AARHUS
UNIVERSITY
SCHOOL OF ENGINEERING

LESSON 3: End-to-end ML

CARSTEN EIE FRIGAARD

AUTUMN 2021

End-to-End Machine Learning Project

In this chapter you will work through an example project end to end, pretending to be a recently hired data scientist at a real estate company. Here are the main steps you will go through:

1. Look at the big picture.
2. Get the data.
3. Discover and visualize the data to gain insights.
4. Prepare the data for Machine Learning algorithms.
5. Select a model and train it.
6. Fine-tune your model.
7. Present your solution.
8. Launch, monitor, and maintain your system.

Working with Real Data

When you are learning about Machine Learning, it is best to experiment with real-world data, not artificial datasets. Fortunately, there are thousands of open datasets to choose from, ranging across all sorts of domains. Here are a few places you can look to get data:

1. The example project is fictitious; the goal is to illustrate the main steps of a Machine Learning project, not to learn anything about the real estate business.

Agenda
End-to-end Machine Learning

1. Admin
   - ▶ O1 Feedback grupper?
2. Sidste del fra L02
   - ▶ Performance metrics (scores)
   - ▶ Opgave: `L02/performance_metrics.ipynb`
3. Generel repetition af § 2,
   - ▶ Opgave: `L03/supergruppe_diskussion.ipynb`
4. End-to-end repetition,
   - ▶ via 'The Map'

Super-group Discussion

Present your oral resume for the [SG]

Supergruppe diskussion

§ 2 "End-to-End Machine Learning Project" [HOML]

Genlæs kapitel § 2 (eksklusiv "*Create the Workspace*" og "*Download the Data*"), og forbered mundtlig præsentation.

Forberedelse inden lektionen

Een eller flere af gruppe medlemmer forbereder et mundtligt resume af § 2:

- i skal kunne give et kort mundtligt resume af hele § 2 til en anden gruppe (på nær, som nævnt, Create the Workspace og Download the Data),
- resume holdes til koncept-plan, dvs. prøv at genfortælle, hvad de overordnede linier i kaptilerne i [HOML].

Lav et kort skriftlig resume af de enkelte underafsnit, ca. 5 til 20 liners tekst, se "TODO"-template herunder (MUST, til O1 aflevering).

Kapitler (incl. underkapitaler):

- *Look at the Big Picture*
- *Get the Data (eksklusiv Create the Workspace og Download the Data),*
- *Discover and Visualize the Data to Gain Insights,*
- *Prepare the Data for Machine Learning Algorithms,*
- *Select and Train a Model,*
- *Fine-Tune Your Model,*
- *Launch, Monitor, and Maintain Your System,*
- *Try It Out!.*

På klassen

Supergruppe [SG] resume af § 2 End-to-End, ca. 30 til 45 min.

- en supergruppe [SG], sammensættes af to grupper [G], on-the-fly på klassen,
- hver gruppe [G] forbereder og giver en anden gruppe [G] et mundtligt resume af § 2 til en anden gruppe,
- tid: ca. 30 min- sammenlagt, den ene grupper genfortæller første haldel af § 2 i ca. 15 min., hvorefter den anden gruppe genfortæller resten i ca. 15 min.
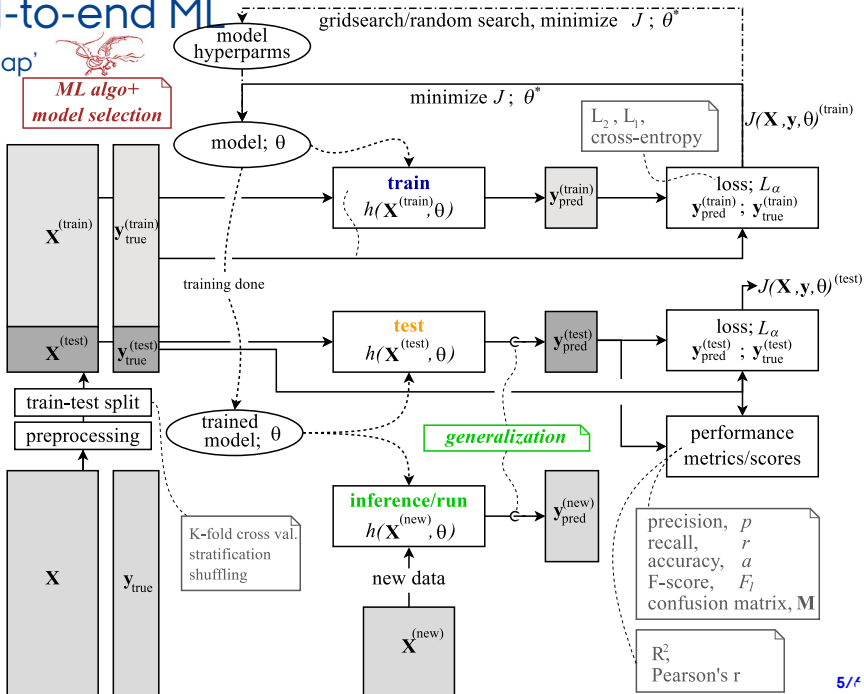
End-to-End Machine Learning Project

End-to-end ML
'The Map'

ML algo+ model selection

gridsearch/random search, minimize J ; θ^*

model hyperparms

minimize J ; θ^*

L_2, L_1, cross-entropy

$J(\mathbf{X}, \mathbf{y}, \theta)^{(\text{train})}$

model; θ

train
$h(\mathbf{X}^{(\text{train})}, \theta)$

$\mathbf{y}_{\text{pred}}^{(\text{train})}$

loss; L_α
$\mathbf{y}_{\text{pred}}^{(\text{train})}$; $\mathbf{y}_{\text{true}}^{(\text{train})}$

$\mathbf{X}^{(\text{train})}$ $\mathbf{y}_{\text{true}}^{(\text{train})}$

training done

$J(\mathbf{X}, \mathbf{y}, \theta)^{(\text{test})}$

test
$h(\mathbf{X}^{(\text{test})}, \theta)$

$\mathbf{y}_{\text{pred}}^{(\text{test})}$

loss; L_α
$\mathbf{y}_{\text{pred}}^{(\text{test})}$; $\mathbf{y}_{\text{true}}^{(\text{test})}$

$\mathbf{X}^{(\text{test})}$ $\mathbf{y}_{\text{true}}^{(\text{test})}$

train-test split

preprocessing

trained model; θ

generalization

performance metrics/scores

inference/run
$h(\mathbf{X}^{(\text{new})}, \theta)$

$\mathbf{y}_{\text{pred}}^{(\text{new})}$

K-fold cross val.
stratification
shuffling

new data

$\mathbf{X}$ $\mathbf{y}_{\text{true}}$

$\mathbf{X}^{(\text{new})}$

precision, p
recall, r
accuracy, a
F-score, F_1
confusion matrix, $\mathbf{M}$

R^2
Pearson's r

ML Supervised Learning, Train/Test