DSC190: Machine Learning with Few Labels

A "standardized" view of ML

Zhiting Hu
Lecture 9, October 21, 2021

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Logistics

- HW1 due extended to Sunday (10/24)
- HW2 out: (much) easier than HW1 !
- Mid-term survey

Outline

- Functional derivative
- A “standardized” view of ML

Functional derivative

- $\nabla_q - \mathbb{H}(q) = \log q + 1$

- Functional $F(y)$: an operator that takes a function $y(x)$ and returns an output value F
- Functional derivative (aka, variational derivative) relates a change in a Functional $F(y)$ to a change in the function y

Functional derivative

- Recall the conventional derivative $\frac{dy}{dx}$
  - Taylor expansion

$$y(x+\epsilon) = y(x) + \frac{\mathrm{d}y}{\mathrm{d}x}\epsilon + O(\epsilon^2)$$

- Functional derivative
  - How much a functional $F[y]$ changes when we make a small change $\varepsilon\eta(x)$ to the function $y(x)$

$$F[y(x) + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)} \eta(x)\,\mathrm{d}x + O(\epsilon^2)$$

  - A function $y(x)$ that maximizes (or minimizes) a functional $F[y]$ must satisfy

$$\frac{\delta F}{\delta y(x)} = 0 \text{ for all } x$$

Functional derivative

$$F[y(x) + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)} \eta(x)\,\mathrm{d}x + O(\epsilon^2)$$

- Consider a functional that is defined by an integral over a function $G(y, x)$

$$F[y] = \int G(y, x) dx$$

- Consider variations in the function $y(x)$,

$$F[y + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\partial G}{\partial y} \eta(x) dx + O(\epsilon^2)$$

Functional derivative

$$F[y(x) + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)} \eta(x)\, \mathrm{d}x + O(\epsilon^2)$$

- Consider a functional that is defined by an integral over a function $G(y, x)$
$$F[y] = \int G(y, x) dx$$
- Consider variations in the function $y(x)$,
$$F[y + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\partial G}{\partial y} \eta(x) dx + O(\epsilon^2)$$
- Ex.1, $-\mathbb{H}(q) = \int q(x) \log q(x)\, dx$
  - $G = q(x) \log q(x)$
- EX.2, posterior regularization

Ex.2: Posterior Regularization

assume single data point $\boldsymbol{x}^*$

$$\min_{q,\boldsymbol{\xi}} \quad -\mathrm{H}\left(q(\boldsymbol{z})\right) - \mathbb{E}_{q(\boldsymbol{z})}\left[\log p(\boldsymbol{x}^*|\boldsymbol{z})\pi(\boldsymbol{z})\right] + \sum_i \xi_i$$

$$s.t. \;\; \mathbb{E}_q\left[T(\boldsymbol{x}^*, \boldsymbol{z})\right] \leq \boldsymbol{\xi}$$

$$\boldsymbol{\xi} \geq 0,$$

- *Lagrangian*

$$\max_{\boldsymbol{\mu}\geq 0, \boldsymbol{\eta}\geq 0, \alpha\geq 0} \min_{q,\boldsymbol{\xi}} \quad -\mathrm{H}\left(q(\boldsymbol{z})\right) - \mathbb{E}_{q(\boldsymbol{z})}\left[\log p(\boldsymbol{x}^*|\boldsymbol{z})\pi(\boldsymbol{z})\right]$$

$$+ \sum_i (1-\mu_i)\xi_i + \sum_i \eta_i \left(\mathbb{E}_q[T_i(\boldsymbol{x}^*; \boldsymbol{z})] - \xi_i\right) + \alpha\left(\sum_z q(\boldsymbol{z}) - 1\right)$$

A “Standardized” View of ML

The general expression as a constrained optimization:

(auxiliary) distribution q

loss

$$\min_{q,\boldsymbol{\theta}} \ \mathcal{L}(q, \boldsymbol{\theta})$$

constrained set

$$s.t. \ \ q \in \mathcal{Q}.$$

MaxEnt perspective

- Supervised MLE and maximum entropy
- Unsupervised MLE and maximum entropy
- Bayesian inference and maximum entropy
  - Bayesian inference as optimization
- Posterior regularization:
  - Constrained Bayesian inference => constrained optimization

