Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

Model Selection

Which Algorithm to Use?

- K-Nearest Neighbor
- Logistic Regression
- Decision tree
- Neural networks
- SVM
- Fisher Linear Discriminant
- Naïve bayes
- …

Which Algorithm is the Best?

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering

Table 3. Normalized scores of each learning algorithm by problem (averaged over eight metrics)

MODEL	CAL	COVT	ADULT	LTR.P1	LTR.P2	MEDIS	SLAC	HS	MG	CALHOUS	COD	BACT	MEAN
BST-DT	PLT	**.938**	.857	**.959**	**.976**	.700	.869	**.933**	.855	**.974**	**.915**	.878*	**.896***
RF	PLT	.876	.930	.897	.941	**.810**	.907*	.884	.883	.937	.903*	.847	.892
BAG-DT	–	.878	.944*	.883	.911	.762	.898*	.856	**.898**	.948	.856	**.926**	.887*
BST-DT	ISO	.922*	.865	.901*	.969	.692*	.878	.927	.845	.965	.912*	.861	.885*
RF	–	.876	.946*	.883	.922	.785	.912*	.871	.891*	.941	.874	.824	.884
BAG-DT	PLT	.873	.931	.877	.920	.752	.885	.863	.884	.944	.865	.912*	.882
RF	ISO	.865	.934	.851	.935	.767*	**.920**	.877	.876	.933	.897*	.821	.880
BAG-DT	ISO	.867	.933	.840	.915	.749	.897	.856	.884	.940	.859	.907*	.877
SVM	PLT	.765	.886	.936	.962	.733	.866	.913*	.816	.897	.900*	.807	.862
ANN	–	.764	.884	.913	.901	.791*	.881	.932*	.859	.923	.667	.882	.854
SVM	ISO	.758	.882	.899	.954	.693*	.878	.907	.827	.897	.900*	.778	.852
ANN	PLT	.766	.872	.898	.894	.775	.871	.929*	.846	.919	.665	.871	.846
ANN	ISO	.767	.882	.821	.891	.785*	.895	.926*	.841	.915	.672	.862	.842
BST-DT	–	.874	.842	.875	.913	.523	.807	.860	.785	.933	.835	.858	.828
KNN	PLT	.819	.785	.920	.937	.626	.777	.803	.844	.827	.774	.855	.815
KNN	–	.807	.780	.912	.936	.598	.800	.801	.853	.827	.748	.852	.810
KNN	ISO	.814	.784	.879	.935	.633	.791	.794	.832	.824	.777	.833	.809
BST-STMP	PLT	.644	**.949**	.767	.688	.723	.806	.800	.862	.923	.622	.915*	.791
SVM	–	.696	.819	.731	.860	.600	.859	.788	.776	.833	.864	.763	.781
BST-STMP	ISO	.639	.941	.700	.681	.711	.807	.793	.862	.912	.632	.902*	.780
BST-STMP	–	.605	.865	.540	.615	.624	.779	.683	.799	.817	.581	.906*	.710
DT	ISO	.671	.869	.729	.760	.424	.777	.622	.815	.832	.415	.884	.709
DT	–	.652	.872	.723	.763	.449	.769	.609	.829	.831	.389	.899*	.708
DT	PLT	.661	.863	.734	.756	.416	.779	.607	.822	.826	.407	.890*	.706
LR	–	.625	.886	.195	.448	.777*	.852	.675	.849	.838	.647	.905*	.700
LR	ISO	.616	.881	.229	.440	.763*	.834	.659	.827	.833	.636	.889*	.692
LR	PLT	.610	.870	.185	.446	.738	.835	.667	.823	.832	.633	.895	.685
NB	ISO	.574	.904	.674	.557	.709	.724	.205	.687	.758	.633	.770	.654
NB	PLT	.572	.892	.648	.561	.694	.732	.213	.690	.755	.632	.756	.650
NB	–	.552	.843	.534	.556	.011	.714	-.654	.655	.759	.636	.688	.481

Comparison of supervised algorithms (decision trees, random forest, SVM, neural networks etc.) over 11 classification tasks

[Caruana & Niculesu-Mizil, ICML 2006]

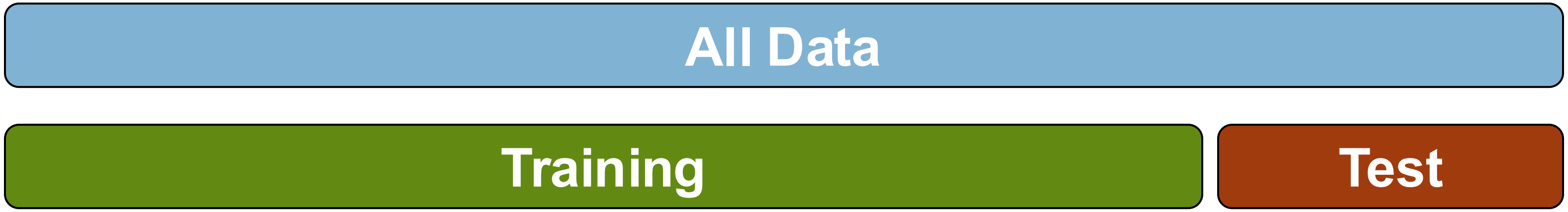
Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

