Machine Learning for Applied Physics and High Energy Physics

Lecture 3 The bias-variance trade-off and error representation

3.1 Bias-variance trade-off.

In this lecture we dig into the central principle that underlies much of machine learning: the bias-variance trade-off. We will discuss it in the context of continuous predictions, such as those involved in a regression problem.

Let us consider a data set

$$\mathcal{D} = \{ \vec{x}_n, \vec{y}_n \} \qquad n = 1, \dots, N$$

consisting of N pairs of features and labels. Let us assume that the true data is generated from a noisy model

$$\vec{y} = f(\vec{x}) + \vec{\varepsilon}$$

where ε is normally distributed with mean zero and standard deviation $\sigma_{\varepsilon}^{\to}$. Let us suppose that we have a statistical procedure (e.g. least square regression) for forming a predictor $f(\vec{x}, \hat{\vec{\vartheta}})$ that gives the prediction of our model for a new data point x. This estimator is chosen by minimizing a cost function

$$l_2(\vec{y}, f(\vec{x}, \vec{\vartheta})) = \frac{1}{N}\sum_{n=1}^{N} (\vec{y}_n - f(\vec{x}_n, \vec{\vartheta}))^2 \equiv MSE(\vec{\vartheta})$$

in such a way that

$$\hat{\vec{\vartheta}}_{\mathcal{D}} = \underset{\vec{\vartheta}}{\operatorname{argmin}} \; MSE(\vec{\vartheta})$$

The best-fit parameters $\hat{\vec{\vartheta}}$ are therefore a function of the data set $\mathcal{D} = \{ \vec{x}_n, \vec{y}_n \}$. We would obtain a different $MSE(\vec{\vartheta})$

if we had a different data set in a Universe of possible data sets obtained by drawing N samples from the true data distribution. We denote as expectation value over all of these datasets as $\mathbb{E}_{\mathcal{D}}$. We would also like to average over different instances of the "noise" ε and we denote the expectation value over the noise by $\mathbb{E}_{\varepsilon}$. We can therefore decompose the expected generalisation error as

$$\mathbb{E}_{\mathcal{D},\varepsilon}\left[\ell_2(\vec{y}, f(\vec{x}, \hat{\vec{\vartheta}}_{\mathcal{D}}))\right] = \mathbb{E}_{\mathcal{D},\varepsilon}\left[\frac{1}{N}\sum_{n=1}^{N}(\vec{y}_n - f(\vec{x}_n, \hat{\vec{\vartheta}}_{\mathcal{D}}))^2\right]$$

$$= \mathbb{E}_{\mathcal{D},\varepsilon}\left[\frac{1}{N}\sum_{n=1}^{N}(\vec{y}_n - f(\vec{x}_n) + f(\vec{x}_n) - f(\vec{x}_n, \hat{\vec{\vartheta}}_{\mathcal{D}}))^2\right]$$

$$= \frac{1}{N}\sum_{n=1}^{N}\Big\{\overbrace{\mathbb{E}_{\varepsilon}\left[(\vec{y}_n - f(\vec{x}_n))^2\right]}^{\sigma_\varepsilon^2} + \frac{1}{N}\mathbb{E}_{\mathcal{D},\varepsilon}\left[(f(\vec{x}_n) - f(\vec{x}_n, \hat{\vec{\vartheta}}_{\mathcal{D}}))^2\right]$$

$$+ \frac{2}{N}\underbrace{\mathbb{E}_{\varepsilon}\left[(\vec{y}_n - f(\vec{x}_n))\right]}_{0}\mathbb{E}_{\mathcal{D}}\left[(f(\vec{x}_n) - f(\vec{x}_n, \hat{\vec{\vartheta}}_{\mathcal{D}}))\right]\Big\}$$

0 because we assumed a Gaussian noise with mean zero

$$= \frac{1}{N}\sum_{n=1}^{N}\sigma_\varepsilon^2 + \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\mathcal{D}}\left[(f(\vec{x}_n) - f(\vec{x}_n, \hat{\vec{\vartheta}}_{\mathcal{D}}))^2\right]$$

We can further decompose the second term as

$$\mathbb{E}_{\mathcal{D}}\left[(f(\vec{x}_n) - f(\vec{x}_n, \hat{\vec{\vartheta}}_{\mathcal{D}}))^2\right]$$

$$= \mathbb{E}_{\mathcal{D}}\left[\left\{f(\vec{x}_n) - \mathbb{E}_{\mathcal{D}}[f(\vec{x}_n, \hat{\vec{\vartheta}}_{\mathcal{D}})] + \mathbb{E}_{\mathcal{D}}[f(\vec{x}_n, \hat{\vec{\vartheta}}_{\mathcal{D}})] - f(\vec{x}_n, \hat{\vec{\vartheta}}_{\mathcal{D}})\right\}^2\right]$$

$$= \mathbb{E}_{\mathcal{D}}\left[\left\{f(\vec{x}_n) - \mathbb{E}_{\mathcal{D}}[f(\vec{x}_n, \hat{\vec{\vartheta}}_{\mathcal{D}})]\right\}^2\right] + \mathbb{E}_{\mathcal{D}}\left[\left\{f(\vec{x}_n, \hat{\vec{\vartheta}}_{\mathcal{D}}) - \mathbb{E}_{\mathcal{D}}[f(\vec{x}_n, \hat{\vec{\vartheta}}_{\mathcal{D}})]\right\}^2\right]$$