MaxEnt perspective

The general expression as a constrained optimization:

(auxiliary) distribution q

loss

$$\min_{q,\boldsymbol{\theta}} \ \mathcal{L}(q, \boldsymbol{\theta})$$

constrained set

$$s.t. \ \ q \in \mathcal{Q}.$$

- Supervised MLE and maximum entropy

$$\min_{q(\boldsymbol{x},\boldsymbol{y})} \ H(q)$$

$$s.t. \ \mathbb{E}_q[T(\boldsymbol{x},\boldsymbol{y})] = \mathbb{E}_{(x^*,y^*)\sim\mathcal{D}}[T(\boldsymbol{x},\boldsymbol{y})]$$

The general expression as a constrained optimization:

MaxEnt perspective

(auxiliary) distribution q

$$\min_{q,\boldsymbol{\theta}} \ \mathcal{L}(q, \boldsymbol{\theta})$$

loss

$$s.t. \ \ q \in \mathcal{Q}.$$

constrained set

- Supervised MLE and maximum entropy
- Unsupervised MLE and maximum entropy

$$\min_{q,\theta} H\big(q(\boldsymbol{y}|\boldsymbol{x}^*)\big) + \mathbb{E}_{q(\boldsymbol{y}|\boldsymbol{x}^*)}[\log p_\theta(\boldsymbol{x}^*, \boldsymbol{y})]$$

The general expression as a constrained optimization:

(auxiliary) distribution q

loss

$$\min_{q,\boldsymbol{\theta}} \; \mathcal{L}(q,\boldsymbol{\theta})$$

constrained set

$$s.t. \;\; q \in \mathcal{Q}.$$

MaxEnt perspective

- Supervised MLE and maximum entropy
- Unsupervised MLE and maximum entropy
- Bayesian inference and maximum entropy

$$\min_{q(\boldsymbol{z})} \; -\mathrm{H}(q(\boldsymbol{z})) + \log p(\mathcal{D}) - \mathbb{E}_{q(\boldsymbol{z})}\left[\log \pi(\boldsymbol{z}) + \sum\nolimits_{\boldsymbol{x}^* \in \mathcal{D}} \log p(\boldsymbol{x}^*|\boldsymbol{z})\right]$$

$$s.t. \; q(\boldsymbol{z}) \in \mathcal{P}$$

The general expression as a constrained optimization:

MaxEnt perspective

(auxiliary) distribution q

loss

$$\min_{q,\boldsymbol{\theta}} \ \mathcal{L}(q, \boldsymbol{\theta})$$

constrained set

$$s.t. \ \ q \in \mathcal{Q}.$$

- Supervised MLE and maximum entropy
- Unsupervised MLE and maximum entropy
- Bayesian inference and maximum entropy
- Posterior regularization

$$\begin{aligned} \min_{q,\boldsymbol{\xi}} \quad & -\mathrm{H}\left(q(\boldsymbol{z})\right) - \mathbb{E}_{q(\boldsymbol{z})}\left[\sum\nolimits_{\boldsymbol{x}^* \in \mathcal{D}} \log p(\boldsymbol{x}^*|\boldsymbol{z})\pi(\boldsymbol{z})\right] + U(\boldsymbol{\xi}) \\ s.t. \quad & q(\boldsymbol{z}) \in \mathcal{Q}(\boldsymbol{\xi}) \\ & \boldsymbol{\xi} \geq 0, \end{aligned}$$

The Standard Equation (SE)

