Generalization and the problem of leakage

Nico Formánek

March 28th, 2024

> *You make a fool of yourself if you declare that you have discovered something, when all you are observing is random chance.*

(David Colquhoun)

ML is an inductive method

> *This observation [i.e. learning from data is impossible] was first made (in somewhat different form) by the philosopher David Hume over 200 years ago, but even today many mistakes in machine learning stem from failing to appreciate it.*

(Pedro Domingos)

ML is an inductive method

> *This observation [i.e. learning from data is impossible] was first made (in somewhat different form) by the philosopher David Hume over 200 years ago, but even today many mistakes in machine learning stem from failing to appreciate it.*

(Pedro Domingos)

Why call it machine "learning" then?

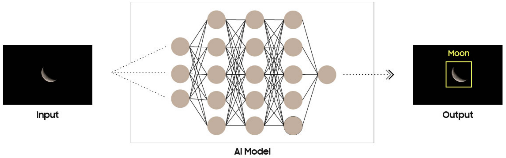
Inductive assumptions

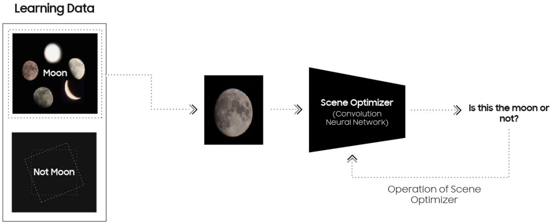
- To learn something you have to make inductive assumptions.
- These assumptions cannot be inferred from the data alone.
- These assumptions must be justified! They must be good, apropriate assumptions.

From abstract to concrete: Moon gate

Figure: Image improvement of moon shot with Samsung smart phone

How Samsung claims it works





How reddit claims it works

r/Android

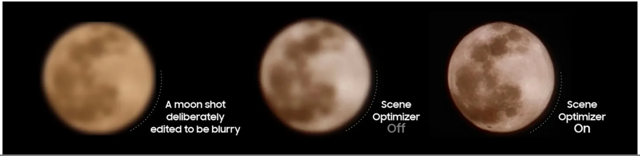
Posted by u/ibreakphotos 4 days ago 2 7 2 2 2 3

12.7k

Samsung "space zoom" moon shots are fake, and here is the proof

Many of us have witnessed the breathtaking moon photos taken with the latest zoom lenses, starting with the S20 Ultra. Nevertheless, I've always had doubts about their authenticity, as they appear almost too perfect. While these images are not necessarily outright fabrications, neither are they entirely genuine. Let me explain.

There have been many threads on this, and many people believe that the moon photos are real ([inputmag]) - even MKBHD has claimed [in this popular youtube short] that the moon is not an overlay, like Huawei has been accused of in the past. But he's not correct. So, while many have tried to prove that Samsung fakes the moon shots, I think nobody succeeded - until now.



Data processing inequality

If random variables have mutual information $I(X;Y)$ then no way of (conditionally independently) processing Y can increase that information. $I(X;f(Y)) \leq I(X;Y)$

Data processing inequality

If random variables have mutual information $I(X;Y)$ then no way of (conditionally independently) processing Y can increase that information. $I(X;f(Y)) \leq I(X;Y)$

- If the information is not there, there is no way to conjecture it.
- What does this mean for techniques such as imputation, oversampling etc.?

Generalization

A classifier generalizes well iff its true risk is close to its empirical risk.

True risk: the average loss with respect to the data generating distribution (which you don't know).
Emp. risk: the loss on the training set (which is easy to calculate).

Moral: we must somehow estimate the true risk.

Cross-validation

CV is widely used in statistics and ML to give generalization guarantees from the data alone.

- Training Test split
- Training Test Validation split
- Leave-one-out CV
- k-fold CV
- Bootstrapping

Some problems with CV

> *Cross-validation is a widely used technique to estimate prediction error, but **its behavior is complex and not fully understood**. Ideally, one would like to think that cross-validation estimates the prediction error for the model at hand, fit to the training data. We prove that this is not the case for the linear model fit by ordinary least squares; rather it estimates the average prediction error of models fit on other unseen training sets drawn from the same population.*

(Bates et al.)

- CV does not do what many people think it does.
- The statistical error of CV cannot be estimated without knowledge of the data generating distribution.
- In ML we generally don't have knowledge of the data generating distribution.