$$+2\,\mathbb{E}_{\mathcal{D}}\left[\left\{f(\vec{x}_n)-\mathbb{E}_{\mathcal{D}}\left[f(\vec{x}_n,\hat{\vec{\vartheta}}_{\mathcal{D}})\right]\right\}\left\{\underbrace{f(\vec{x}_n,\hat{\vec{\vartheta}}_{\mathcal{D}})-\mathbb{E}_{\mathcal{D}}\left[f(\vec{x}_n,\hat{\vec{\vartheta}}_{\mathcal{D}})\right]}_{0}\right\}\right]$$

$$=\left\{f(\vec{x}_n)-\mathbb{E}_{\mathcal{D}}\left[f(\vec{x}_n,\hat{\vec{\vartheta}}_{\mathcal{D}})\right]\right\}^2+\mathbb{E}_{\mathcal{D}}\left[\left\{f(\vec{x}_n,\hat{\vec{\vartheta}}_{\mathcal{D}})-\mathbb{E}_{\mathcal{D}}\left[f(\vec{x}_n,\hat{\vec{\vartheta}}_{\mathcal{D}})\right]\right\}^2\right]$$

We call

$$\mathrm{Bias}^2=\frac{1}{N}\sum_{n=1}^{N}\left\{f(\vec{x}_n)-\mathbb{E}_{\mathcal{D}}\left[f(\vec{x}_n,\hat{\vec{\vartheta}}_{\mathcal{D}})\right]\right\}^2$$

the measure of the deviation of the expectation value of an estimator from the true value

We call (as usual)

$$\mathrm{Var}=\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\mathcal{D}}\left[\left\{f(\vec{x}_n,\hat{\vec{\vartheta}}_{\mathcal{D}})-\mathbb{E}_{\mathcal{D}}\left[f(\vec{x}_n,\hat{\vec{\vartheta}}_{\mathcal{D}})\right]\right\}^2\right]$$

the measure of the fluctuations of the estimator due to finite-sample effects.

We call noise

$$\mathrm{Noise}=\frac{1}{N}\sum_{n=1}^{N}\sigma_{\varepsilon}^2$$

We therefore arrive at the following decomposition

$$\mathbb{E}_{\mathcal{D},\varepsilon}\left[l_2(\vec{y},f(\vec{x},\hat{\vec{\vartheta}}_{\mathcal{D}}))\right]=\mathrm{Bias}^2+\mathrm{Var}+\mathrm{Noise}$$

The minimum of $\mathbb{E}_{\mathcal{D},\varepsilon}$, which is sought by good ML algorithms, does not generally correspond to the minimum of each of its components.

3.2 Overview of Bayesian inference

To solve a problem using Bayesian methods, we have to specify two functions

- the likelihood $p(\mathcal{D}/\vec{\vartheta})$, which describes the probability of observing a data set $\mathcal{D}$ for given values of the model parameters $\vec{\vartheta}$;
- the prior $p(\vec{\vartheta})$, which describes any (prior) knowledge about the parameters before we collect the data.

The posterior distribution follows from Bayes' theorem

$$p(\vec{\vartheta}/\mathcal{D}) = \frac{p(\vec{\vartheta})\, p(\mathcal{D}/\vec{\vartheta})}{\int d\vec{\vartheta}'\, p(\vec{\vartheta}')\, p(\mathcal{D}/\vec{\vartheta}')}$$

Many common statistical procedures can be cast as MAXIMUM LIKELIHOOD ESTIMATION (MLE). In MLE, one chooses the best-fit parameters $\hat{\vec{\vartheta}}$ that maximise the likelihood (or equivalently, tha minimise the negative log-likelihood)

$$\vec{\vartheta}_{MLE} = \underset{\vec{\vartheta}}{\arg\max} \ln p(\mathcal{D}/\vec{\vartheta}) = \underset{\vec{\vartheta}}{\arg\min} \Big\{ \underbrace{- \ln p(\mathcal{D}/\vec{\vartheta})}_{NLL(\vec{\vartheta})} \Big\}$$

In other words, MLE consists in maximising the probability of seeing the observed data, given a generative model.

The prior distribution is genuinely Bayesian.

- If we do not have any specialised knowledge of $\vec{\vartheta}$ before seeing the data, we should select an "uninformative" prior
- If we do have knowledge of $\vec{\vartheta}$ before seeing the data,

we should choose an informative prior.
Using informative priors tends to decrease the variance of the posterior distribution while, potentially, increasing its bias. This is beneficial if the decrease in variance is larger than the increase in bias. A commonly used prior is the Gaussian prior:

$$p(\vec{\vartheta}/\lambda) = \prod_j \sqrt{\frac{\lambda}{2\pi}}\, e^{-\lambda \vartheta_j^2}$$

which is used to express the assumption that many of the parameters will be small.
We often summarise the posterior probability distribution with a single estimator:

- the mean $\langle \vec{\vartheta} \rangle = \int d\vec{\vartheta}\, \vec{\vartheta}\, p(\vec{\vartheta}/\mathcal{D})$ (Bayes estimate)
- the mode $\hat{\vartheta}_{MAP} = \arg\max_{\vec{\vartheta}} p(\vec{\vartheta}/\mathcal{D})$ (MAXIMUM A POSTERIOR

The reason being that a probability distribution (which is often multi-dimensional) is an object difficult to manipulate.
The mean or MAP estimators come with a credibility interval which quantifies the uncertainty associated to them.
The credibility interval $100(1-\alpha)\%$ is the interval that contains a fraction $1-\alpha$ of the posterior probability.

CREDIBILITY INTERVAL (Bayesian) CONFIDENCE LEVEL (frequentist)

In the expression of the Gaussian prior distribution, λ is a hyperparameter (or nuisance variable). One could

define another prior distribution for λ, usually using an
uninformative prior, and to average the posterior distribution over all choices of λ. This is called a hierarchical prior.