The bias-variance tradeoff

Generalisation error

General for supervised machine learning: Split the available data into a training and independent test set



Estimate the generalisation error based on the independent test set

The bias-variance tradeoff

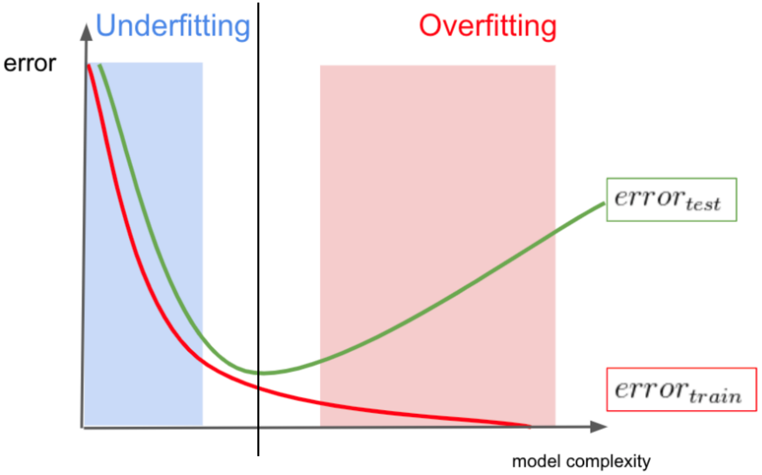
Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering

Typically, we would like to choose our model complexity to trade bias off with variance in such a way as to minimize the test error

High bias
Low variance

optimal model

Low bias
High variance



The **model bias** refers to the limitation due to the model flexibility. More flexible models result in less bias

The **variance of the model** refers to the amount by which $\hat{f}$ would change, if a different training dataset was used.
More flexible models have higher variance, i.e. small changes in the training data can result in large changes in $\hat{f}$.

Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

Hyperparameters

Model Selection - Hyperparameters

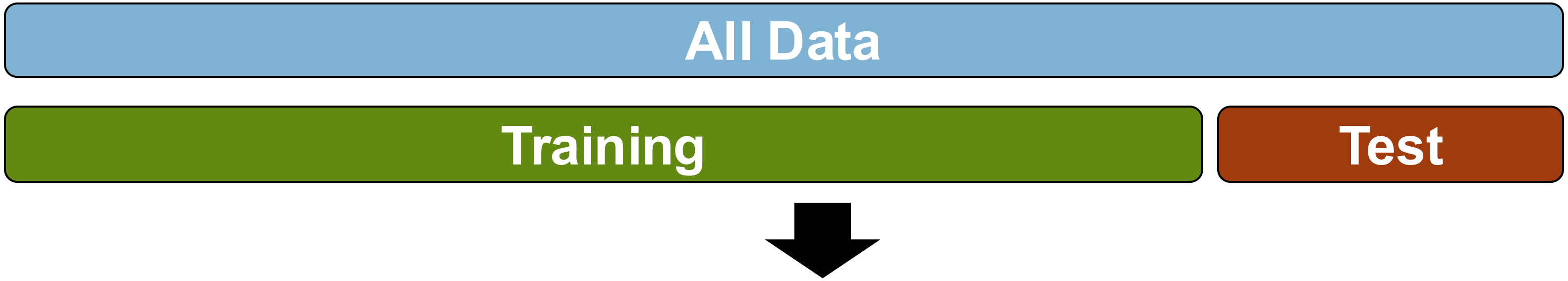
The goal of Model Selection is to find good hyperparameters for a given algorithm. Hyperparameters are not optimised during training (they do not belong to the «trainable parameters»). How to choose the best values?

Algorithm	Hyperparameter	Parameters
k-Means	- Number of clusters k	- Coordinates of the k centroids
Polynomial Regression	- Polynomials of the input variables - Regularization parameter λ	- Values of $\theta_0, \theta_1, \dots \theta_N$
Neural Networks	- Number of hidden layers - Size of hidden layers - Links between neurons - Activation function	- Weights and biases

School of Engineering

Data partitioning

Data Partitioning



Independent test set. Do not use during model building and selection

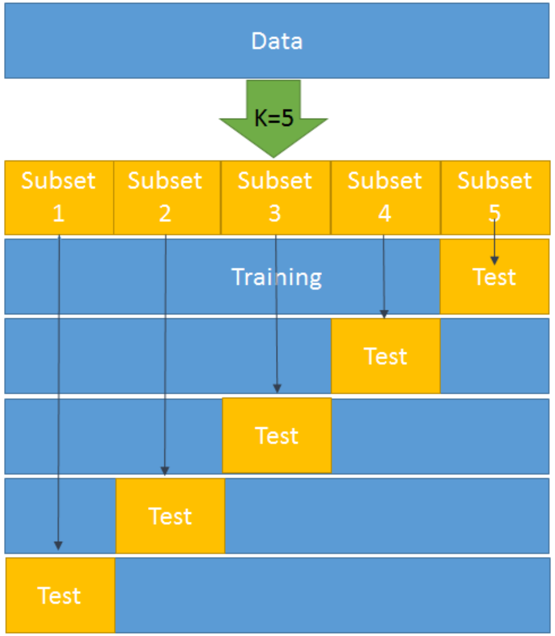
Default approaches split into training, test, and validation set (if possible)