- Let t be the variable of interest
  - E.g., the input-output pair $\boldsymbol{t} = (\boldsymbol{x}, \boldsymbol{y})$ in a prediction task
  - or $\boldsymbol{t} = \boldsymbol{x}$ in generative modeling
- $p_\theta(\boldsymbol{t})$: the target model to be learned
- $q(\boldsymbol{t})$: auxiliary distribution
- The SE: $$\min_{q, \boldsymbol{\theta}, \boldsymbol{\xi}} - \alpha \mathbb{H}(q) + \beta \mathbb{D}\Big(q(\boldsymbol{t}), p_\theta(\boldsymbol{t})\Big) + U(\boldsymbol{\xi})$$
$$s.t. \ -\mathbb{E}_{q(\boldsymbol{t})}\Big[f_k(\boldsymbol{t}) \Big] < \xi_k, \quad k = 1, \dots, K$$
  - Experience function f represents external experiences of different kinds for training the model
    - $f_k(\boldsymbol{t}) \in \mathbb{R}$: measures the goodness of a configuration $\boldsymbol{t}$ in light of any given experiences
    - Data, constraints, reward, adversarial discriminators, etc., can all be formulated as an experience function (later)
    - Maximizing $\mathbb{E}_{q(\boldsymbol{t})}[f_k(\boldsymbol{t})]$ -> q is encouraged to produce samples receiving high scores

[Hu & Xing, 2021]

The Standard Equation (SE)

- Let t be the variable of interest
  - E.g., the input-output pair $\boldsymbol{t} = (\boldsymbol{x}, \boldsymbol{y})$ in a prediction task
  - or $\boldsymbol{t} = \boldsymbol{x}$ in generative modeling
- $p_\theta(\boldsymbol{t})$: the target model to be learned
- $q(\boldsymbol{t})$: auxiliary distribution
- The SE:

$$\min_{q, \boldsymbol{\theta}, \boldsymbol{\xi}} -\alpha \mathbb{H}(q) + \beta \mathbb{D}\left(q(\boldsymbol{t}), p_\theta(\boldsymbol{t})\right) + U(\boldsymbol{\xi})$$

$$s.t. \; -\mathbb{E}_{q(\boldsymbol{t})}\left[f_k(\boldsymbol{t}) \right] < \xi_k, \quad k = 1, \ldots, K$$

  - Divergence $\mathbb{D}$: measures the distance between the target model p_θ to be trained and the auxiliary model q
    - E.g., cross entropy

[Hu & Xing, 2021]

The Standard Equation (SE)

- Let t be the variable of interest
  - E.g., the input-output pair $\boldsymbol{t} = (\boldsymbol{x}, \boldsymbol{y})$ in a prediction task
  - or $\boldsymbol{t} = \boldsymbol{x}$ in generative modeling
- $p_\theta(\boldsymbol{t})$: the target model to be learned
- $q(\boldsymbol{t})$: auxiliary distribution
- The SE: $$\min_{q,\boldsymbol{\theta},\boldsymbol{\xi}} -\alpha\mathbb{H}(q) + \beta\mathbb{D}\left(q(\boldsymbol{t}), p_\theta(\boldsymbol{t})\right) + U(\boldsymbol{\xi})$$ $$s.t. \ -\mathbb{E}_{q(\boldsymbol{t})}\left[f_k(\boldsymbol{t})\right] < \xi_k, \quad k = 1, \ldots, K$$
  - Uncertainty $\mathbb{H}$: controls the compactness of the model
    - E.g., Shannon entropy

[Hu & Xing, 2021]

The Standard Equation (SE)

$$\min_{q,\boldsymbol{\theta},\boldsymbol{\xi}} -\alpha\mathbb{H}(q) + \beta\mathbb{D}\Big(q(\boldsymbol{t}), p_\theta(\boldsymbol{t})\Big) + U(\boldsymbol{\xi})$$

$$s.t.\ -\mathbb{E}_{q(\boldsymbol{t})}\Big[f_k(\boldsymbol{t})\Big] < \xi_k, \quad k = 1, \ldots, K$$

Assuming penalty $U = \sum_k \xi_k$, and $f = \sum_k f_k$:

$$\min_{q,\theta} -\alpha\mathbb{H}(q) + \beta\mathbb{D}\Big(q(\boldsymbol{t}), p_\theta(\boldsymbol{t})\Big) - \mathbb{E}_{q(\boldsymbol{t})}\Big[f(\boldsymbol{t})\Big]$$

3 terms:

Uncertainty
(self-regularization)
e.g., Shannon entropy

Divergence
(fitness)
e.g., Cross Entropy

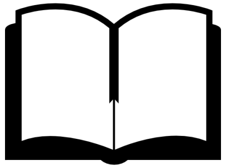
Experiences
(exogenous regularizations)
e.g., data examples, rules

