Iterative Approximate Cross Validation in High Dimensions

Sina Alsharif

School of Computer Science and Engineering, University of New South Wales

Thesis A (Term 2, 2023)

Section 1

Background

Background

To set out the notation used throughout this seminar, we can define the Empirical Risk Minimisation (ERM) framework to solve a supervised learning problem.

Background

To set out the notation used throughout this seminar, we can define the Empirical Risk Minimisation (ERM) framework to solve a supervised learning problem.

General Problem Setting

- Input space $\mathcal{X}$ and an output space $\mathcal{Y}$.
- Data is "generated" by a *true* distribution $P(\mathcal{X}, \mathcal{Y})$.
- Aim is to find a mapping $h : \mathcal{X} \to \mathcal{Y}$ (called a hypothesis).
- Denote all possible combinations of input and output space as $\mathcal{D} = \{(X, Y) \in (\mathcal{X}, \mathcal{Y})\}$.

Risk

To measure the error (or loss) we make on a data point, define a function $\ell(h; D_i)$,

$$\ell(h; D) = \sum_{i=1}^{n} \ell(h; D_i)$$

as the loss for the dataset. Examples of ℓ are 0-1 loss (for classification) and a squared error (for regression).

Risk

To measure the error (or loss) we make on a data point, define a function $\ell(h; D_i)$,

$$\ell(h; D) = \sum_{i=1}^{n} \ell(h; D_i)$$

as the loss for the dataset. Examples of ℓ are 0-1 loss (for classification) and a squared error (for regression).

Risk

Define the **risk** of a hypothesis for the complete space of inputs and outputs as,

$$R(h) = \mathbb{E}_{\mathcal{X},\mathcal{Y}}[\ell(h; \mathcal{D})]$$

The optimal hypothesis for the data is,

$$h_{\mathcal{H}} = \arg\min_{h \in \mathcal{H}} R(h)$$

Empirical Risk

As we cannot measure **true** risk, we seek an approximation using the observed data D.

Empirical Risk

We define **empirical risk** of a hypothesis given data D as,

$$R_{\text{emp}}(h; D) = \frac{1}{n} \sum_{i=1}^{n} \ell(h; D_i)$$

The optimal hypothesis *given observed data* D is,

$$h_D = \arg\min_{h \in \mathcal{H}} R_{\text{emp}}(h; D)$$

where $\mathcal{H}$ is a hypothesis space which a *learning algorithm* picks a hypothesis from. We will describe the parameters which describe h as $\theta \in \mathbb{R}^p$, assuming the case of a Generalised Linear Model (GLM).

Issues with ERM

By the (weak) law of large numbers $R_{\text{emp}}(h; D) \to R(h; \mathcal{D})$ as $n \to \infty$, so it is reasonable to assume that h_D converges to a minimiser of true risk.

[1] Refer to the bias-variance (or approximation-estimation) tradeoff.

Issues with ERM

By the (weak) law of large numbers $R_{\text{emp}}(h; D) \to R(h; \mathcal{D})$ as $n \to \infty$, so it is reasonable to assume that h_D converges to a minimiser of true risk. However, if we fit the explicit minimiser of empirical risk, we will not always find the minimiser of true risk [1]. We can restrict the learning algorithm's ability to fully minimise empirical risk through **regularisation**.

[1] Refer to the bias-variance (or approximation-estimation) tradeoff.

Issues with ERM

By the (weak) law of large numbers $R_{\text{emp}}(h; D) \to R(h; \mathcal{D})$ as $n \to \infty$, so it is reasonable to assume that h_D converges to a minimiser of true risk. However, if we fit the explicit minimiser of empirical risk, we will not always find the minimiser of true risk [1]. We can restrict the learning algorithm's ability to fully minimise empirical risk through **regularisation**.

Regularised Empirical Risk

To restrict the hypothesis space, we define the formulation of **regularised empirical risk** as,

$$R_{\text{reg}}(h; D) = R_{\text{emp}}(h; D) + \lambda\pi(\theta)$$

Here, $\pi : \mathbb{R}^p \to \mathbb{R}$ is a regulariser and $\lambda \in \mathbb{R}^+$ is a hyper-parameter to control the strength of regularisation. The solution becomes $(h_D)_\lambda = \arg\min_{h \in \mathcal{H}_\lambda} R_{\text{emp}}(h; D)$ where $\mathcal{H}_\lambda$ is a restricted hypothesis space.

[1] Refer to the bias-variance (or approximation-estimation) tradeoff.

Cross Validation

To avoid "overfitting" to the observed data (i.e blindly minimising empirical risk), we can attempt to define an approximation of true risk to measure the efficacy of a learned hypothesis.

[2] Arlot & Celisse (2008)

Cross Validation

To avoid "overfitting" to the observed data (i.e blindly minimising empirical risk), we can attempt to define an approximation of true risk to measure the efficacy of a learned hypothesis.

The most common way to estimate the true risk of a hypothesis is to run Cross Validation (CV) for a hypothesis . This is where we break the observed data into small subsets to run multiple "validation" experiments (training the data on a subset and testing on an unseen subset).

[2] Arlot & Celisse (2008)

Cross Validation

To avoid "overfitting" to the observed data (i.e blindly minimising empirical risk), we can attempt to define an approximation of true risk to measure the efficacy of a learned hypothesis.

The most common way to estimate the true risk of a hypothesis is to run Cross Validation (CV) for a hypothesis . This is where we break the observed data into small subsets to run multiple "validation" experiments (training the data on a subset and testing on an unseen subset).

One of the most effective methods for risk approximation is Leave One Out Cross Validation (LOOCV) [2]. This method is computationally expensive as we repeat the learning task n times (where n is the size of the observed data).

[2] Arlot & Celisse (2008)

Section 2

Literature Review

Approximate CV Methods

To reduce the computational cost of CV (specifically LOOCV), we turn to Approximate Cross Validation (ACV), which attempts to approximate (rather than solve) individual CV experiments.

LOOCV

The definition of leave-one-out (LOO) regularised empirical risk is,

$$R_{\text{reg}}(\theta; D_{-j}) = \sum_{i=1, i\neq j}^{n} \ell(\theta; D_i) + \lambda\pi(\theta)$$

where we leave out a point with index j for this experiment.

Section 3

Preliminary Work

Section 4

Future Plans