Some problems with CV

> *Cross-validation is a widely used technique to estimate prediction error, but **its behavior is complex and not fully understood**. Ideally, one would like to think that cross-validation estimates the prediction error for the model at hand, fit to the training data. We prove that this is not the case for the linear model fit by ordinary least squares; rather it estimates the average prediction error of models fit on other unseen training sets drawn from the same population.*

(Bates et al.)

- CV does not do what many people think it does.
- The statistical error of CV cannot be estimated without knowledge of the data generating distribution.
- In ML we generally don't have knowledge of the data generating distribution.

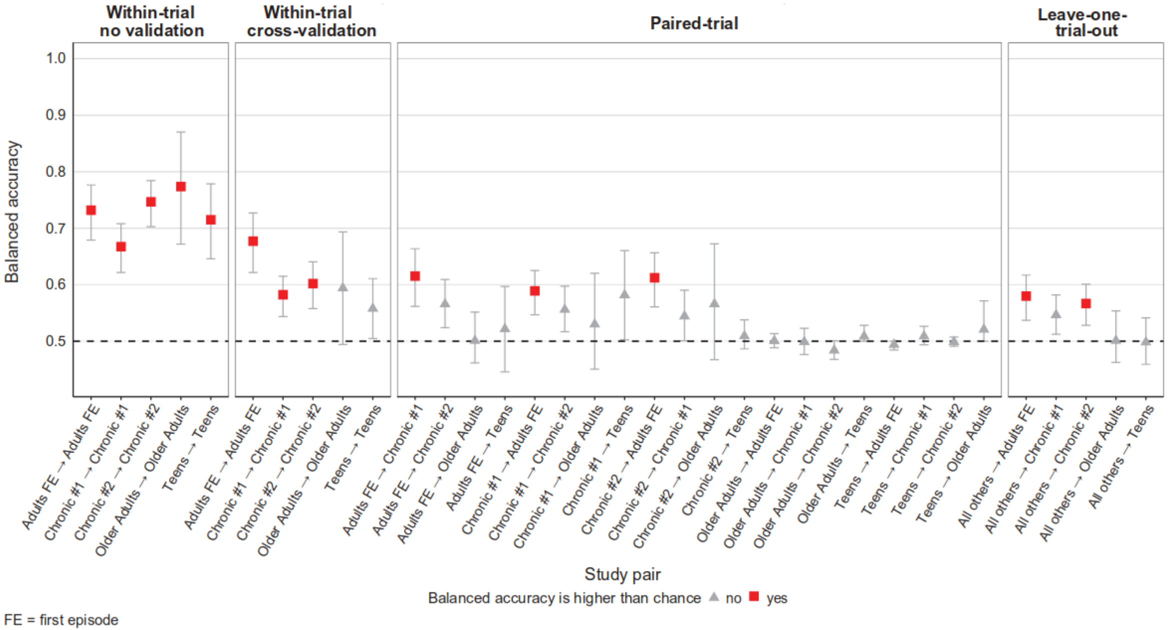
NB: This is already the high art of CV, assuming that everything went according to the textbook. Practice looks often different.

Data leakage

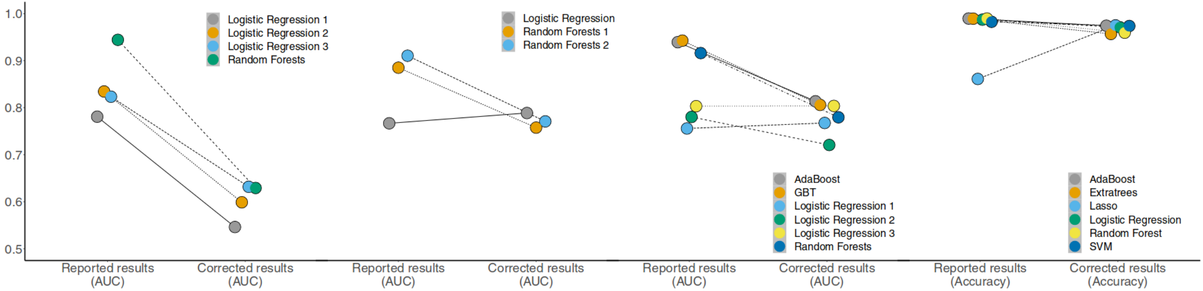
Data leakage is a **spurious relationship** between the independent variables and the target variable that arises as an artifact of the data collection, sampling, or pre-processing strategy.

