Examen Advanced Machine Learning (AMAL)
Masters DAC et M2A – Sorbonne Université
2023-02-10
Durée 1h30 – Documents papier autorisés

MCQ: answer Y/N

1. The double descent phenomenon in deep NNs characterizes a gradient acceleration technique N
2. In ResNets, skip connections have been introduced to improve the stability of classical NNs Y
3. The gating mechanism in GRUs makes use of a form of skip connection Y
4. The skip gram model is a language model Y
5. The transformers are language models N
6. Transformers are encoder-decoder architectures Y
7. Gaussian processes are trained to predict conditional distributions over functions Y
8. Gaussian processes are fully defined by a mean function and a covariance function N
9. Neural processes allow to predict a mean value and an uncertainty on this mean value Y
10. Neural processes make use of a series of datasets for training Y

The whole exercise is about ensemble methods.

We consider data from a source space $\mathcal{X}$ and label space $\mathcal{Y}$, $l: \mathcal{Y}^2 \rightarrow R^+$ a loss function, a sample is denoted $(x, y) \in \mathcal{X} x \mathcal{Y}$, and its distribution is denoted $p(x, y)$ or p for short. Let us consider a neural network $f(., \theta): \mathcal{X} \rightarrow \mathcal{Y}$, with parameters θ. The objective of training consists in minimizing the generalization error defined as $\boldsymbol{\mathcal{L}(f, \theta) = E_{(x,y)\sim p}[l(f(x, \theta), y)]}$. For that the network will be trained on a dataset D of size N, sampled from p using a training configuration c that may include different sources of randomness such as hyperparameters, etc. We denote $e = (D, c)$ the corresponding learning procedure, $\theta(e) = argmin_\theta \frac{1}{N} \sum_{(x,y)\in D} l(f(x, \theta), y)$ denotes the weights learned through e. The generalization error of functions learned under the learning procedure $e \sim p_e$, with p_e a distribution on e, is defined as $\boldsymbol{\mathcal{L}_e(f) = E_{e\sim p_e}[\mathcal{L}(f, \theta(e))]}$. Being an expectation, $\mathcal{L}_e(f)$ does not depend on a particular training set D or configuration c. It characterizes the performance of the class of estimators learned through e. In the following we will consider real valued functions f and MSE loss i.e. $\boldsymbol{l(f(x, \theta), y) = (f(x, \theta) - y)^2}$.

1. Show that $\mathcal{L}_e(f) = E_{e\sim p_e}[\mathcal{L}(f, \theta(e))] = \mathrm{E}_{(x,y)\sim p}[bias^2(f|(x, y)) + var(f|x)]$ (1)

With $bias(f|(x, y)) = y - \bar{f}(x), var(f|x) = E_e\left[\left(f(x, \theta(e)) - \bar{f}(x)\right)^2\right], \bar{f}(x) = E_e[f(x, \theta(e))]$

2. We now consider an ensemble estimator defined as $f_{ens}(., \theta_{1:M}) = \frac{1}{M}\sum_{m=1}^{M} f(., \theta_m)$ where $\theta_{1:M} = \{\theta_1, \ldots, \theta_M\}$ and each θ_m or equivalently $f(., \theta_m)$ is a sample from the learning procedure. Let us denote $e_{1:M} = \{e_1, \ldots, e_M\}$, so that $E_{e_{1:M}}[.] = E_{e_M}\left[E_{e_{M-1}}\left[\ldots E_{e_1}[.]\right]\right]$.

Let us define the bias and variance for f_{ens}:

$$bias(f_{ens}|(x, y)) = y - E_{e_{1:M}}\left[\frac{1}{M}\sum_{m=1}^{M} f(x, \theta_m)\right]$$

$$var(f_{ens}|x) = E_{e_{1:M}}\left[\left(\frac{1}{M}\sum_{m=1}^{M} f(x, \theta_m) - E_{e_{1:M}}\left[\frac{1}{M}\sum_{m=1}^{M} f(x, \theta_m)\right]\right)^2\right]$$

With $E_{e^M}[.] = E_{e_1 \ldots e_M}[.]$ the expectation over the distribution of the training procedures $e_1 \ldots e_M$.

