OVERFITTING

DETECTION & TROUBLSHOOTING

OVERFIT VS. UNDERFIT

PROBLEM

UNDERFITTING

WHEN

Happens when a model is too simple to capture the underlying patterns in the data. This can occur if the model is not sufficiently complex or if the data has not been properly preprocessed.

INDICATORS

An underfitted model performs poorly on both training and unseen data because it fails to capture the inherent complexities.

IMPACTS

High bias, meaning the model makes consistent errors regardless of the input. While this results in a model that is less capable, it is often easier to diagnose and correct than overfitting.

OVERFIT VS. UNDERFIT

PROBLEM

OVERFITTING

WHEN

When a model learns the training data too well, capturing noise and outliers rather than just the underlying patterns. This usually happens when the model is too complex, such as having too many parameters relative to the amount of training data.

INDICATORS

Performs exceptionally well on training data but poorly on unseen data (testing or validation data). This means the model fails to generalize, leading to poor performance in real-world applications.

IMPACTS

High variance, meaning the model's predictions can vary significantly with small changes in the input data. This makes the model unreliable in practical scenarios.

WHICH IS LESS RISKY?

Underfitting is generally less risky than overfitting because it indicates that the model is not complex enough, which can often be addressed by increasing model complexity, adding more features, or tuning hyperparameters.

Overfitting is more problematic because it may give the illusion of high performance during training, leading to overconfidence in the model’s abilities. Detecting overfitting can be harder, and correcting it often requires more careful tuning, regularization, or simplifying the model.

TROUBLESHOOT

CROSS-VALIDATION

Cross-validation is a statistical method used in machine learning to evaluate the performance and generalizability of a model. It involves dividing the dataset into multiple subsets (or folds) and systematically training and testing the model on different subsets of the data. The primary goal of cross-validation is to ensure that the model performs well on unseen data and to minimize issues like overfitting and underfitting.

OVERFIT VS. UNDERFIT

SPLITTING

- The dataset is divided into k equally sized (or nearly equal) subsets, called "folds."
- Typically, k is chosen to be 5 or 10, but it can vary depending on the size of the dataset and the specific use case.

TRAINING

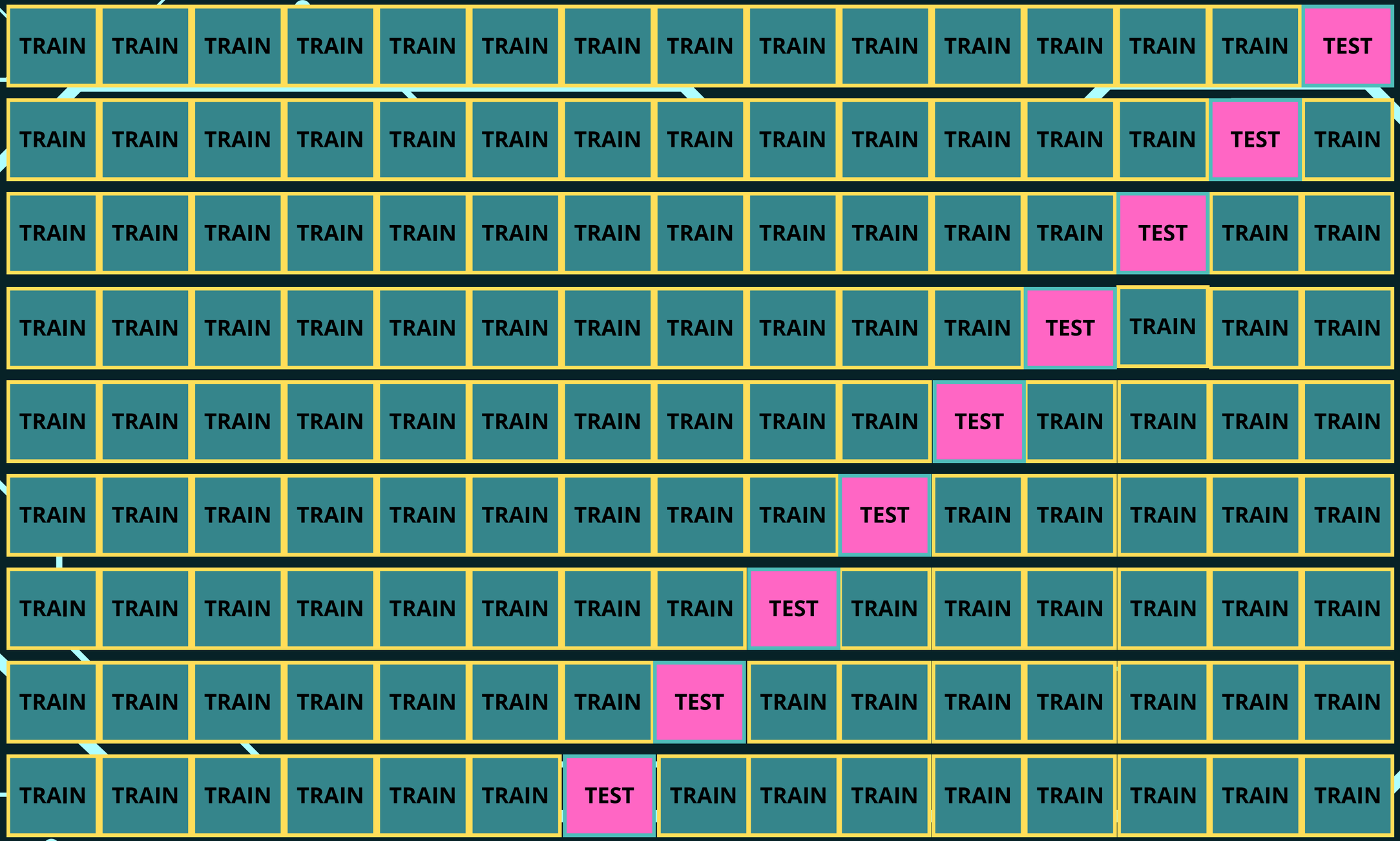
The model is trained k times, each time using k-1 folds for training and the remaining 1 fold for testing.