Uncertainty

Teacher
$q(\boldsymbol{t})$

Student
$p_\theta(\boldsymbol{t})$

Textbook
$f(\boldsymbol{t})$



The Standard Equation (SE)

$$\min_{q,\theta} -\alpha\mathbb{H}(q) + \beta\mathbb{D}\Big(q(\boldsymbol{t}), p_\theta(\boldsymbol{t})\Big) - \mathbb{E}_{q(\boldsymbol{t})}\Big[f(\boldsymbol{t})\Big]$$

- The introduction of the auxiliary distribution q relaxes the learning problem of p_θ, originally only over $\boldsymbol{\theta}$, to be now alternating between q and $\boldsymbol{\theta}$
  - Recall in EM, we introduced q to deal with the intractable marginal log-likelihood
- q acts as a conduit between the exogenous experience and the target model
  - subsumes the experience, by maximizing the expected f value
  - passes it incrementally to the target model, by minimizing the divergence $\mathbb{D}$
- E.g., assume $\mathbb{D}$ is cross entropy, and $\mathbb{H}$ is Shannon entropy
  - The above optimization, at each iteration n:

$$q^{(n+1)}(\boldsymbol{t}) = \exp\left\{\frac{\beta\log p_{\theta^{(n)}}(\boldsymbol{t}) + f(\boldsymbol{t})}{\alpha}\right\} / Z$$

$$\boldsymbol{\theta}^{(n+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{E}_{q^{(n+1)}(\boldsymbol{t})}\left[\log p_\theta(\boldsymbol{t})\right],$$

The Standard Equation (SE)

$$\min_{q,\theta} -\alpha\mathbb{H}(q) + \beta\mathbb{D}\Big(q(\boldsymbol{t}), p_\theta(\boldsymbol{t})\Big) - \mathbb{E}_{q(\boldsymbol{t})}\Big[f(\boldsymbol{t})\Big]$$

- The introduction of the auxiliary distribution q relaxes the learning problem of p_θ, originally only over $\boldsymbol{\theta}$, to be now alternating between q and $\boldsymbol{\theta}$
  - Recall in EM, we introduced q to deal with the intractable marginal log-likelihood
- q acts as a conduit between the exogenous experience and the target model
  - subsumes the experience, by maximizing the expected f value
  - passes it incrementally to the target model, by minimizing the divergence $\mathbb{D}$
- E.g., assume $\mathbb{D}$ is cross entropy, and $\mathbb{H}$ is Shannon entropy
  - The above optimization, at each iteration n:

$$\text{Teacher:} \quad q^{(n+1)}(\boldsymbol{t}) = \exp\left\{\frac{\beta\log p_{\theta^{(n)}}(\boldsymbol{t}) + f(\boldsymbol{t})}{\alpha}\right\} \,/\, Z$$

$$\text{Student:} \quad \boldsymbol{\theta}^{(n+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{E}_{q^{(n+1)}(\boldsymbol{t})}\left[\log p_\theta(\boldsymbol{t})\right],$$

The Standard Equation (SE)

$$\min_{q,\theta} -\alpha\mathbb{H}(q) + \beta\mathbb{D}\Big(q(\boldsymbol{t}), p_\theta(\boldsymbol{t})\Big) - \mathbb{E}_{q(\boldsymbol{t})}\Big[f(\boldsymbol{t})\Big]$$

- Formulates a large space of learning algorithms, which encompasses many well-known algorithms

SE encompasses many well-known algorithms (more later)