Used for model selection and finding the optimal hyperparameters

By partitioning the available data into three sets, we drastically reduce the number of samples which can be used for learning the model, and the results can depend on a particular random choice for the pair of (train, validation) sets. A solution to this problem is **cross-validation**.

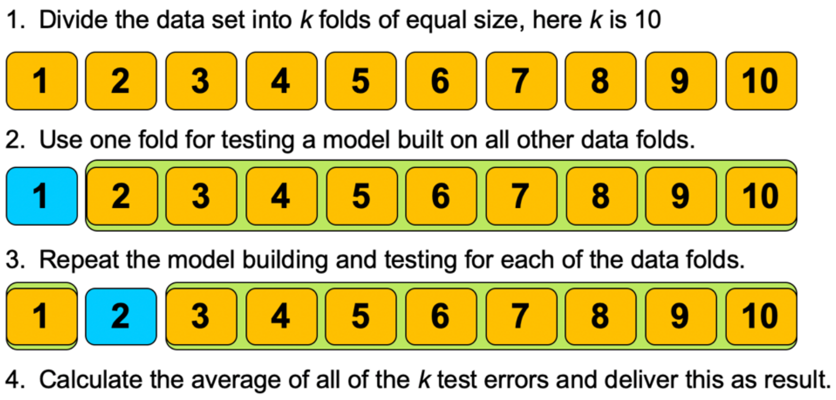
k-Fold Cross Validation

Typical choice is k = 5 or k = 10



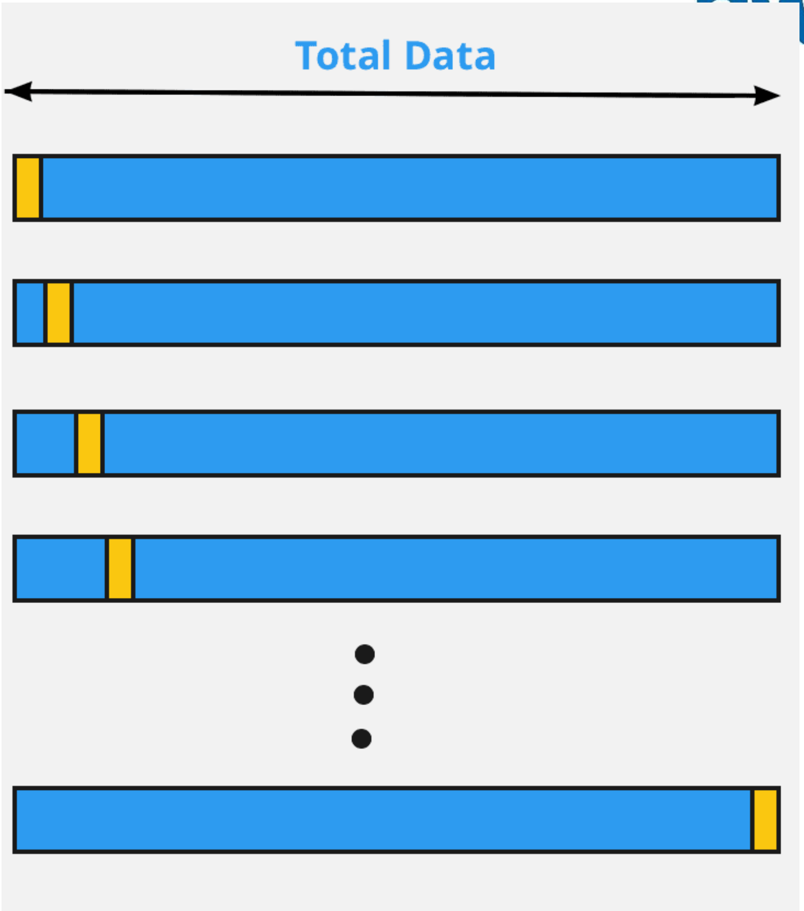
k-fold Cross Validation (CV)

- Cross calidation is for model checking and finding good hyperparameter, but NOT for model building
- Cross calidation needs to run several times to give reliable results (e.g. 100 times for 5-fold CV)
- After cross calidaton the entire training data can be used to build the model



Leave-one-out Cross Validation LOOCV

- De-facto n-fold cross validation, with n number of training samples
- Train the model on *n-1* samples and evaluate it on the single data point
- Computationally expensive
- Useful especially for small datasets



Cross validation score

The cross validation score

The Cross validation score is the average of a given accuracy metric over all these runs for all data folds.

