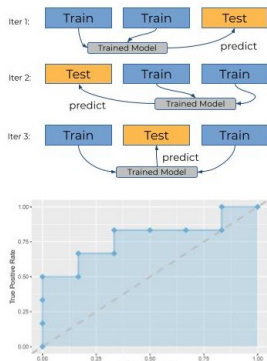


Introduction to Machine Learning

Evaluation: Introduction and Remarks



Learning goals

- Understand the goal of performance estimation
- Know the definition of generalization error
- Understand the difference between outer and inner loss

PERFORMANCE ESTIMATION

- After training our model, we are naturally interested in its **performance**.
- We recall what supervised learning is about:

$$\mathcal{I} : \mathbb{D} \times \Lambda \rightarrow \mathcal{H}, \quad (\mathcal{D}, \lambda) \mapsto \hat{f}_{\mathcal{D}, \lambda}$$

- \mathcal{I} minimizes the empirical risk resulting from $L(y, f)$.
- This so-called **inner loss**, however, is only a statistical proxy to what we are really interested in: the **true expected loss** for new, unlabeled data.
- After all, we chose our model precisely so it would be loss-minimal on the data we trained it on, but we cannot hope for it to perform equally well on general data from \mathbb{P}_{xy} .

→ Evaluation based on the inner loss would be **optimistically biased**.

GENERALIZATION ERROR

- The true expected loss for a model $\hat{f}_{\mathcal{D}_n, \lambda}$, learned on $\mathcal{D}_n \sim \mathbb{P}_{xy}$, is measured w.r.t. to previously unseen data $(\mathbf{x}, y) \sim \mathbb{P}_{xy}$.
- We refer to this as **generalization error** or **outer loss**:

$$\text{GE}(\hat{f}_{\mathcal{D}_n, \lambda}) := \mathbb{E}_{(\mathbf{x}, y) \sim \mathbb{P}_{xy}} \left[L \left(y, \hat{f}_{\mathcal{D}_n, \lambda}(\mathbf{x}) \right) \right]$$

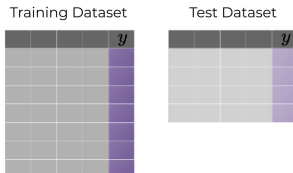
- The goal of **performance evaluation** is to measure $\text{GE}(\hat{f}_{\mathcal{D}_n, \lambda})$.
→ As \mathbb{P}_{xy} is unknown to us, we can only estimate it.
- Bild (Caro)

INNER VS OUTER LOSS

- Supervised learning thus implies the following dichotomy:
 - **Learning**: train $\hat{f}_{\mathcal{D}_n, \lambda}$ minimizing *inner* loss
 - **Evaluation**: evaluate $\hat{f}_{\mathcal{D}_n, \lambda}$ estimating *outer* loss
- Beyond evaluating a single learner, the outer loss lends itself to comparing different types of learners, or learners with varying hyperparameter configurations λ .
- Ideally, we have **inner loss = outer loss**.
- This is not always possible – sometimes we use inner losses that are hard to optimize or do not even specify one directly, as in:
 - Logistic regression: minimization of binomial loss
 - k-NN: no explicit loss minimization
- On the other hand, there are some special metrics for evaluation, such as those derived from ROC curves.

TRAINING AND TEST DATA

- For reliable estimates of $GE(\hat{f}_{\mathcal{D}_n, \lambda})$ we need **test data** that are independent of the data we trained our model on.
- Such test sets are not always available, but we will learn about techniques of **resampling** that allow us to carve out test sets from the data at hand.



- Note that this paradigm is somewhat different from traditional statistical model diagnosis where models are judged by their goodness-of-fit rather than their generalization ability.