Experience type	Experience function f	Divergence $\mathbb{D}$	α	β	Algorithm
Data instances	$f_{\text{data}}(\boldsymbol{x}; \mathcal{D})$	CE	1	1	Unsupervised MLE
	$f_{\text{data}}(\boldsymbol{x}, \boldsymbol{y}; \mathcal{D})$	CE	1	ϵ	Supervised MLE
	$f_{\text{data-self}}(\boldsymbol{x}, \boldsymbol{y}; \mathcal{D})$	CE	1	ϵ	Self-supervised MLE
	$f_{\text{data-w}}(\boldsymbol{t}; \mathcal{D})$	CE	1	ϵ	Data re-weighting
	$f_{\text{data-aug}}(\boldsymbol{t}; \mathcal{D})$	CE	1	ϵ	Data Augmentation
	$f_{\text{active}}(\boldsymbol{x}, \boldsymbol{y}; \mathcal{D})$	CE	1	ϵ	Active Learning (Ertekin et al., 2007)
Knowledge	$f_{rule}(\boldsymbol{x}, \boldsymbol{y})$	CE	1	1	Posterior Regularization (Ganchev et al., 2010)
	$f_{rule}(\boldsymbol{x}, \boldsymbol{y})$	CE	$\mathbb{R}$	1	Unified EM (Samdani et al., 2012)
Reward	$\log Q^{\theta}(\boldsymbol{x}, \boldsymbol{y})$	CE	1	1	Policy Gradient
	$\log Q^{\theta}(\boldsymbol{x}, \boldsymbol{y}) + Q^{in,\theta}(\boldsymbol{x}, \boldsymbol{y})$	CE	1	1	+ Intrinsic Reward
	$Q^{\theta}(\boldsymbol{x}, \boldsymbol{y})$	CE	$\tau > 0$	$\tau > 0$	RL as Inference
Other advanced	binary classifier	JSD	0	1	Vanilla GAN (Goodfellow et al., 2014)
	discriminator	f-divg.	0	1	f-GAN (Nowozin et al., 2016)
	1-Lipschitz discriminator	W_1 dist.	0	1	WGAN (Arjovsky et al., 2017)
	1-Lipschitz discriminator	KL	0	1	PPO-GAN (Wu et al., 2020)

SE with supervised data experience

$$\min_{q,\theta} -\alpha\mathbb{H}(q) + \beta\mathbb{D}\Big(q(\boldsymbol{t}), p_\theta(\boldsymbol{t})\Big) - \mathbb{E}_{q(\boldsymbol{t})}\Big[f(\boldsymbol{t}) \Big]$$

- Input-output variables $\boldsymbol{t} = (\boldsymbol{x}, \boldsymbol{y})$
- Experience: dataset $\mathcal{D} = \{(\boldsymbol{x}^*, \boldsymbol{y}^*)\}$ of size N
  - defines the empirical distribution

$$\tilde{p}(\boldsymbol{x}, \boldsymbol{y}) = \frac{m(\boldsymbol{x}, \boldsymbol{y})}{N} = \mathbb{E}_{(\boldsymbol{x}^*, \boldsymbol{y}^*)\sim\mathcal{D}}[\mathbb{1}_{(\boldsymbol{x}^*, \boldsymbol{y}^*)}(\boldsymbol{x}, \boldsymbol{y})]$$

The expected similarity between $(\boldsymbol{x}, \boldsymbol{y})$ and observed data $(\boldsymbol{x}^*, \boldsymbol{y}^*)$, with similarity measure $\mathbb{1}_a(b)$, i.e., an indicator function (1 if a=b, 0 otherwise)

SE with supervised data experience

$$\min_{q,\theta} -\alpha\mathbb{H}(q) + \beta\mathbb{D}\left(q(\boldsymbol{t}), p_\theta(\boldsymbol{t})\right) - \mathbb{E}_{q(\boldsymbol{t})}\left[f(\boldsymbol{t}) \right]$$

- Input-output variables $\boldsymbol{t} = (\boldsymbol{x}, \boldsymbol{y})$
- Experience: dataset $\mathcal{D} = \{(\boldsymbol{x}^*, \boldsymbol{y}^*)\}$ of size N
  - defines the empirical distribution

$$\tilde{p}(\boldsymbol{x}, \boldsymbol{y}) = \frac{m(\boldsymbol{x}, \boldsymbol{y})}{N} = \mathbb{E}_{(\boldsymbol{x}^*, \boldsymbol{y}^*)\sim\mathcal{D}}[\mathbb{1}_{(\boldsymbol{x}^*, \boldsymbol{y}^*)}(\boldsymbol{x}, \boldsymbol{y})]$$