2.1 Preliminaries, show:

$$E_{e_{1:M}}\left[\frac{1}{M}\sum_{m=1}^{M} f(x, \theta_m)\right] = \frac{1}{M}\sum_{m=1}^{M} E_{e_m}[f(x, \theta_m)]$$

$(\sum_{m=1}^{M} a_m - b_m)^2 = \sum_{m=1}^{M}(a_m - b_m)^2 + \sum_{m=1}^{M}\sum_{m' \neq m}(a_m - b_m)(a_{m'} - b_{m'})$

Then using (1) for f_{ens}, show:

$$\mathcal{L}_{e_{1:M}}(f_{ens}) \triangleq E_{e_{1:M}}[\mathcal{L}(f_{ens}, \theta_{1:M})] = E_{(x,y)\sim p}[B^2 + \frac{1}{M}V + \frac{M-1}{M}C]$$

Where:

For simplification in the following $f_m(.)$ denotes $f(., \theta_m)$

$$B = \frac{1}{M}\sum_{m=1}^{M} bias(f_m|(x,y))$$

$$V = \frac{1}{M}\sum_{m=1}^{M} var(f_m|x)$$

$$C = \frac{1}{M(M-1)}\sum_m \sum_{m' \neq m} cov(f_m, f_{m'}|x) \text{ with } cov(f, f'|x) = E_{e,e'}[(f - E_e[f])(f' - E_{e'}[f'])]$$

2.2 Let us now suppose that the f_m are identically distributed. This means that the bias, variance and expectation are the same for all the f_m. Let us denote respectively $bias(f|(x,y)), var(f|x), \bar{f}$ these different variables – defined as above (question 1). Show that:

$$E_{e_{1:M}}[\mathcal{L}(f_{ens}, \theta_{1:M})] = E_{(x,y)\sim p}[bias^2(f|(x,y)) + \frac{1}{M}var(f|x) + \frac{M-1}{M}cov(f, f'|x)] \quad (2)$$

Where $cov(f, f'|x) = E_{e,e'}[\left(f(x, \theta(e)) - \bar{f}(x)\right)\left(f(x, \theta(e')) - \bar{f}(x)\right)]$.

2.3 Interpret this last result. What should be the property of the f functions for minimizing this generalization error?

2.4 We will now suppose that the f_m are independent, meaning that $cov(f, f'|x) = 0$, what is the expression of $E_{e_{1:M}}[\mathcal{L}(f_{ens}, \theta_{1:M})]$? What is the benefit of using an ensemble method compared to a single function estimator f?

3. We will now consider an alternative to building ensembles, that emerged with the use of large pre-trained networks. We start from a pre-trained network (for example on ImageNet). Then we fine tune it on a given dataset using some learning procedure e. For different but related learning procedures e, one will get different set of parameters $\theta(e)$ that will be close one to the other. Suppose we fine tune M functions $f_m, m = 1 \ldots, M$ Let us define the ensemble function:

$$f_{wa} \triangleq f(., \theta_{wa}) \text{ with } \theta_{wa} = \frac{1}{M}\sum_{m=1}^{M}\theta_m$$

This means that the ensemble is defined by a unique function f_{wa}, its weights being the average of the weights of the M functions f_m. θ_{wa} is supposed close to θ_m, $\forall m$. The f_m are supposed identically distributed so that (2) applies.

3.1 Using a first order Taylor expansion of $f(., \theta_m)$ around $f(., \theta_{wa})$, show that $f_{ens} - f_{wa} = O(\Delta_{1:M}^2)$ with $\Delta = \max_m \|f_m - f_{wa}\|_2$

3.2 Using a zeroth order Taylor expansion w.r.t. its first argument of $l(f_{ens}(x), y)$ around $f_{wa}(x)$ to show that $l(f_{ens}(x), y) = l(f_{wa}(x), y) + O(\Delta_{1:M}^2)$

3.3 Using this result show that $\mathcal{L}(f_{wa}, \theta_{1:M}) = \mathcal{L}(f_{ens}, \theta_{1:M}) + O(\Delta_{1:M}^2)$, with

$$\Delta_{1:M} = \max_m \|\theta_m - \theta_{wa}\|_2$$

Show then $E_{e_{1:M}}[\mathcal{L}(f_{wa}, \theta_{1:M})] = E_{e_{1:M}}[\mathcal{L}(f_{ens}, \theta_{1:M})] + O(\bar{\Delta}^2)$ with $\bar{\Delta}^2 = E_{e^M}[\Delta_{1:M}^2]$

3.4 Interpret this result. What could be the benefit of the f_{wa} estimator compared to the f_{ens} estimator?

Note - Taylor expansion. Let $f: \mathbb{R}^p \to \mathbb{R}$ a differentiable function, the Taylor expansion of order 1 around point a is: $f(a + h) = f(a) + \nabla f(a).h + O(||h||^2)$