(Kapoor and Narayanan)


Robberies in Alaska
correlates with
Professor salaries in the US
Robbery rate
128.7
114.9
101.1
87
74
Salary
$139.6K
$137.3K
$135.0K
$132.6K
$130.3K
2009
2011
2013
2015
2017
2019
2021
The robbery rate per 100,000 residents in Alaska · Source: FBI Criminal Justice Information Services
Average salary of full-time instructional faculty on 9-month contracts in degree-granting postsecondary institutions, by academic rank of Professor · Source: National Center for Education Statistics
2009-2021, r=0.922, r²=0.851, p<0.01 · tylervigen.com/spurious/correlation/2723



Within-trial no validation
Within-trial cross-validation
Paired-trial
Leave-one-trial-out
Balanced accuracy
1.0
0.9
0.8
0.7
0.6
0.5
Adults FE → Adults FE
Chronic #1 → Chronic #1
Chronic #2 → Chronic #2
Older Adults → Older Adults
Teens → Teens
Adults FE → Chronic #1
Adults FE → Chronic #2
Adults FE → Older Adults
Adults FE → Teens
Chronic #1 → Adults FE
Chronic #1 → Chronic #2
Chronic #1 → Older Adults
Chronic #1 → Teens
Chronic #2 → Adults FE
Chronic #2 → Chronic #1
Chronic #2 → Older Adults
Chronic #2 → Teens
Older Adults → Adults FE
Older Adults → Chronic #1
Older Adults → Chronic #2
Older Adults → Teens
Teens → Adults FE
Teens → Chronic #1
Teens → Chronic #2
Teens → Older Adults
All others → Adults FE
All others → Chronic #1
All others → Chronic #2
All others → Older Adults
All others → Teens
Study pair
Balanced accuracy is higher than chance
no
yes
FE = first episode



1.0
0.9
0.8
0.7
0.6
0.5
Logistic Regression 1
Logistic Regression 2
Logistic Regression 3
Random Forests
Logistic Regression
Random Forests 1
Random Forests 2
AdaBoost
GBT
Logistic Regression 1
Logistic Regression 2
Logistic Regression 3
Random Forests
AdaBoost
Extratrees
Lasso
Logistic Regression
Random Forest
SVM
Reported results (AUC)
Corrected results (AUC)
Reported results (AUC)
Corrected results (AUC)
Reported results (AUC)
Corrected results (AUC)
Reported results (Accuracy)
Corrected results (Accuracy)

A taxonomy of data leakage

L1 Lack of clean separation of training and test dataset

L2 Model uses features that are not legitimate.

L3 Test set is not drawn from the distribution of scientific interest.

 :: Nico Formánek ::

Preventing data leakage with model sheets

- A model sheet is a questionaire.
- Questions are designed to pin critical points of leakage.
- Has to be filled out after training.
- Should be supplemented with publication.

Stuttgart S-Bahn

Fill the model sheet for the Stuttgart S-Bahn model you trained on the first day. Start with Section 2. Work in Groups of 2.

- You might need to consult slides and code from day 1.
- If you think a question does not apply please say why.

Chekroud, A. M., Hawrilenko, M., Loho, H., Bondar, J., Gueorguieva, R., Hasan, A., Kambeitz, J., Corlett, P. R., Koutsouleris, N., Krumholz, H. M., Krystal, J. H., & Paulus, M. (2024). Illusory generalizability of clinical prediction models. *Science*, *383*(6679), 164–167. https://doi.org/10.1126/science.adg8538

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

Lones, M. A. (2021, August 5). *How to avoid machine learning pitfalls: A guide for academic researchers*. arXiv.org. Retrieved January 10, 2024, from https://arxiv.org/abs/2108.02497v4

Powell, M., Hosseini, M., Collins, J., Callahan-Flintoft, C., Jones, W., Bowman, H., & Wyble, B. (2020, February 14). *I Tried a Bunch of Things: The Dangers of Unexpected Overfitting in Classification*. https://doi.org/10.1101/078816

von Luxburg, U., & Schoelkopf, B. (2008, October 27). *Statistical Learning Theory: Models, Concepts, and Results*. arXiv: 0810.4752 `[math, stat]`. https://doi.org/10.48550/arXiv.0810.4752