- Define the experience function

$$f := f_{data}(\boldsymbol{x}, \boldsymbol{y}; \mathcal{D}) = \log \mathbb{E}_{(\boldsymbol{x}^*, \boldsymbol{y}^*)\sim\mathcal{D}}\left[\mathbb{1}_{(\boldsymbol{x}^*, \boldsymbol{y}^*)}(\boldsymbol{x}, \boldsymbol{y}) \right]$$

- Let $\mathbb{D}$ cross entropy, $\mathbb{H}$ Shannon entropy, $\alpha = 1, \beta = \epsilon$ (a very small value)

$$\min_{q,\theta} -H(q) - \epsilon\, \mathbb{E}_q\left[\log p_\theta(\boldsymbol{x}, \boldsymbol{y}) \right] - \mathbb{E}_q\left[f_{data}(\boldsymbol{x}, \boldsymbol{y}; \mathcal{D}) \right]$$

SE with supervised data experience

$$f := f_{data}(\boldsymbol{x}, \boldsymbol{y}; \mathcal{D}) = \log \mathbb{E}_{(\boldsymbol{x}^*, \boldsymbol{y}^*) \sim \mathcal{D}} \left[\mathbb{1}_{(\boldsymbol{x}^*, \boldsymbol{y}^*)}(\boldsymbol{x}, \boldsymbol{y}) \right]$$

$$\min_{q, \boldsymbol{\theta}} \; -H(q) - \epsilon \, \mathbb{E}_q \left[\log p_\theta(\boldsymbol{x}, \boldsymbol{y}) \right] - \mathbb{E}_q \left[f_{data}(\boldsymbol{x}, \boldsymbol{y}; \mathcal{D}) \right]$$

- At each iteration n:

$$\text{Teacher:} \quad q^{(n+1)}(\boldsymbol{t}) = \exp\left\{ \frac{\beta \log p_{\theta^{(n)}}(\boldsymbol{t}) + f(\boldsymbol{t})}{\alpha} \right\} / Z \;\; \approx \tilde{p}(\boldsymbol{x}, \boldsymbol{y})$$

$$\text{Student:} \quad \boldsymbol{\theta}^{(n+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{E}_{q^{(n+1)}(\boldsymbol{t})} \left[\log p_\theta(\boldsymbol{t}) \right],$$

q reduces to the empirical distribution

Maximizes data log-likelihood

- Recovers **supervised MLE**!

SE with unsupervised data experience

$$\min_{q,\theta} -\alpha\mathbb{H}(q) + \beta\mathbb{D}\Big(q(\boldsymbol{t}), p_\theta(\boldsymbol{t})\Big) - \mathbb{E}_{q(\boldsymbol{t})}\Big[f(\boldsymbol{t})\Big]$$

- Input-output variables $\boldsymbol{t} = (\boldsymbol{x}, \boldsymbol{y})$
- Experience: dataset $\mathcal{D} = \{(\boldsymbol{x}^*)\}$ of size N, I,e., we only observe the $\boldsymbol{x}$ part
  - defines the empirical distribution

$$\tilde{p}(\boldsymbol{x}) = \frac{m(\boldsymbol{x})}{N} = \mathbb{E}_{\boldsymbol{x}^*\sim\mathcal{D}}[\mathbb{1}_{\boldsymbol{x}^*}(\boldsymbol{x})]$$

- Define the experience function

$$f := f_{data}(\boldsymbol{x}\,;\mathcal{D}) = \log \mathbb{E}_{\boldsymbol{x}^*\sim\mathcal{D}}[\mathbb{1}_{\boldsymbol{x}^*}(\boldsymbol{x})]$$

- Let $\mathbb{D}$ cross entropy, $\mathbb{H}$ Shannon entropy, $\alpha = 1, \beta = 1$

$$\min_{q,\theta} -H(q) - \mathbb{E}_q\Big[\log p_\theta(\boldsymbol{x}, \boldsymbol{y})\Big] - \mathbb{E}_q\Big[\ f_{data}(\boldsymbol{x}; \mathcal{D})\ \Big]$$

  - Assume $q(\boldsymbol{x}, \boldsymbol{y}) = \tilde{p}(\boldsymbol{x})q(\boldsymbol{y}|\boldsymbol{x})$

