Mini-project overview

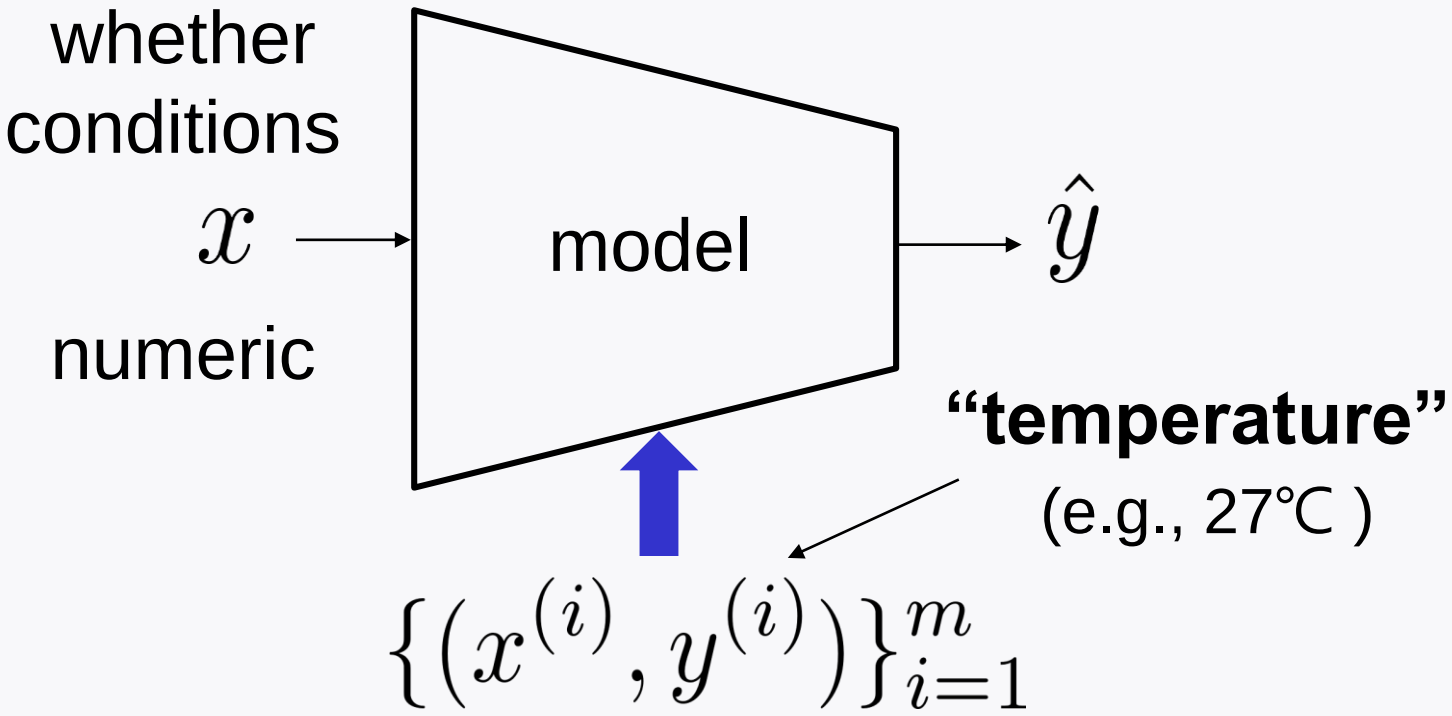
Lecture 17

Changho Suh

January 29, 2024

Mini-project #2

Task: Weather prediction



Jena climate dataset

Data:

