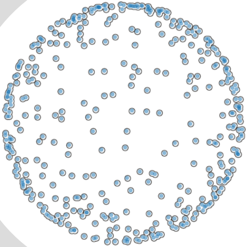
WiSe 2024/25

Machine Learning for Data Science

Lecture by G. Montavon

FREIE UNIVERSITÄT BERLIN



Lecture 8b

Reproducibility (cont.)

Outline

Classical Model Selection

- Holdout
- Cross-validation (CV)

Limits of Holdout/CV for non-iid. Data

- Domain shifts
- Learning domain-invariant representations
- Spurious correlations
- Detecting spurious correlations

Interpretable Models

- Sparse linear models

Part 1

Classical Model Selection

Motivations

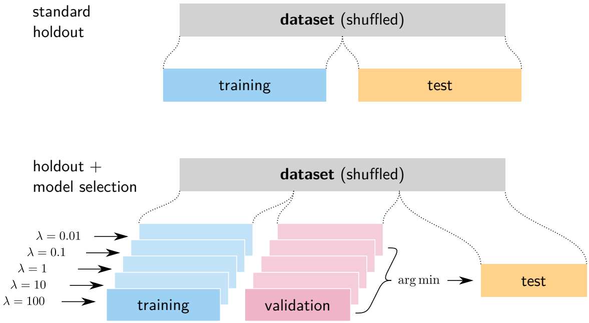
Observation:

- Feature reduction/regularization provides a way of controlling model complexity, but there are no generally applicable heuristics for setting the hyperparameters (e.g. λ).
- VC dimension provides a bound on the generalization error, however, the bound is quite loose. Choosing regularization hyperparameters in a way that optimizes the VC bound will likely lead to suboptimal choices.

Question:

- Can we select regularization hyperparameters by directly assessing their performance on the data?

Holdout with Model Selection

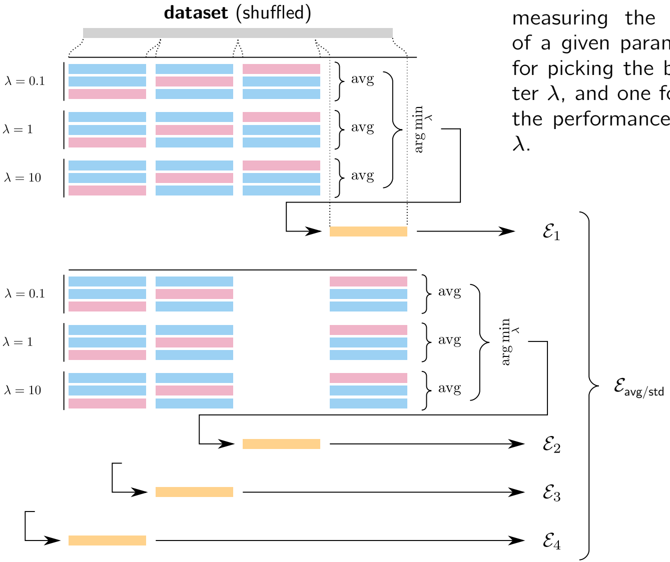


Notes:

- If we want to both select and evaluate the model, the dataset has to be split in *three* parts. A first part for training the model, a second part for choosing the best hyperparameter, and a last part for evaluating the selected model.
- Splitting the dataset in three implies that we have less data for training/selection/evaluation.

Question: Can we apply cross-validation to be more data efficient?

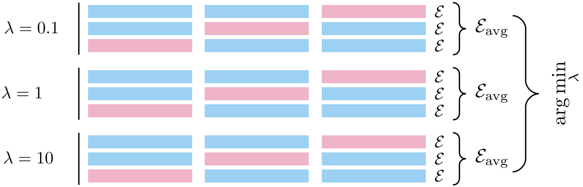
Cross-Validation with Model Selection



Three nested loops: one for measuring the performance of a given parameter λ, one for picking the best parameter λ, and one for evaluating the performance of the best λ.

Cross-Validation with Model Selection

A note on the '$\arg\min_\lambda$':



- The '$\arg\min_\lambda$' finds the optimal parameter $\lambda^\star$, but there are a several models that have been trained with $\lambda = \lambda^\star$.