The actual accuracy metric used to compute the score depends on the task
- regression: MSE, MAE, RMSD, etc.
- classification: Precision, Recall, F1-score, Accuracy

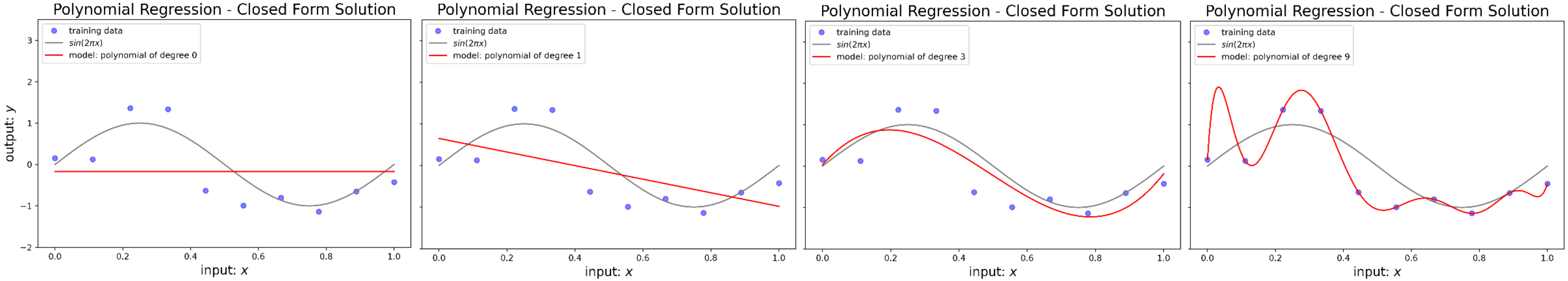
Attention – do not get confused!
- if the metric represents an error: smaller is better
- If the metric represents a type of accuracy: larger is better

→ always check what is ACTUALLY plottet!

Model Selection and generalisation error in polynomial regression

Which degree of the polynomial is appropriate?

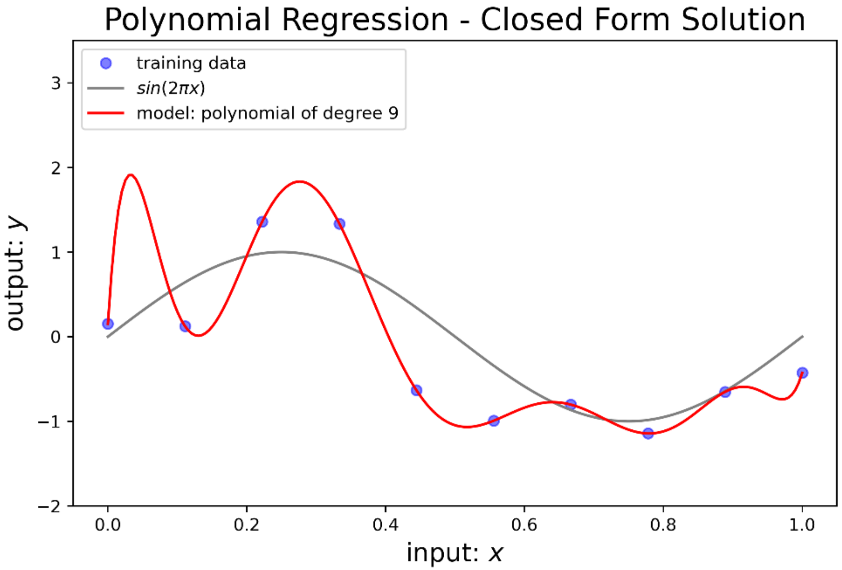
10 data points sampled from $y(x) = \sin 2\pi x$ with small additive Gaussian noise



The agreement with the training data improves with increasing model flexibility (degree of the polynomial).

Generalisability: An Important Aspect in Model Selection

Illustration for **Overfitting**: The model with 10 degrees of feedom manages to fit the 10 data points exactly. However, visual inspection suggests that its **generalisability** is poor, i.e. large deviations from the underlying structure in observations different from the training set.

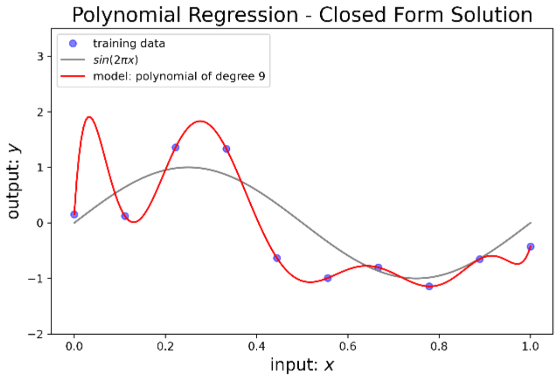


Overfitting vs. underfitting

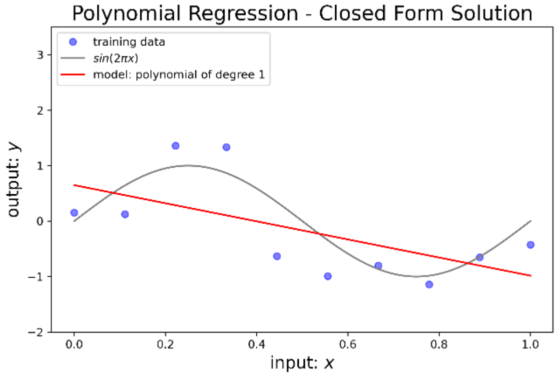
Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering

An **overfitted** model corresponds too closely or exactly to the training data. It may include some of the residual variation of the data, i.e. the noise, in the model structure and therefore fail to predict observations from data not included in the training set reliably, i.e. **generalise poorly**.