weather features collected from 2009 to 2016 $\in \mathbf{R}^{14}$

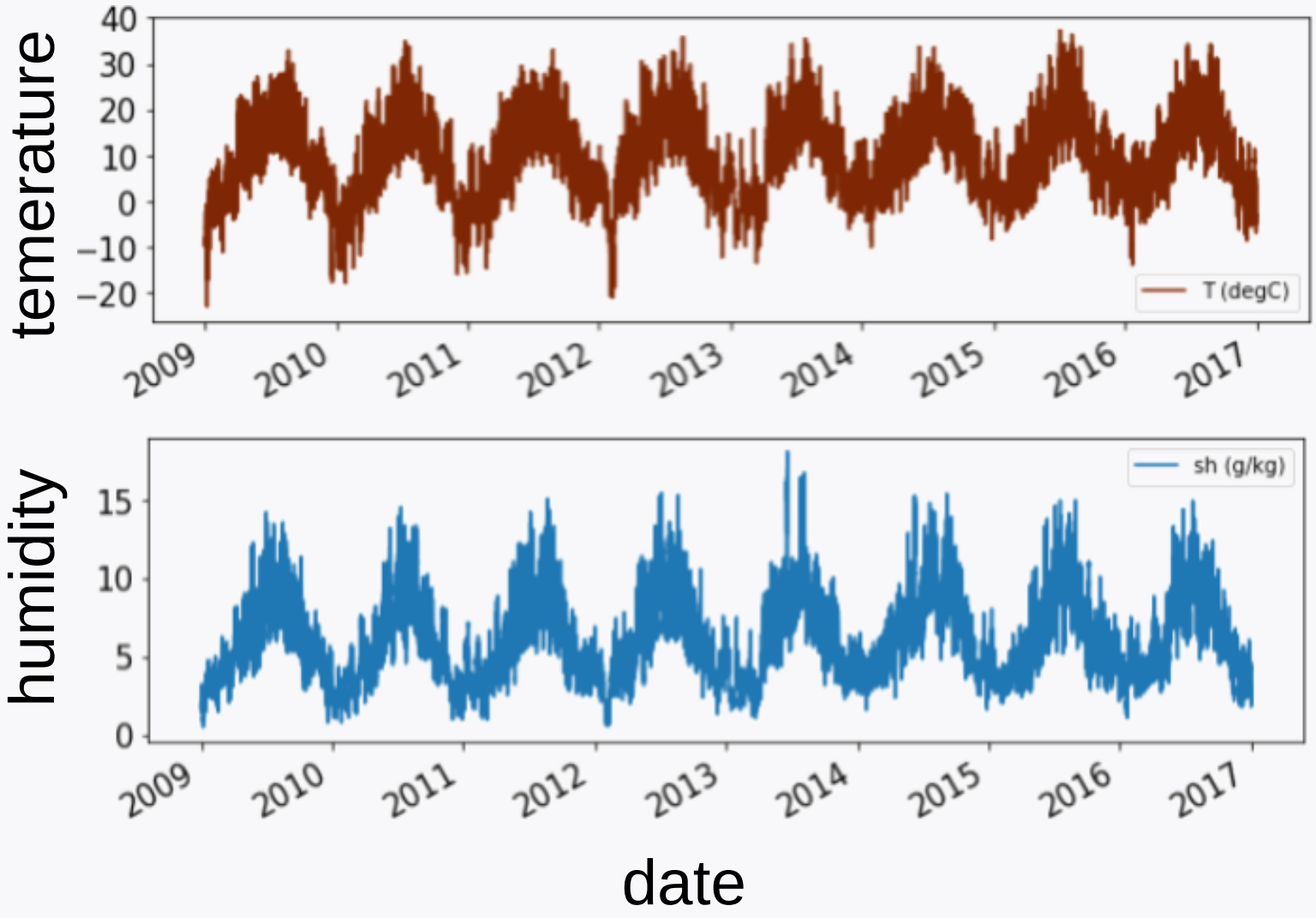
Example: pressure, humidity, …

Label: celcius temperature $\in \mathbf{R}$

Measured in Jena, Germany, every 10 minutes:

420,551 samples

Visualization



Data organization

Load data in **csv** file using pandas.

Note: **Time series** data

Split dataset into **train/val/test** sets with:

7:2:1 (in *chronological* order)

To this end, will use `train_test_split`

```
from sklearn.model_selection import train_test_split
```

Model selection

Will try two models:

DNN and **RNN**

Performance measure

Will use another measure instead of MSE:

Root-mean-square error (RMSE):

$$\sqrt{\frac{1}{m_{\text{test}}}\sum_{i=1}^{m_{\text{test}}}\|y^{(i)}-\hat{y}^{(i)}\|^2}$$

Normalized RMSE:

$$\frac{\text{RMSE}}{\sqrt{\frac{1}{m_{\text{test}}}\sum_{i=1}^{m_{\text{test}}}\|y^{(i)}-\mu\|^2}} \leftarrow \sigma_{\text{test}}$$