Several possibilities:

- Train a new model from scratch with the whole training+validation data using this parameter $\lambda^\star$.
- Take all existing models that have been trained with $\lambda^\star$ and construct (if possible) an average model, e.g. by averaging the model parameters or the predictions.

Model Selection / Evaluation

The model selection/evaluation techniques above are usually applied to minimize/evaluate the error of a model, however, they are also applicable to other metrics.

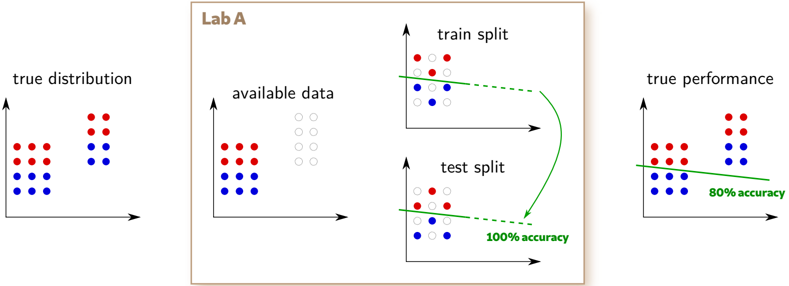
Examples:

- Calinsky-Harabasz or Dunn's index of some clustering model (where the number of clusters or other clustering hyperparameter needs to be selected).
- Correlation coefficient found by a CCA or related model (where some subset of input features or some robustness parameter σ_n^2 needs to be selected).

Part 2

Limits of Holdout / Cross-Validation

Unrepresentative Data



- Holdout/cross-validation are only expected to work well if we assume that available data is drawn i.i.d. from the 'true' distribution of interest.
- In practice, it often doesn't hold.

Example: Histological Images

Example: Analysis of histopathological images

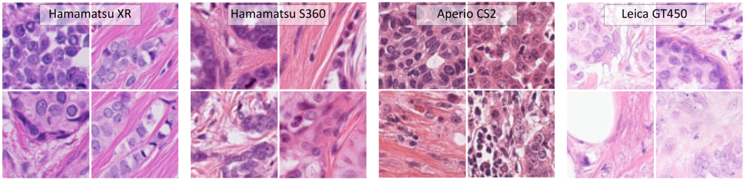
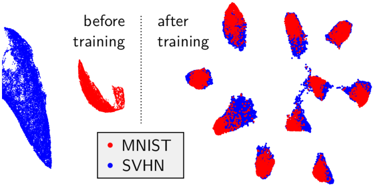


Image source: Aubreville et al. Quantifying the Scanner-Induced Domain Gap in Mitosis Detection. CoRR abs/2103.16515 (2021)

- Different labs may use different acquisition devices for collecting histopathological data. Different acquisition devices may have different resolutions, sharpness, or color profiles.
- Such variations can strongly affect the predictions of a model.
- The problem can be mitigated by learning a *domain invariant representation* $\boldsymbol{x} \mapsto \Phi(\boldsymbol{x})$, and perform the analysis on top of it.

Learning Domain Invariant Representations



Moment matching:

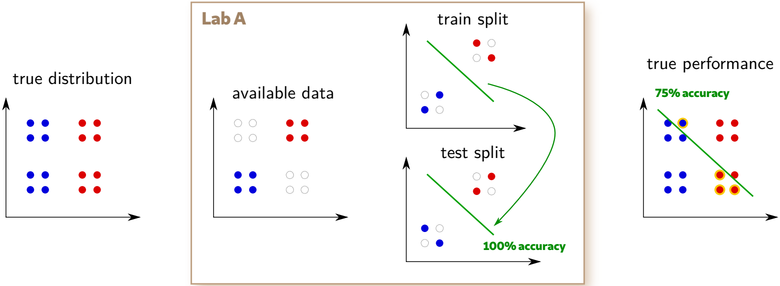
- If $\mathcal{D}, \mathcal{D}'$ are the two domains, perform the analysis in representation space and penalize solutions in which different domains have different means, e.g.

$$\min_{\boldsymbol{w}} \quad \mathcal{E}(\boldsymbol{w}) + \lambda \cdot (\mathrm{E}_{\mathcal{D}}[\boldsymbol{w}^\top \boldsymbol{x}] - \mathrm{E}_{\mathcal{D}'}[\boldsymbol{w}^\top \boldsymbol{x}])^2$$

  The penalty can be enhanced to include higher-order moments.

