

Mini-project overview

Lecture 16

Changho Suh

January 29, 2024

Plan for Week 2

1. Mini-project # 1 (Day 6):

Learn how to: architect a machine learning project;
do data processing (via pandas);
improve a model.

2. Mini-project # 2 (Day 7):

Learn how to: do data processing for RNN;
improve a model.

3. Proposal (Days 8,9,10):

Learn how to write a proposal.

Write your own proposal and do rehearsal.

Do final presentation.

Mini-project #1

(Day 6)

Guideline for a machine learning project

Four steps illustrated via:

“STAR” method

Will explain the STAR method while describing the contents of mini-project #1.

1. Situation

Describe the **situation**:

1. Project context (배경)
2. Challenge (도전적 과제)

1. Situation: Project context

Mini-project #1 is about:

Vehicle manufacturing

There are many car options.

The vehicle test time varies significantly across different options.

1. Situation: Challenge

The challenge that we will address is about:

Time taken for testing a vehicle

Some car options may require long test time.

→ Incurs significant cost and environmental issues.

Hence: Crucial to figure out car options taking short test time.

2. Task: Come up with an ML task

Will consider an ML task that Mercedes-Benz took into account:

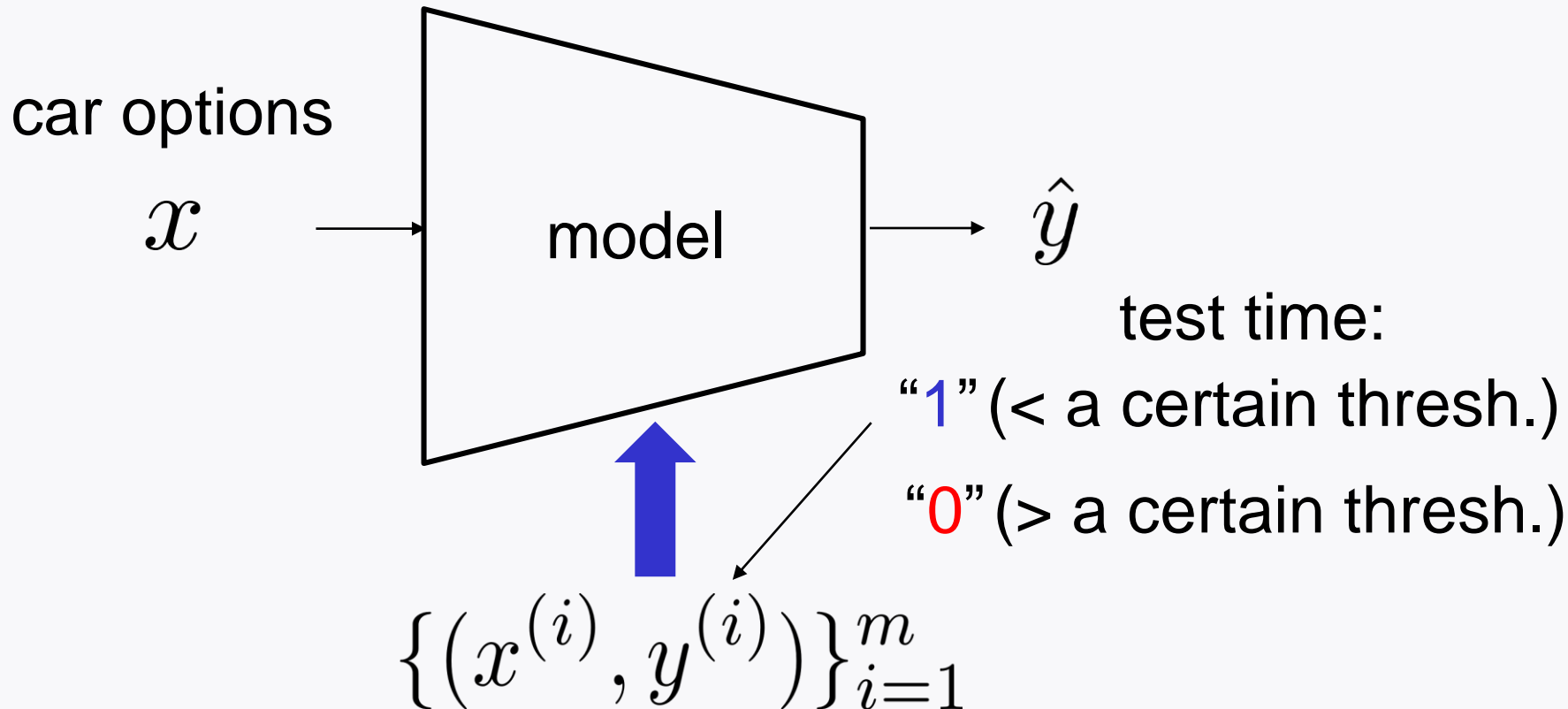
Task: Predict vehicle test time.