In other words: The model contains more paramaters than can be justified by the structure underlying the training data.

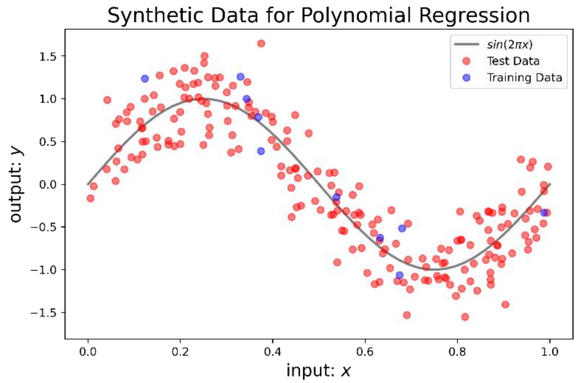


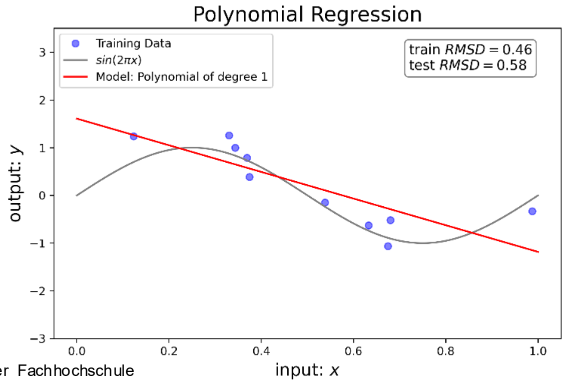
Underfitting occurs when the model is not flexible enough to capture the underlying structure of the training data. This results in both, **poor agreement with the training data and generalisability**.

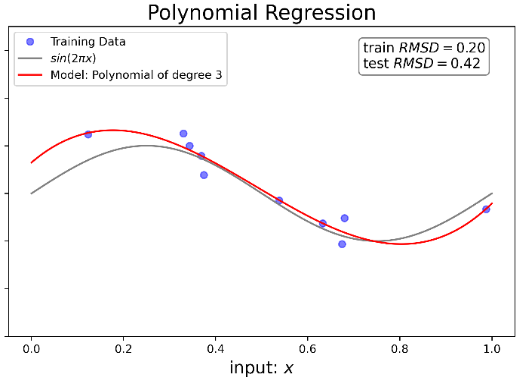


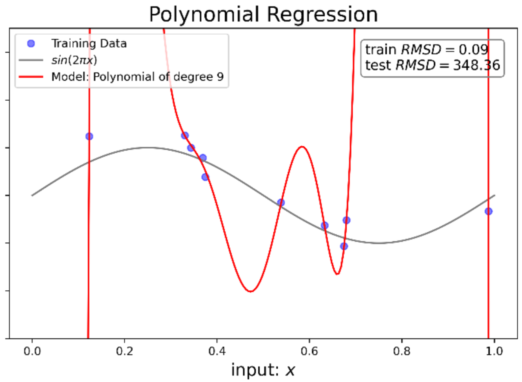
Measuring the generalisation error

If the underlying structure of the data is unknown, an **independent test** dataset is used to **estimate the generalisation error** by means of the evaluation **metrics**, e.g. $RMSD$.