Recovers **unsupervised MLE (EM)!**

SE with manipulated data experience

- Input-output variables $\boldsymbol{t} = (\boldsymbol{x}, \boldsymbol{y})$
- Experience: dataset $\mathcal{D} = \{(\boldsymbol{x}^*, \boldsymbol{y}^*)\}$ of size N
  - defines the empirical distribution

$$\tilde{p}(\boldsymbol{x}, \boldsymbol{y}) = \frac{m(\boldsymbol{x}, \boldsymbol{y})}{N} = \mathbb{E}_{(\boldsymbol{x}^*, \boldsymbol{y}^*)\sim\mathcal{D}}[\mathbb{1}_{(\boldsymbol{x}^*, \boldsymbol{y}^*)}(\boldsymbol{x}, \boldsymbol{y})]$$

- Define the experience function

$$f := f_{data}(\boldsymbol{x}, \boldsymbol{y}; \mathcal{D}) = \log \mathbb{E}_{(\boldsymbol{x}^*, \boldsymbol{y}^*)\sim\mathcal{D}}\left[\mathbb{1}_{(\boldsymbol{x}^*, \boldsymbol{y}^*)}(\boldsymbol{x}, \boldsymbol{y})\right]$$

- The similarity measure $\mathbb{1}_a(b)$ is too restrictive. Let's enrich it:
  - Don't have to be 0/1, we can scale it

$$f := f_{data-w}(\boldsymbol{x}, \boldsymbol{y}; \mathcal{D}) = \log \mathbb{E}_{(\boldsymbol{x}^*, \boldsymbol{y}^*)\sim\mathcal{D}}\left[w(\boldsymbol{x}^*, \boldsymbol{y}^*) \cdot \mathbb{1}_{(\boldsymbol{x}^*, \boldsymbol{y}^*)}(\boldsymbol{x}, \boldsymbol{y})\right]$$

  - Plug f_{data-w} into SE, keep all other configurations the same as supervised MLE, we recover **data re-weighting** in the "student" step

$$\max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{t}^*\sim\mathcal{D}}\left[w(\boldsymbol{t}^*) \cdot \log p_\theta(\boldsymbol{t}^*)\right]$$

SE with manipulated data experience

- Input-output variables $\boldsymbol{t} = (\boldsymbol{x}, \boldsymbol{y})$
- Experience: dataset $\mathcal{D} = \{(\boldsymbol{x}^*, \boldsymbol{y}^*)\}$ of size N
  - defines the empirical distribution

$$\tilde{p}(\boldsymbol{x}, \boldsymbol{y}) = \frac{m(\boldsymbol{x}, \boldsymbol{y})}{N} = \mathbb{E}_{(\boldsymbol{x}^*, \boldsymbol{y}^*) \sim \mathcal{D}}[\mathbb{1}_{(\boldsymbol{x}^*, \boldsymbol{y}^*)}(\boldsymbol{x}, \boldsymbol{y})]$$

- Define the experience function

$$f := f_{data}(\boldsymbol{x}, \boldsymbol{y}; \mathcal{D}) = \log \mathbb{E}_{(\boldsymbol{x}^*, \boldsymbol{y}^*) \sim \mathcal{D}}\left[\mathbb{1}_{(\boldsymbol{x}^*, \boldsymbol{y}^*)}(\boldsymbol{x}, \boldsymbol{y}) \right]$$

- The similarity measure $\mathbb{1}_a(b)$ is too restrictive. Let's enrich it:
  - Don't have to match exactly, we can relax it

$$f := f_{data-aug}(\boldsymbol{x}, \boldsymbol{y}; \mathcal{D}) = \log \mathbb{E}_{(\boldsymbol{x}^*, \boldsymbol{y}^*) \sim \mathcal{D}}\left[a_{(\boldsymbol{x}^*, \boldsymbol{y}^*)}(\boldsymbol{x}, \boldsymbol{y}) \right]$$

    - $a_{(\boldsymbol{x}^*, \boldsymbol{y}^*)}(\boldsymbol{x}, \boldsymbol{y})$: assigns non-zero probability to not only the exact $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ but also other $(\boldsymbol{x}, \boldsymbol{y})$ configurations
  - Plug $f_{data-aug}$ into SE, keep all other configurations the same as supervised MLE, we recover **data augmentation** in the "student" step $\max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{t}^* \sim \mathcal{D},\, \boldsymbol{t} \sim a_{\boldsymbol{t}^*}(\boldsymbol{t})} [\log p_\theta(\boldsymbol{t})]$.