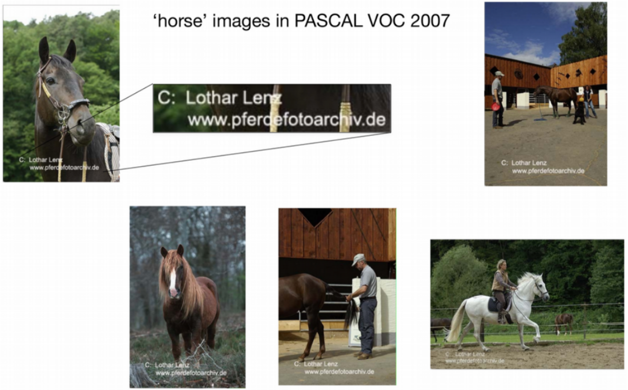
- In practice, more powerful tools exist to match distributions in representation space, such as the Wasserstein distance (e.g. [1]).

Unrepresentative Data



- Lack of access to the true distribution may give rise to *spurious correlations*. I.e. one feature correlates with red/blue groups only because of the limited amount of data available.
- Very common in practice (see e.g. [2]).

Example: Images & Horses

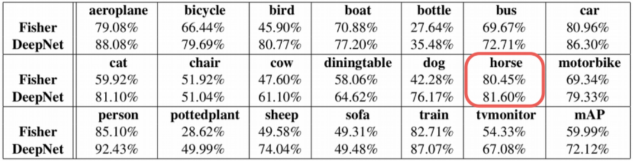


- Lots of images of horse in the PASCAL VOC 2007 dataset come with a copyright tag. Images of other classes do not have such copyright tag.
- Copyright tags may enable to correlate images with the concept 'horse' on this data, but such experiment may not be reproducible on a different dataset without copyright tags.

Example: Images & Horses

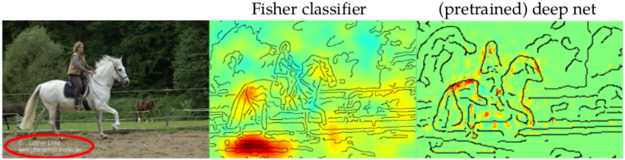
- Test set accuracy doesn't give much information on whether the model bases its decision on the correct features or exploits the spurious correlation.

Fisher Vector Classifier vs. DeepNet pretrained on ImageNet



	aeroplane	bicycle	bird	boat	bottle	bus	car
Fisher	79.08%	66.44%	45.90%	70.88%	27.64%	69.67%	80.96%
DeepNet	88.08%	79.69%	80.77%	77.20%	35.48%	72.71%	86.30%
	cat	chair	cow	diningtable	dog	horse	motorbike
Fisher	59.92%	51.92%	47.60%	58.06%	42.28%	80.45%	69.34%
DeepNet	81.10%	51.04%	61.10%	64.62%	76.17%	81.60%	79.33%
	person	pottedplant	sheep	sofa	train	tvmonitor	mAP
Fisher	85.10%	28.62%	49.58%	49.31%	82.71%	54.33%	59.99%
DeepNet	92.43%	49.99%	74.04%	49.48%	87.07%	67.08%	72.12%

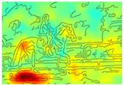
- Only an inspection of the decision structure by the user (e.g. using LRP heatmaps) enables the detection of the flaw in the model [3, 4].



Improving Reproducibility

Improving Reproducibility:

- Provide the data along with the meta-data. Then other people can verify the conditions for reproducibility.
- Spurious correlations can be detected by heatmaps or analyzing the weights of the model.
- Meta-data can be useful to correct some properties of the model or the dataset, e.g. reweighting the data distribution, forcing the representation to be invariant to some subgroups, etc.



Question:

- The model's decision strategy is easier to inspect/verify by a human if the model relies on few features. Can we force the model to rely on few features?

