A Future Empirical Risk Minimizers Transfer Risk Performance

Author1, Author2
Affiliation
Univ
City
{Author1, Author2}email@email

Author3
Affiliation
Univ
City
email@email

Abstract

***K*eywords** First keyword · Second keyword · More

1 Introduction

2 Related Work

3 Experimental Setting

3.1 Problem Setting

[1]

$$y = a^{\mathrm{T}}x + \underbrace{\varepsilon}_{\mathcal{N}(0,\sigma=1)} \tag{1}$$

- We are looking at linear regression problem in meta learning setting where each task is represented by the slope (a).
- For the sake of simplicity task distribution is assumed to be originating from the multivariate normal distribution $p_A \sim \mathcal{N}(m\boldsymbol{I}, c\boldsymbol{I})$, where m and c are constants used for parametrizing the experiments.
- Samples drawn are represented by $Z := (x_i, y_i)_{i=1}^{N}$. x_i, y_i and N represent the i^{th} feature, i^{th} label and the number of samples respectively. Moreover, the distribution of these samples are represented by p_Z.
- For an estimator $\hat{a}_N$ trained with N samples from Z we are after the expected error over the whole task distribution.
- Input distribution is given by a multivariate normal distribution $p_x \sim \mathcal{N}(\boldsymbol{0}, b\boldsymbol{I})$
- The expected overall the loss over the task distribution can be formulated as,

$$\int\int\int(\hat{a}_N(Z)^{\mathrm{T}}x - y)^2 p(x,y)dxdyp_Z dZ p_A da. \tag{2}$$

[1] It should be noted that the roman letters represent scalars, lower case bold letters represent vectors and the upper case bold letters represent the matrices. Moreover, calligraphic letters are designated for distributions.(*e.g.* $\{u, N\} \rightarrow$ scalar, $\{\boldsymbol{v}\} \rightarrow$ vector and $\{\boldsymbol{M}\} \rightarrow$ matrix

3.2 Models

3.2.1 Linear Regression Model

Assume we have a model in the form $m(x) := \mathbf{w}^{\mathrm{T}}x + \mathbf{b}$. The free-parameters $\mathbf{w}$ and $\mathbf{b}$ can be estimated upon observing N training samples via squared loss which gives the least squares solution as

$$\mathbf{W} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{3}$$

,where $\mathbf{W}$ represents the stacked free parameters $\mathbf{X}$ represents design matrix and the $\mathbf{y}$ represents the stacked N labels.

3.2.2 Ridge Regression Model

Assuming the same model given in Section 3.2.1 the L_2-regularized version of it can be found as.

$$\mathbf{W} = (\lambda\mathbf{I} + \mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \tag{4}$$

where λ is the regularization parameter that is regularizing both $\mathbf{w}$ and the $\mathbf{b}$.

3.3 Bayes

Knowing the distribution of the tasks give us the exact posterior that one would expect for $N \to \infty$ after starting even from an uninformative prior such as $\mathcal{N} \sim (\mathbf{0}, \mathbf{I})$.

3.4 MAML

This model considers the approach taken in MAML paper. For this simple setting what MAML implicitly does is that it takes the intermadiate model near the vicinity of the mean of the task distribution and leave the model there for a quick adaptation.

Experiments

- All the models are trained with same training points and tested with the same training points. n

4 Results and Discussion

5 Conclusion

Acknowledgments

References