DHBW Ravensburg

Intelligence Engineering

Solving Real World Problems At Scale

Prof. Dr.-Ing. Mark Schutera | November 9, 2026 · *drifting stream*

LECTURE IV

Training Data

This lecture in one slide

IV Training Data

Sampling
Labeling
Class Imbalance
Data Augmentation

Let's move from a systems and infrastructure perspective to a more,

Data Science perspective.

Let's move from a systems and infrastructure perspective to a more,

Data Science perspective.

Let's move from a systems and infrastructure perspective to a more,
Data Science perspective.
MACHINE LEARNING
DATA ENGINEERING
What is moving, not how.

MACHINE LEARNING
DATA

Complex, messy, tricky, surprising

* complex
* messy
* tricky
* surprising.
MACHINE LEARNING
DATA
... but there are techniques to acquire or create good data.

* complex
* messy
* tricky
* surprising.
MACHINE LEARNING
DATA
... but there are techniques to acquire or create good data.
* sampling
* labeling
* balancing
* extending

1

Sampling

Sampling from all possible real world data

Into different splits

Into batches

Into individual samples

Sampling is necessary or plainly beneficial

SAMPLING

is necessary due to resource limitations, or it's plainly beneficial to the objective.

Sampling can introduce biases

Non-probability sampling: so much selection bias

Convenience: take what is easily accessible

Snowball: select based on already selected

Judgment: experts decide, not cherry picking

Quota: based on quotas for certain domains

Do LLMs suffer and deploy any of these?

PROBABILITY
SAMPLING

Random: equal probabilities to be selected

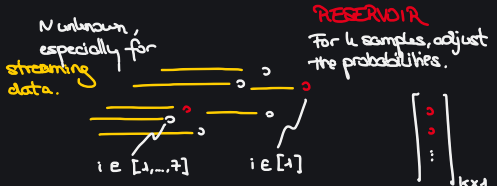
Rare events might drop

Stratified: groups, then random sampling

Clustering is a problem in its own right

Weighted: configure selection probabilities

N unknown, especially for streaming data



Reservoir: k samples, adjust the probabilities

Replace if i is in 1 to k

Further sampling methods: let me know

2

Labeling

Most models today are trained supervised



LABELING

Most models today are trained SUPERVISED

... is not an auxiliary task, this is a core function. The quality and quantity of labeled data is not to be underestimated.

Labeling is a core function, not an auxiliary task

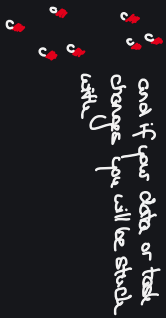
Labeling by hand

Humans with domain knowledge

Humans with
Domain
Knowledge
... motivation
... time.
... data clearance.
LABELING
BY HAND

Difficult, slow and expensive

Humans with Domain Knowledge
... motivation
... time.
... data clearance.
LABELING BY HAND
€
... is difficult, slow and expensive.
but we wanted to iterate fast, always.



If your data or task changes, you are stuck

Relabeling, or even recollecting data

Two annotators, two labels

Different levels of annotator expertise

Annotator variability gives rise to

$\tilde{y}_{A1}$
$\tilde{y}_{A2}$
3 Different levels of annotator expertise
LABELING BY HAND
and annotator variability
gives rise to ...
... AMBIGUITY or MULTIPLICITY

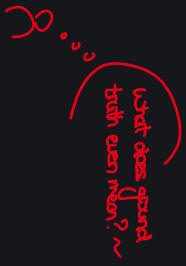


It's a nine

It's a six

...this is extremely common.
And rises even with the complexity of the task or domain.

Extremely common, rises with complexity

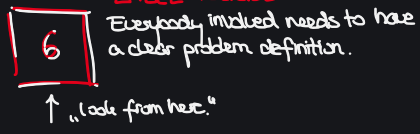


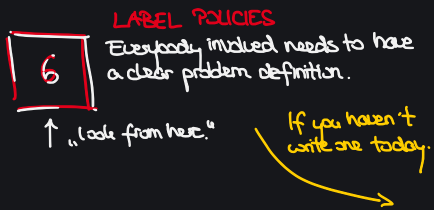
What does ground truth even mean?