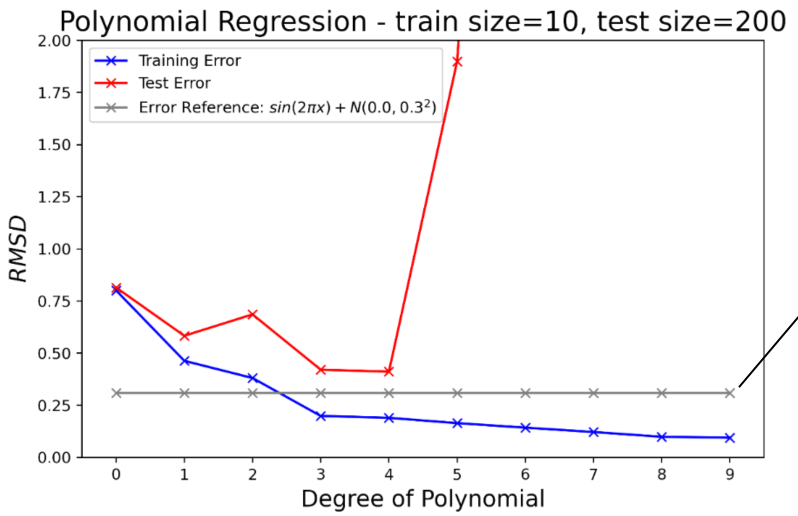








Generalisation error as a function of model flexibility



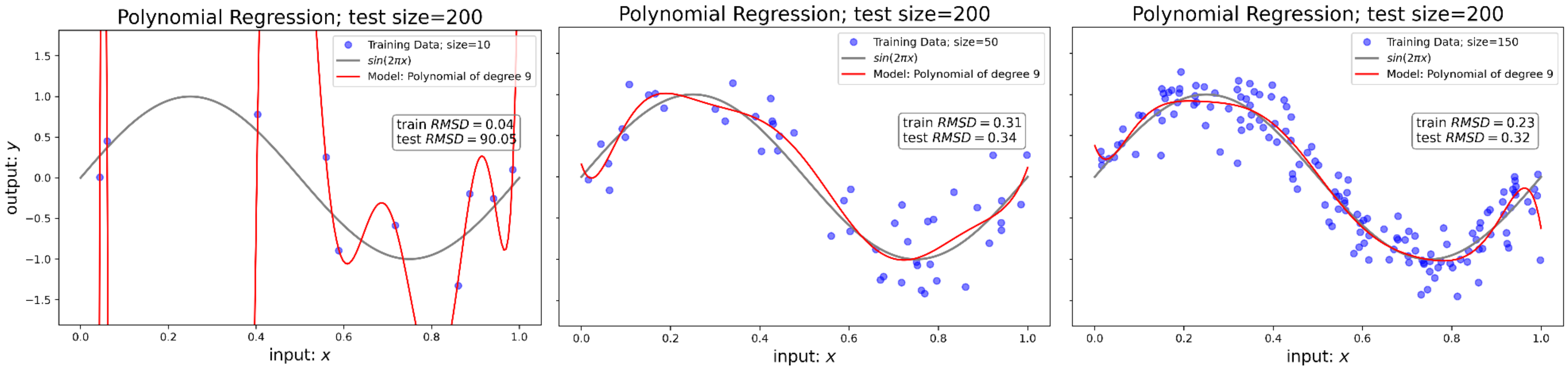
Reference error due to the additive noise in this synthetic dataset

- Monotonously decreasing training error
- and U-shaped test error

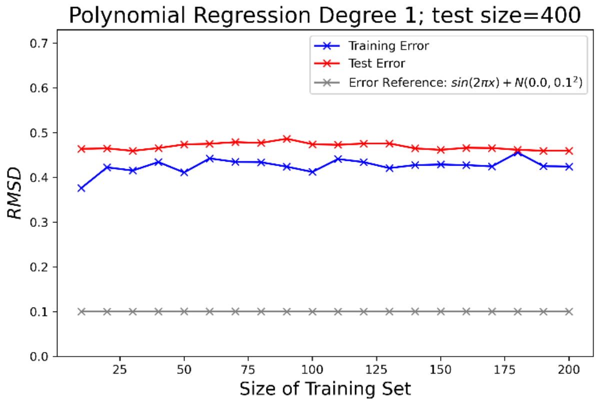
with increasing flexibility of the model (and constant trainingset size)

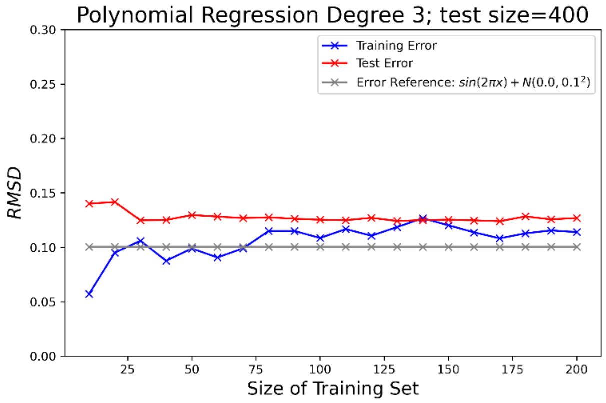
Generalisation Error as a Function of the Amount of Training Data

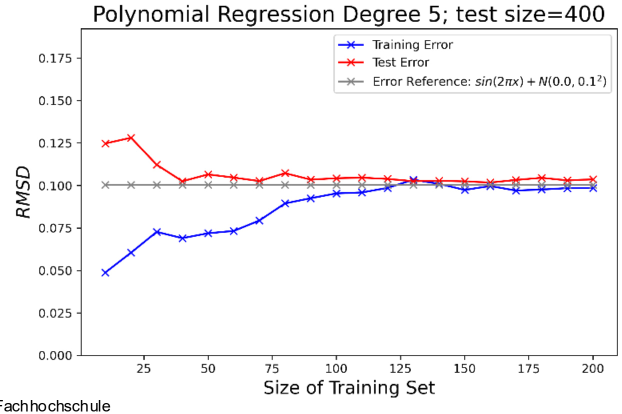
Typically, the overfitting problem decreases with increasing amount of training data and constant model flexibility.