SE with actively supervised experience

- Have access to a vast pool of unlabeled data instances
- Can select instances (queries) to be labeled by an oracle (e.g., human)

- Experiences:
  - $u(\boldsymbol{x})$ measures *informativeness* of an instance $\boldsymbol{x}$
    - e.g., Uncertainty on $\boldsymbol{x}$, measured by predictive entropy
  - Instances + oracle labels:

$$f(\boldsymbol{x}, \boldsymbol{y}\,;\, Oracle) = \log \mathbb{E}_{\boldsymbol{x}^* \sim \mathcal{D},\, \boldsymbol{y}^* \sim \mathrm{Oracle}(\boldsymbol{x}^*)}\left[\, \mathbb{1}_{(\boldsymbol{x}^*, \boldsymbol{y}^*)}(\boldsymbol{x}, \boldsymbol{y}) \,\right]$$

SE with actively supervised experience

$$\min_{q,\theta} -\alpha H(q) - \beta \mathbb{E}_q\left[\log p_\theta(\boldsymbol{x},\boldsymbol{y})\right] - \mathbb{E}_{q(\boldsymbol{x},\boldsymbol{y})}\left[\ f(\boldsymbol{x},\boldsymbol{y})\ \right]$$

$$f := f(\boldsymbol{x},\boldsymbol{y}\,;\, Oracle) + u(\boldsymbol{x}) \qquad \alpha = 1, \beta = \epsilon$$

- Teacher $q(\boldsymbol{x},\boldsymbol{y}) = \exp\left\{\dfrac{\beta \log p_\theta(\boldsymbol{x},\boldsymbol{y}) + f(\boldsymbol{x},\boldsymbol{y}\,;Oracle) + u(\boldsymbol{x})}{\alpha}\right\} / Z$

- Student $\min_{\theta} -\mathbb{E}_q\left[\log p_\theta(\boldsymbol{x},\boldsymbol{y})\right]$

Equivalent to **active learning** [e.g., Ertekin et al., 07]:
- Randomly draw a subset $\mathcal{D}_{sub} = \{\boldsymbol{x}^*\}$
- Draw a query $\boldsymbol{x}^*$ from $\mathcal{D}_{sub}$ according to $\exp\{u(\boldsymbol{x})\}$
- Get label $\boldsymbol{y}^*$ for $\boldsymbol{x}^*$ from the oracle
- Maximize log likelihood on $(\boldsymbol{x}^*, \boldsymbol{y}^*)$

Key Takeaways

(auxiliary) distribution q loss

$$\min_{q,\boldsymbol{\theta}} \ \mathcal{L}(q,\boldsymbol{\theta})$$

constrained set

$$s.t. \ \ q \in \mathcal{Q}.$$

- The MaxEnt perspective converts learning into a constrained optimization problem
- The standard equation (SE):

$$\min_{q,\theta} -\alpha \mathbb{H}(q) + \beta \mathbb{D}\Big(q(\boldsymbol{t}), p_\theta(\boldsymbol{t})\Big) - \mathbb{E}_{q(\boldsymbol{t})}\Big[f(\boldsymbol{t})\Big]$$

3 terms:

Uncertainty
(self-regularization)
e.g., Shannon entropy

Uncertainty

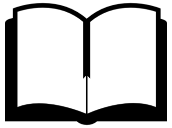
Divergence
(fitness)
e.g., Cross Entropy

Teacher
$q(\boldsymbol{t})$

Student
$p_\theta(\boldsymbol{t})$

Experiences
(exogenous regularizations)
e.g., data examples, rules

Textbook
$f(\boldsymbol{t})$



- Functional derivative

$$F[y(x)+\epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)} \eta(x)\,\mathrm{d}x + O(\epsilon^2)$$

Questions?