A binary classifier:

Outputs 1	if test time < a certain threshold
0	if test time > a certain threshold

2. Task: ML model

Test-time prediction:



3. Activity: Describe activities

3.1. Dataset:

What are x and y in detail? What is m ?

How to collect if not prepared yet?

3.2. Model:

Choose a model (LS, LR, DNN, CNN, RNN, RF etc)

Explain the rationale behind the choice.

3.3. Target performance:

Accuracy (classification) MSE (regression)

Explain the rationale behind the choice using domain knowledge

3.1. Activity: Dataset

Source: Mercedes-Benz provides **4209** examples

Data: anonymized **376** car options

Label: test time

Raw data is in **csv** file.



mercedes_test.csv

x_1	x_2	x_3	y
K	v	at	130.81
K	t	av	88.53
az	w	n	76.26

Will learn how to **load data from csv file**

3.1. Activity: How to load MB dataset

To this end, will employ:



3.1. Activity: Pre-processing



for data manipulation & analysis

x1	...	y
K	...	130.81
K	...	88.53



x1_aa	...	x1_K	...	x1_z	...	y
0	...	1	...	0	...	0
0	...	1	...	0	...	1

1. One hot encoding of categorical data $\in \{aa, \dots, z\}$
2. Binary label w/ a median threshold

3.1. Activity: Train/val/test sets

of examples: **4209**

Recall: It belongs to the middle range:

$$1,000 \leq m \leq 10,000$$

Hence: Will take the following split ratio:

80:10:10

3.1. Activity: Data frame → numpy array



for large, multi-dimensional arrays and matrices

x1_aa	...	x1_K	...	y
0	...	1	...	0
0	...	1	...	1



$$X = \begin{pmatrix} 0 & \dots & 1 & \dots \\ 0 & \dots & 1 & \dots \end{pmatrix}$$

$$y = \begin{pmatrix} 0 & 1 \end{pmatrix}$$

3.2. Activity: Model selection

Will try three models for exercise purpose:

LS, LR and DNN

3.3. Activity: Target performance

LS: ~89% accuracy (w/ regularization)

LR: ~89% (w/ regularization)

3.3. Activity: Target performance

DNN: ~89% (w/ some techniques)

Techniques that we will apply:

Regularization, early stopping

He's initialization

Hyperparameter search w/ cross validation:

of layers, learning rate, ...

4. Result: Describe results (impacts)

Explain what your ML model can do:

1. Quantitatively: In terms of numbers (money)
2. Qualitatively: In many other aspects

Three more things to learn from MP1

1. Plotting

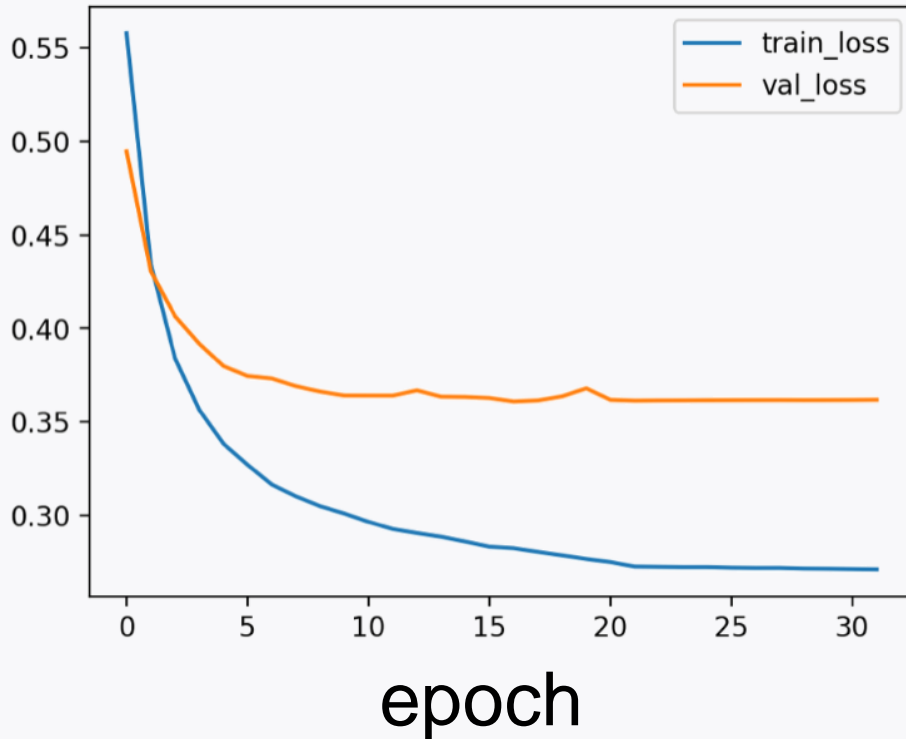
2. Logging

Track intermediate results

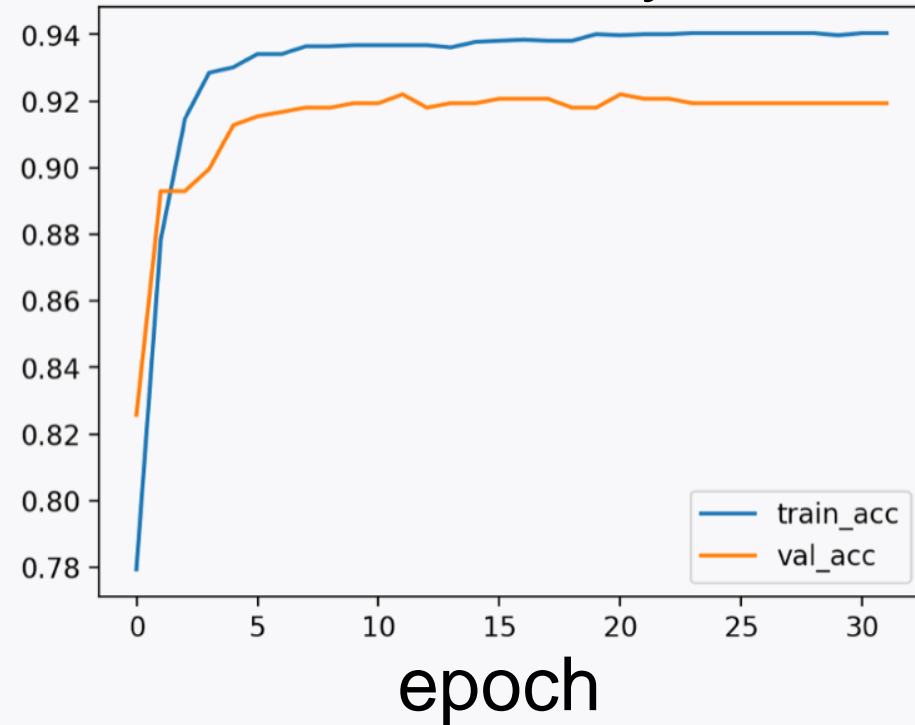
3. Saving

Plotting: A traditional way

loss

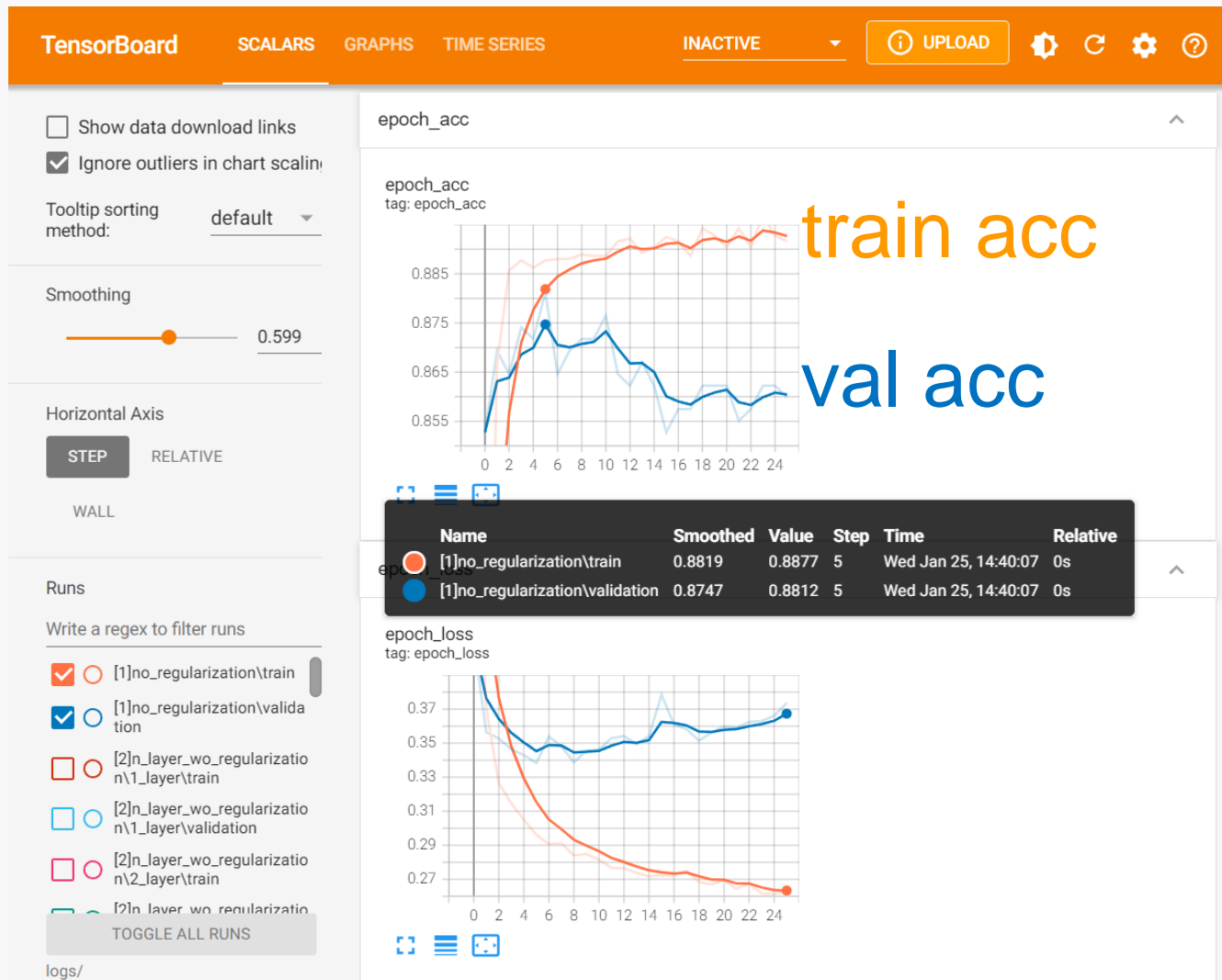


accuracy



Plotting: A more convenient way

Using TensorBoard (visualization tool)



Logging: Track intermediate results

Example:

(LR) regularization_factor=0.9 → acc: 89%

(DNN) learning rate: 0.01, batch size: 512 → acc: 91%

To this end, will use **pandas**.

Saving

Save parameters of trained models.

Sklearn models: `from joblib import dump`

Keras models: `model.save()`

Look ahead

Figure out how mini-project #2 is organized.