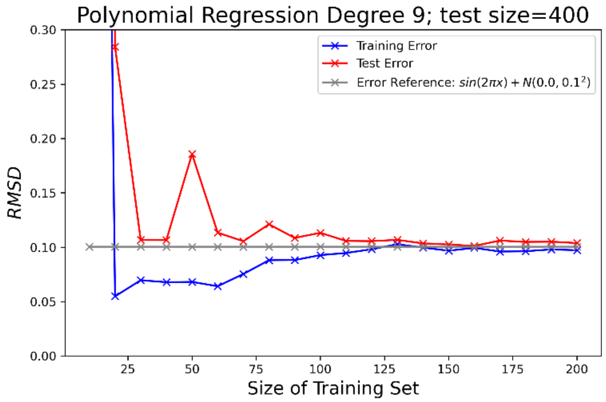


More flexible models need more training data



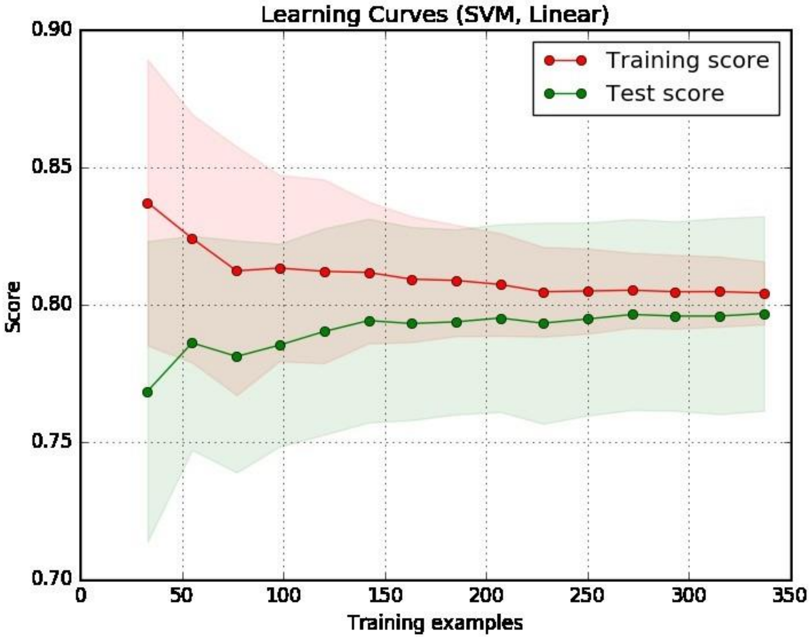






Learning curve analysis in classification

Evaluation of Learning Curves



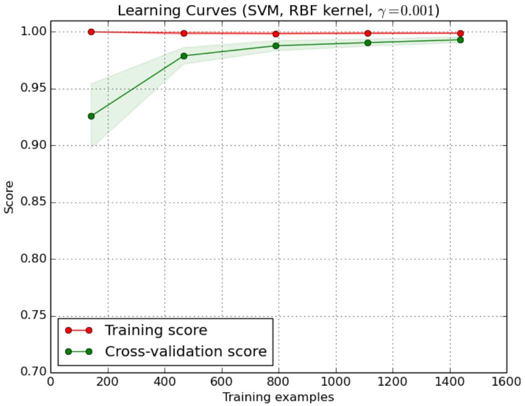
Evaluate tendencies of training and test curves for sufficient examples (right side of plot):

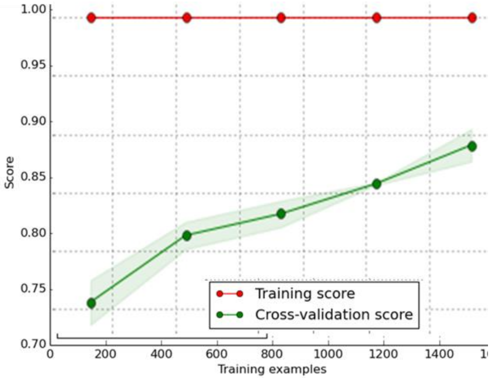
General Rules:

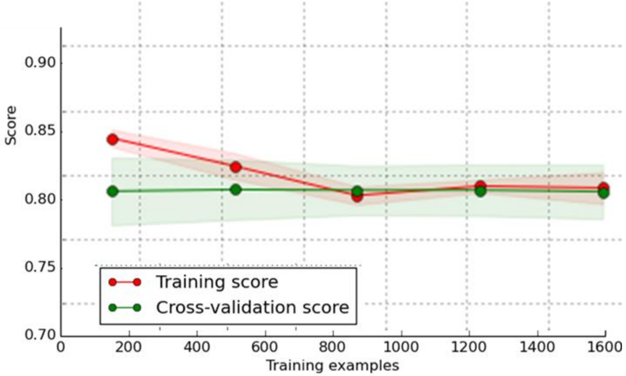
- If both curves are "close to each other" and both of them have a low score -> potential *underfitting* (High Bias)
- If training curve has a much better score than testing curve -> potential *overfitting* (High Variance)

https://stats.stackexchange.com/questions/220827/how-to-know-if-a-learning-curve-from-svm-model-suffers-from-bias-or-variance/220852#220852