TESTING

For each iteration, a different fold is used as the test set, while the remaining folds form the training set.

RESULTS

- After completing all k iterations, the performance metrics (such as accuracy, precision, recall, etc.) are averaged over the k test sets.
- This average provides a more reliable estimate of the model's performance on unseen data


TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TEST
TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TEST TRAIN
TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TEST TRAIN TRAIN
TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TEST TRAIN TRAIN TRAIN
TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TEST TRAIN TRAIN TRAIN TRAIN
TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TEST TRAIN TRAIN TRAIN TRAIN TRAIN
TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TEST TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN
TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TEST TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN
TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TEST TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN TRAIN

ASSUMPTION: Training, Test and Validation sets are independent of each other.

Which of the two scenarios violate the assumption?

1. **Predicting the age of a person using facial features of people, in the data there are picture of identical twins, one was used in training, the other is used in testing.**
2. **Predicting criminal activities of streets using hoouse features, on the training set are houses at the right side of the road, on the testing set the houses on the left side of the road.**

GENERALIZATION BOUNDARIES

GENERALIZATION

The trained model should perform fairly well in new datasets.

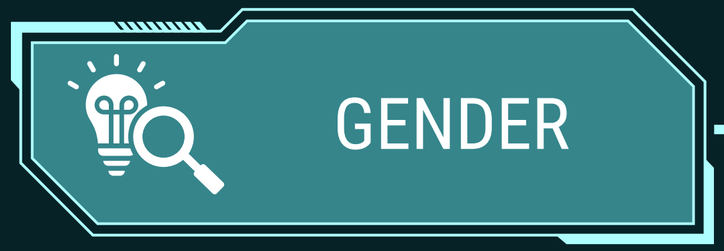
GENERALIZATION BOUNDARIES

- Affected by your sampling methods.
- Affected by the innate characteristics of your data.

Current Research

Aquaculture Biotechnology: Effects of Feed Crude Protein, Water Temperature, and Dissolved Oxygen on Fecundity of Female Freshwater Lobsters (Charex quadricarinatus).

GENERALIZATION BOUNDARIES



Works only on female freshwater lobsters.

Works only on adult stages.

Works only if the data is sampled in tropical climates.

Works only for Cherax quadricarinatus (not for Cherax cainii, Cherax destructor, Procambarus clarkii).

KEY TAKE-AWAYS

- Model is as good as the data it is trained on.
- Choose what kind of features you want to train your models with.
- Be transparent with the features you trained into your model.
- Your model will work good based on the generalization boundaries, but its performance will decrease if your data starts to drift from what it is trained on.

THANK YOU!

A presentation is a method of sharing information, ideas, or arguments with an audience using spoken words and visual aids. It typically involves a speaker conveying content to inform, persuade, or entertain the listeners.

www.linkedin.com/in/gerard-ompad

gerard.ompad@gmail.com

https://github.com/gerard-ompad