泛化理论

序章 泛化理论简介

§0.2.1 ERM 模型

@ 滕佳烨

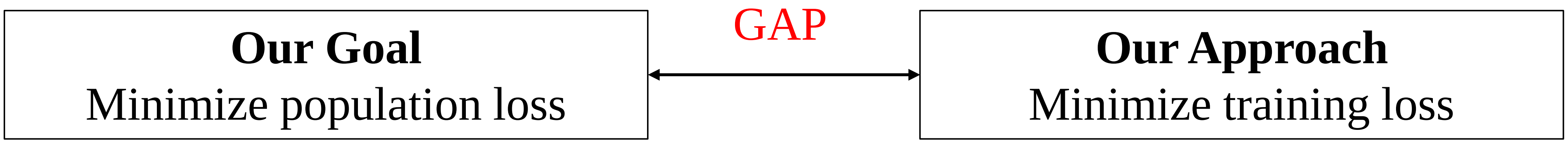
[ref] Understanding Machine Learning: From Theory to Algorithms, Shai Shalev-Shwartz and Shai Ben-David (2014)

Recall:

Generalization: Measuring how model performs on *unseen* data.

Goal: minimize the **population loss**.
Technique: minimize the **training loss** (since we only have finite training samples).

ERM (empirical risk minimization, informally): train a model to minimize the training loss(, and evaluate it via the test loss).



Formal Definition:

Population Distribution

Feature ($x \in \mathcal{X} \subset \mathbb{R}^d$): e.g., an image

Label ($y \in \mathcal{Y} \subset \mathbb{R}$): e.g., cat or dog

Population distribution $(x, y) \sim \mathcal{P}$ unseen

iid sampled →

Training Dataset (x, y)

Training data ($S = \{(x_i, y_i)\}_{i=1}^{n}$):

iid sampled from distribution

Train a model with ERM ↓

Empirical Risk Minimization

Empirical risk: $\mathcal{L}_S(\hat{f}) = \mathbb{P}_{(x,y)\sim U(S)} L(\hat{f}(x), y)$

Algorithm ($\mathcal{A}: \mathcal{X} \times \mathcal{Y} \to \Theta$): take data and return trained parameters

← Training

Model $\mathrm{y} = \mathrm{f}_\theta(\mathrm{x})$

Parameter ($\theta \in \Theta$): to be trained

Model ($f_\theta, \hat{f}$): trained model indexed by the parameter θ

Model performance ↓

Evaluation

population risk $\mathcal{L}_\mathcal{D}(\hat{f}) = \mathbb{P}_{(x,y)\sim\mathcal{D}} L(\hat{f}(x), y)$ unseen

Loss $\boldsymbol{L}(\cdot)$

0-1 loss $L(\hat{y}, y) = \mathbb{I}(\hat{y} \neq y)$

MSE loss $L(\hat{y}, y) = (\hat{y} - y)^2$

Formal Definition:

Our ultimate goal is to train a model with small test loss (population loss).
However, we only train the model on the training set, and attains small training loss.
Does the small training loss ***generalize*** to the test set?

Generalization gap: $\mathcal{L}_{\mathcal{D}}(\hat{f}) - \mathcal{L}_S(\hat{f}) = \mathcal{L}(\hat{f}) - \hat{\mathcal{L}}(\hat{f})$

Note that we have

$$\mathcal{L}(\hat{f}) = \underbrace{[\mathcal{L}(\hat{f}) - \hat{\mathcal{L}}(\hat{f})]}_{\text{Generalization Gap}} + \underbrace{\boxed{\hat{\mathcal{L}}(\hat{f})}}_{\text{Optimization}}$$

Weakness: when there is label noise, $\mathcal{L}(\hat{f})$ does not converge to 0.
Therefore, either generalization gap or optimization loss does not converge to zero.

Generalization Gap: $\mathcal{L}(\hat{f}) - \hat{\mathcal{L}}(\hat{f})$

Note that we have

$$\mathcal{L}(\hat{f}) = \underbrace{[\mathcal{L}(\hat{f}) - \hat{\mathcal{L}}(\hat{f})]}_{\text{Generalization Gap}} + \underbrace{\boxed{\hat{\mathcal{L}}(\hat{f})}}_{\text{Optimization}}$$

Weakness: when there is label noise, $\mathcal{L}(\hat{f})$ does not converge to 0.
Therefore, either generalization gap or optimization loss does not converge to zero.

For example,

Under-para linear reg: small generalization gap ($\sim d/n$), large optimization error ($\sim \frac{\mathrm{n}-d}{\mathrm{n}}\sigma^2$)

Over-para linear reg: large generalization gap ($\geq \sigma^2$), small optimization error ($= 0$)

Therefore, generalization research usually rely on realizable assumption $\inf_{f \in \mathcal{F}} \mathcal{L}(f) = 0$,

or we need to focus on the excess risk (e.g., benign overfitting).

Generalization Gap in Another view: $\mathcal{L}(\hat{f}) - \hat{\mathcal{L}}(\hat{f})$

What we want: small test loss on trained parameter $\hat{\theta}$ compared to the best parameter

$$\mathcal{L}(\hat{f}) - \inf_{f} \mathcal{L}(f)$$

Firstly, under ERM, with good approximation (excess risk)

$$\mathcal{L}(\hat{f}) - \inf_{f} \mathcal{L}(f) = \boxed{\left(\mathcal{L}(\hat{f}) - \inf_{f \in \mathcal{F}} \mathcal{L}(f)\right)} + \boxed{\left(\inf_{f \in \mathcal{F}} \mathcal{L}(f) - \inf_{f} \mathcal{L}(f)\right)}$$

Approximation error

Secondly, with good optimization

$$\mathcal{L}(\hat{f}) - \inf_{f \in \mathcal{F}} \mathcal{L}(f) = [\mathcal{L}(\hat{f}) - \hat{\mathcal{L}}(\hat{f})] + \boxed{[\hat{\mathcal{L}}(\hat{f}) - \hat{\mathcal{L}}(f^*)]} + \boxed{[\hat{\mathcal{L}}(f^*) - \mathcal{L}(f^*)]}$$

Generalization Gap　　ERM, ≤ 0　　Concentration

Take-away messages

(a) The formal definition of machine learning (notations).
(b) Relationship between generalization gap and population risk.
(c) Generalization gap v.s. excess risk (label noise).

All the slides will be available at www.tengjiaye.com/generalization.

@ 滕佳烨

Thanks!