Part 3

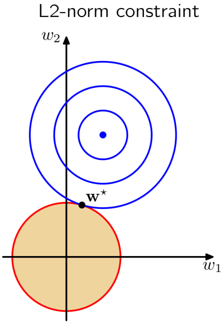
Making the Model Sparse

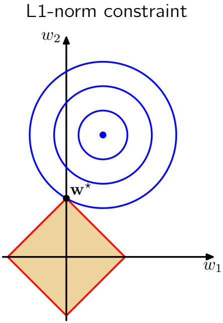
L1-Norm Regularization

Idea:

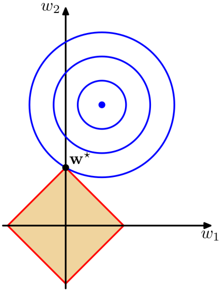
- Reduce the number of features used by the model by constraining the L1-norm of the parameters (instead of the L2-norm).
- For example, for linear regression:

$$\text{Minimize} \quad \mathrm{E}[(\boldsymbol{w}^\top \boldsymbol{x} - y)^2] \quad \text{s.t.} \quad \|\boldsymbol{w}\|_1 \leq C.$$





L1-Norm Regularization



$$\text{Minimize} \quad \mathrm{E}[(\boldsymbol{w}^\top \boldsymbol{x} - y)^2] \quad \text{s.t.} \quad \|\boldsymbol{w}\|_1 \leq C$$

Note:

- The L1 norm can be casted into 2^d linear constraints. For example, for $d = 3$, we have:

$$\begin{array}{ll} +w_1 + w_2 + w_3 \leq C & +w_1 + w_2 - w_3 \leq C \\ +w_1 - w_2 + w_3 \leq C & +w_1 - w_2 - w_3 \leq C \\ -w_1 + w_2 + w_3 \leq C & -w_1 + w_2 - w_3 \leq C \\ -w_1 - w_2 + w_3 \leq C & -w_1 - w_2 - w_3 \leq C. \end{array}$$

- This can be solved by quadratic programming solvers (e.g. cvxopt), however, it is computationally infeasible when d is large.

A Practical Algorithm [5]

Observation:

- It is easy to (1) test whether all exponentially many constraints are satisfied (just test $\|\boldsymbol{w}\|_1 \leq C$) and (2) find the most strongly violated constraint (it is $\boldsymbol{h}^\top \boldsymbol{w} \leq C$ with $\forall_{t=1}^d : \; h_t = \text{sign}(w_t)$).

Idea:

- Iteratively add violated constraints and re-optimize until $\|\boldsymbol{w}\|_1 \leq C$.

Algorithm:

- Initialize the set of constraints to $S \leftarrow \emptyset$.
- **Repeat** until $\|\boldsymbol{w}\|_1 \leq C$
  - Optimize the objective $\mathcal{E}(\boldsymbol{w})$ subjects to constraints S
  - Find the most violated constraint s among all 2^d constraints.
  - $S \leftarrow S \cup \{s\}$
- **return** $\boldsymbol{w}$

- In practice, it was shown that the algorithm converges after somewhere between $0.5d$ and $0.75d$ constraints have been added.

Summary

Summary

- Model selection can be achieved objectively by witholding some of the data from training, and using the rest of the data for selection.
- Cross-validation can be applied both for model evaluation and model selection. If one wishes to do both, one needs to use a nested cross-validation procedure.
- Holdout/cross-validation are working most reliably when the data is drawn iid. from the 'true' distribution.
- Non-iid data is very frequent in practice due to heterogeneities in the construction of datasets. It may give rise to domain shifts and spurious correlations.
- Various techniques are needed to achieve reproducibility in a non-iid. setting, sometimes requiring additional penalty terms, or requiring user feedback. Making models interpretable is important in order to enable user feedback.

References

L. Andéol, Y. Kawakami, Y. Wada, T. Kanamori, K. Müller, and G. Montavon.
Learning domain invariant representations by joint wasserstein distance minimization.
Neural Networks, 167:233–243, 2023.

C. S. Calude and G. Longo.
The deluge of spurious correlations in big data.
Foundations of Science, 22(3):595–612, Mar. 2016.

S. Lapuschkin, A. Binder, G. Montavon, K. Müller, and W. Samek.
Analyzing classifiers: Fisher vectors and deep neural networks.
In *CVPR*, pages 2912–2920. IEEE Computer Society, 2016.

S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller.
Unmasking clever hans predictors and assessing what machines really learn.
Nature Communications, 10(1), Mar. 2019.

R. Tibshirani.
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288, Jan. 1996.