Target performance

1. DNN: NRMSE ~ 0.09

2. RNN (LSTM): NRMSE ~ 0.08

Advanced techniques we will apply

Regularization, early stopping

Weight initialization

Learning rate decaying

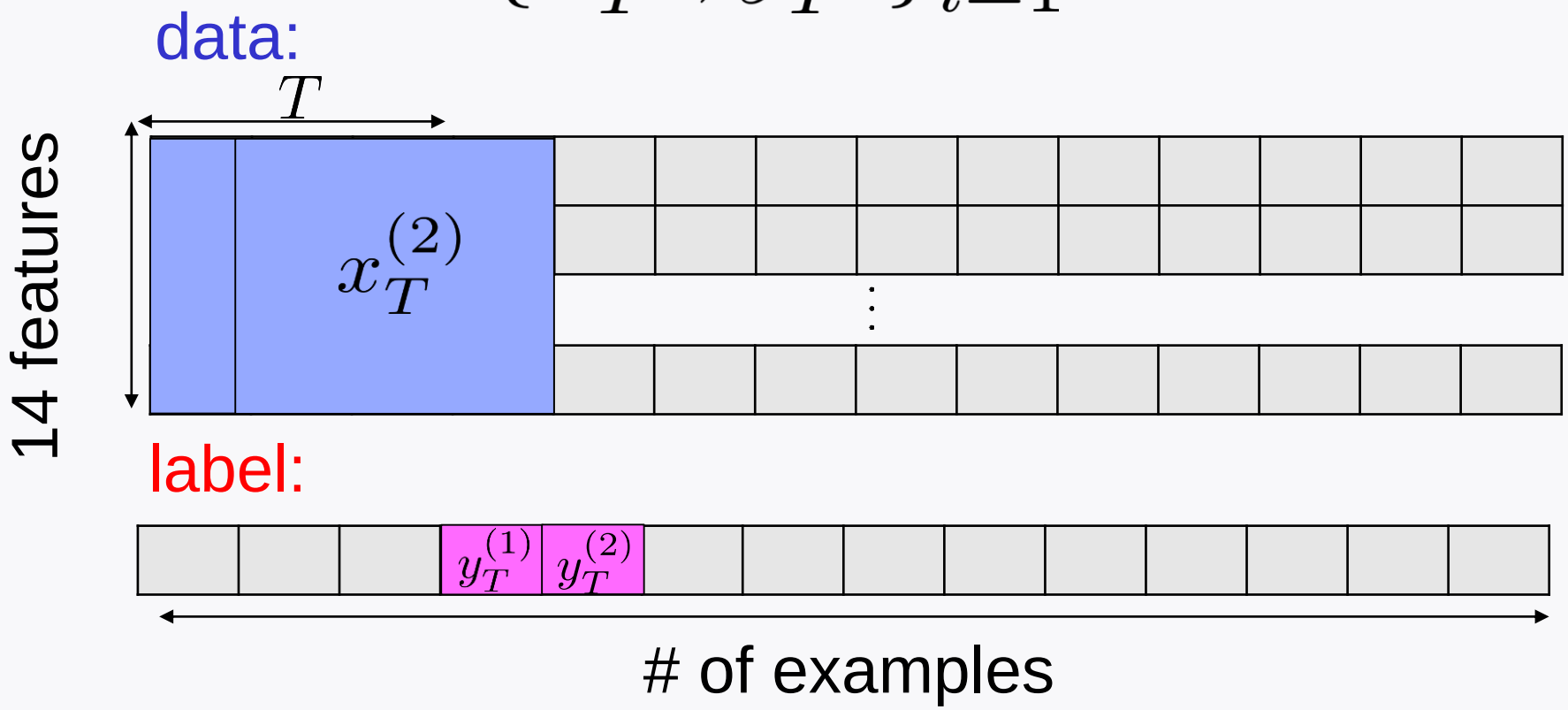
Hyperparameter search:

T (window size), learning rate, …

Hyperparameter: time window *T*

Will generate time series dataset according to *T:*

$$\{x_T^{(i)}, y_T^{(i)}\}_{i=1}^{m_T}$$



Saving

1. “loss” curves
2. log files
3. parameters of trained models