gamma})^2)^{t-2}(1+\eta)^2\Delta_0^0 + 16\sum_{t=0}^{\infty}(1 - \frac{\eta^2}{\gamma^2})^t\eta^3\Delta_0^0$$
$$= (1 - (\frac{\eta}{\gamma})^2)^{t-2}(1+\eta)^2\Delta_0^0 + \mathcal{O}(\eta\gamma^2\Delta_0^0)$$

Last iterate convergence: In particular, as t $\to \infty$, the last iterate of OMD is within $\mathcal{O}(\gamma\sqrt{\eta.\Delta_0^0})$ distance from the space of equilibrium points, where $\sqrt{\Delta_0^0}$ is the distance of (x_0, y_0) to the equilibrium space and where the distance is taken w.r.t $\sqrt{x^T A A^T x + y^T A^T A y}$.

Learning the mean of a multivariate normal distribution

$\mathbb{P}_r \approx N(v, I), v \in \mathbb{R}^d$, input noise z drawn from $N(0, I)$. The goal of the generator is to figure out the true distribution, i.e. to converge to v.

$$D_w(x) = \langle w, x \rangle$$
$$G_\theta(z) = z + \theta$$

The WGAN takes the form:

$$L(\theta, w) = \mathbb{E}_{x \sim N(v,I)}[\langle w, x \rangle] - \mathbb{E}_{z \sim N(0,I)}[\langle w, z + \theta \rangle]$$

Expected zero-sum game: $\inf_\theta \sup_w \langle w, v - \theta \rangle$
The unique equilibrium is for the generator to choose $\theta = v$, and for the discriminator to choose w=0.

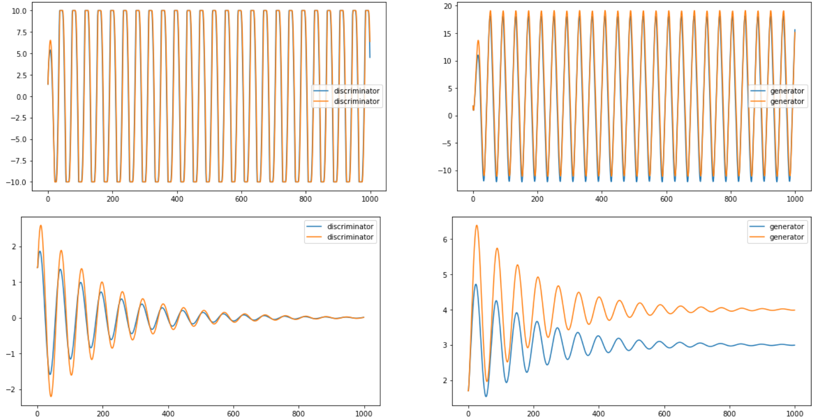
Learning the mean of a multivariate normal distribution

We have $\nabla_{w,t} = v - \theta_t$ and $\nabla_{\theta,t} = -w_t$, the update rules for OMD are:

$$w_{t+1} = w_t + 2\eta.(v - \theta_t) - \eta.(v - \theta_{t-1})$$
$$\theta_{t+1} = \theta_t + 2\eta.w_t - \eta.w_{t-1}$$

1. Using different update rules (Adagrad, Momentum, Nesterov momentum) GD always leads to a limit cycle
2. Robustness of last-iterate convergence for OMD and SOMD
3. Similar results when learning a co-variance matrix

Learning the mean of a multivariate normal distribution



Experimental results

Promising results for OMD and variants compared to GD with modifications confirmed in other experiments:

1. Generating DNA sequences: CNNs networks using SOMD achieve lower KL divergence than SGD
2. Generating images from CIFAR10 with optimistic Adam: highest inception score

Weight clipping limitations

1. Very deep WGAN critics often fail to converge: vanishing or exploding gradients
2. They learn simple functions: weight clipping ignores higher moments of the data

$\Rightarrow$ solution: add a regularization term using the L2 norm of gradient of D(x)

Gradient Penalty:

$$\mathbb{E}_{x\sim Q}[D_w(x)] - \mathbb{E}_{z\sim F}[D_w(G_\theta(z))] - \lambda \mathbb{E}_{\hat{x}\sim Q_\epsilon}[(\|\nabla_x D_w(\hat{x})\| - 1)^2]$$

Q_ϵ is the uniform distribution of points along $\epsilon.x + (1-\epsilon).G(z)$ when $x \sim Q$ and $z \sim F$.

Conclusion

1. Stability of GAN training yet challenging!
   Many attempts which led to various flavors of GANs:
   MGAN (multi-discriminators), EBGAN (Energy-based),
   F-GAN (f-divergence-based),...
   No unified strategy: deep boosting applied to GANs
   (AdaGan)
2. Interesting open question: non convex-concave settings in zero-sum two players game