Principles of Machine Learning

Methodology

Academic year 2021/2022

Queen Mary University of London

EXERCISES

EXERCISE ♯1. A fraction F of a dataset is used for training a regression model, leaving the remaining $1 - F$ for testing. Discuss the impact on the training MSE and the test MSE of:

- Values of F close to 1.
- Values of F close to 0.

EXERCISE ♯2. In a dataset consisting of 1,000 samples, it has been shown that a 70-30 split (i.e. 70% of the samples used for training, 30% for validation) will provide a good estimation of the deployment performance of the trained models. If the dataset size increases to 10,000 samples, what split would you suggest?

EXERCISE ♯3. Discuss the following scenarios:

- Your company will start manufacturing self-driving cars that use computer vision algorithms to detect pedestrians. Instead of creating a computer vision team, your company decides to buy an external solution. What would be a good approach to identify the best solution?
- After testing many solutions, you discover that none of them are good enough. Specifically, you realise that all of them fail when presented with images taken in adverse weather conditions. You are still convinced that it is better to have someone else developing the solution for you. What could you do to help external teams to develop the algorithm that you need?
- Your company is a leading developer of computer vision algorithms. You have found out that a self-driving car company is looking for a computer vision algorithm to detect pedestrians and has made a dataset available for anyone to build a solution. What would you do with this dataset?
- After developing and validating 10 different algorithms, you obtain the following validation performances 0.91, 0.5, 0.45, 0.92, 0.89, 0.905, 0.7, 0.4, 0.52 and 0.9, where 0 is the lowest performance and 1 the highest. What would you do with these results?

EXERCISE ♯4. Consider four polynomial models of degrees 1, 2, 3 and 4 and three datasets consisting of 8 samples $\{(x_i, y_i), 1 \leq i \leq N\}$ and created as follows:

- **Dataset 1**: x_i is randomly drawn from a uniform distribution $U(0, 1)$ and $y_i = 3x_i$.
- **Dataset 2**: x_i is randomly drawn from a uniform distribution $U(0, 1)$ and $y_i = 3x_i + n_i$, where n_i is randomly drawn from a Gaussian distribution $N(2, \sigma^2)$.
- **Dataset 3**: x_i and y_i are both randomly drawn from a uniform distribution $U(0, 1)$.

Assume that the first four samples of each dataset are used to train the four polynomial models using the MMSE criterion and the remaining four samples are used for validation. Answer the following questions for each dataset:

- How will the training MSE and validation MSE change with the order of the polynomial?
- What do you expect the coefficients of each MMSE polynomial to be?
- What would be the impact of increasing the number of training and validation samples from 4 to 50, on the MSE values and the coefficients of each MMSE polynomial?

EXERCISE $\sharp$5. Consider a dataset consisting of 100 samples $\{(x_i, y_i), 1 \leq i \leq 100\}$. The following options are proposed to create the training dataset and the test dataset:

- The first 50 samples in the dataset will be used for training, the remaining for testing.
- 50 numbers between 1 and 100 will be randomly generated without repetition. Samples whose index i is one of those numbers will be used for training, the remaining 50 samples will be used for testing.
- Samples whose index i is an even number will be used for training, whereas samples indexed by an odd number will be used for testing.

Which option would you implement and why?

EXERCISE $\sharp$6. Let x_k be a time series representing the changes in the price of a stock, where k denotes the time point at which x_k was recorded. Your goal is to build a model that predicts a price on a given day based on the N most recent, previous prices. Assume that your time series consists of K consecutive prices.

- After choosing a reasonable value for N, you are considering using the first K_1 prices in the time series for training and the remaining $K - K_1$ prices for validation. Under which assumptions will the validation performance be a good estimation of the deployment performance?
- A second option you are considering is to extract all the segments of size $N + 1$ from the time series and assign them randomly to either the training stage or the validation stage. Under which assumptions will the validation performance be a good estimation of the test performance?
- What factors will you consider when choosing the value of N?