Is human level performance the gold standard?

TALK TO ME
let's work on how to speed up robot operations, often first trained on human demonstrations.

Talk to me: speed up robot operations


LABEL POLICIES
6
Everybody involved needs to have a clear problem definition.
„I code from here."


6
LABEL POLICIES
Everybody involved needs to have a clear problem definition.
„I code from here."
If you haven't write one today.

ANNOTATOR VARIABILITY

Embrace the complexity and the distribution.
Report it and put your model in context.

ANNOTATOR VARIABILITY

Embrace the complexity and the distribution.
Report it and put your model in context.

+) Some label noise has been shown to help generalization.

Some label noise has been shown to help

... DATA LINEAGE

data is valuable so we will often tap into multiple sources.

Data lineage: we tap into multiple sources

... DATA LINEAGE

data is valuable so we will often tap into multiple sources.

And we will make sure to version our data and validate / test those sources separately.

They might have labelled different things

Further reading: frugal labeller, smolVLA

LABELING
BY THE NATURE OF
THINGS

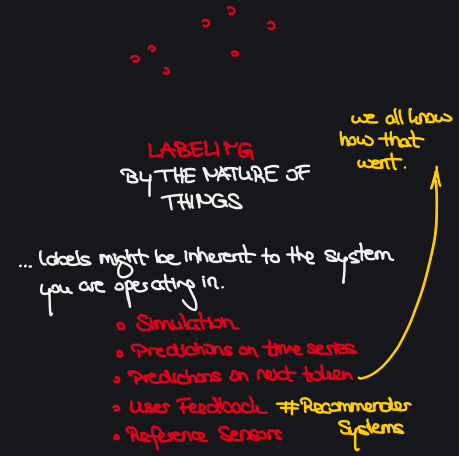
... labels might be inherent to the system you are operating in.



LABELING
BY THE NATURE OF
THINGS

... labels might be inherent to the system you are operating in.

- Simulation
- Predictions on time series
- Predictions on next token
- User Feedback
- Reference Sensors

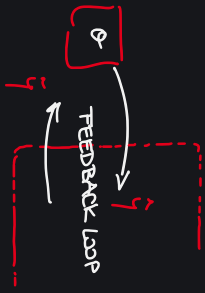

LABELING BY THE NATURE OF THINGS
we all know how that went.
... labels might be inherent to the system you are operating in.
Simulation
Predictions on time series
Predictions on next token
User Feedback #Recommender Systems
Reference Sensors

Framing: many things can collect feedback

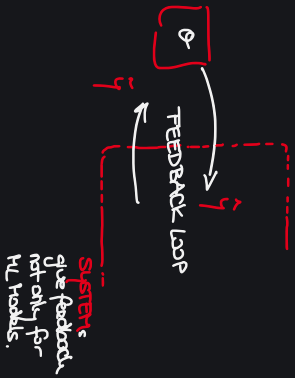
Press like if you agree



No feedback is an implicit label



Feedback loop



Systems give feedback not only for ML models

SYSTEMS give feedback not only for ML models.

LOOP TIME
Can be from seconds to weeks and are non-configurable, for others you can engineer it.

Loop time: from seconds to weeks

HANDLING THE
LACK OF
LABELS

HANDLING THE
LACK OF

LABELS

REFRESH THISE
UNSUPERVISED LEARNING SKILLS,
AND METHODS.

WEAK SUPERVISION

Leverage heuristics to generate noisy labels, from label functions.

WEAK SUPERVISION

Leverage heuristics to generate noisy labels, from label functions.

∴ Combine, denoise, reweight.

A few real labels guide development

So why do we need an ML model anyways?

Semi supervision: structural assumptions

High probability predictions, smoothed, iterate

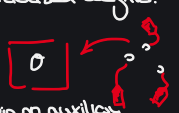
$\tilde{y}' \approx \hat{y}$
SELF SUPERVISION
Leverage structural assumptions. Label as seeds are powerfull.
#Clustering
#self-supervision
... high probability predictions get one-hot smoothed. Iterate.

TRANSFER LEARNING
Zero-Shot or Fine-Tuning from initialized weights.
Train on auxiliary base-task, then tune for x-shot.

TRANSFER LEARNING
zero-shot or Fine-Tuning from initialized weights.
Train on auxiliary base-task, then tune or x-shot.
#anchor, inject, ..
LoRA

TRANSFER LEARNING

Zero-Shot or Fine-Tuning from initialized weights.



Train on auxiliary base task, finetune or x-shot.

anchor, inject, ... LoRA

With the rise of FOUNDATION MODELS zero-shot becomes even more powerful. and distillation is central.

Active learning: samples near the boundary

Increase label efficiency

Cues are usually uncertainty measures

Samples near the decision boundaries count most.
ACTIVE LEARNING
Label data that is most useful to your model dev.
Increase (dev) efficiency.
Ques are usually uncertainty measures.
READ UP on Monte Carlo Dropout and epistemic uncertainty.

Efficiency might seem irrelevant with foundation models and abundant compute. But is handy for real-time model adaptation.

Handy for real time model adaptation

Efficiency might seem irrelevant with foundation models and abundant compute. But is handy for real-time model adaptation.

#Embodied AI.

3

Class Imbalance

CLASS IMBALANCE
refers to a substantial difference
in the numbers of samples per domain.

CLASS IMBALANCE

refers to a substantial difference in the numbers of samples per domain.

Domains A and B

CLASS IMBALANCE

refers to a substantial difference in the numbers of samples per domain.

due to..
.. poor sampling
.. probabilities
.. label errors

CLASS IMBALANCE

refers to a substantial difference in the numbers of samples per domain.

due to..

.. poor sampling
.. probabilities
.. label errors

Take a moment to reflect on the impact on overall metrics.

It also is a challenge to learning

It also is a challenge to learning

It also is a challenge to learning

- Due to weak or entirely missing signals.

End up in a weak local optimum

More likely with more complex problems

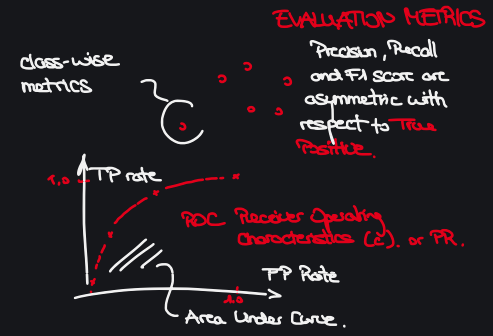
HANDLING
CLASS IMBALANCE

on data-level.

EVALUATION METRICS

EVALUATION METRICS
class-wise
metrics

Precision, recall, F1 are asymmetric


EVALUATION METRICS
class-wise metrics
Precision, Recall and F1 score are asymmetric with respect to True Positive.
1.0
TP rate
ROC Receiver Operating Characteristics (e). or PR.
TP Rate
1.0
Area Under Curve.

RESAMPLING

to modify the underlying distribution.

Resampling modifies the underlying distribution

RESAMPLING

to modify the underlying distribution.

#Undersampling

Tomek deletes closest to minority class first.

Undersampling: Tomek deletes closest first

x' = gamma times x

Oversampling: copy, augment, synthesize

4

Data Augmentation

AUGMENTATION.

Label preserving perturbation, online during training.

Augmentation: label preserving perturbation

References

1 Chip Huyen. *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications*. Sebastopol, CA: O'Reilly Media, 2022.

2 Mark Schutera et al. "Methods for the Frugal Labeler: Multi-Class Semantic Segmentation on Heterogeneous Labels". In: *PLoS ONE* 17.2 (2022), e0263656.

3 Mark Schutera and Stefan Bühler. *!Imperio, smolVLA: The Implications of Data Poisoning on Open Source Robotics*. 2026. arXiv: 2607.04146.