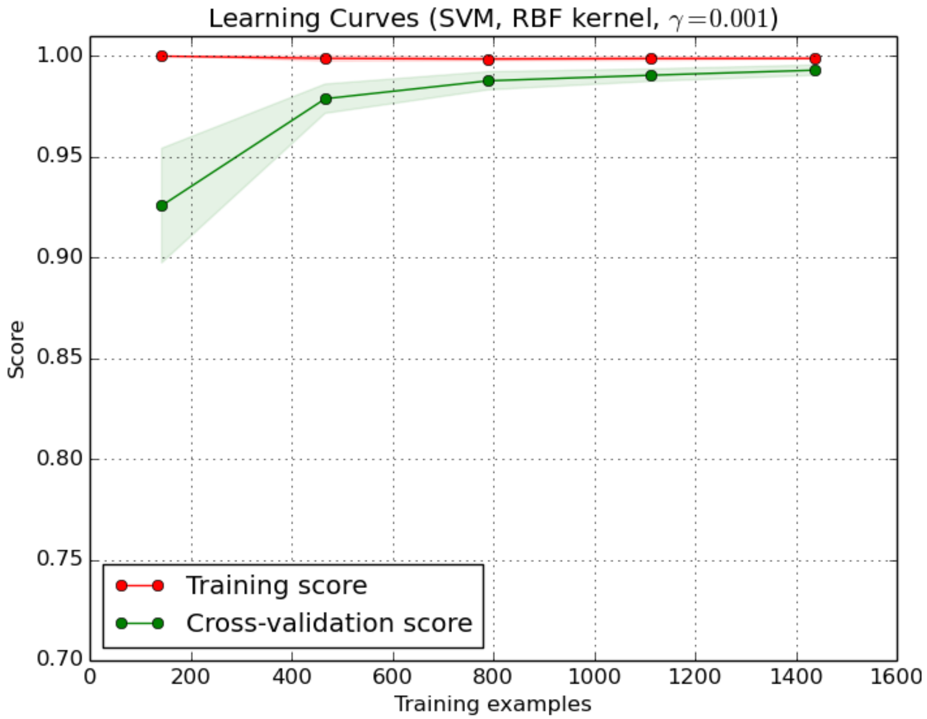
Exercise: Analyze the following Learning Curve





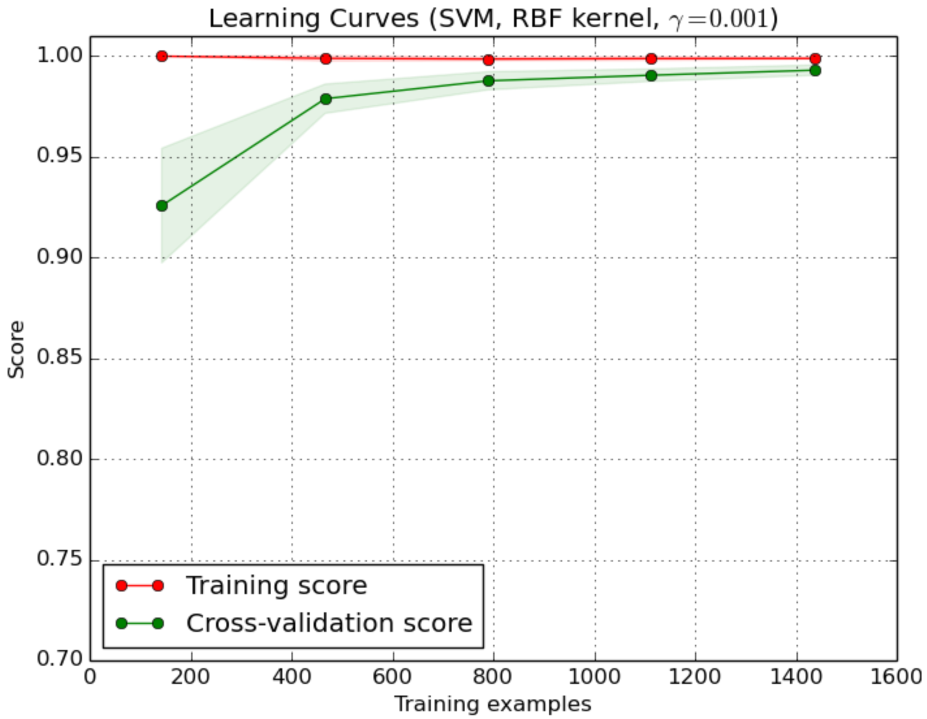


Exercise: Analyze the following Learning Curve



http://scikit-learn.org/0.15/auto_examples/plot_learning_curve.html

SOLUTION: Analyze the following Learning Curve



Observations:

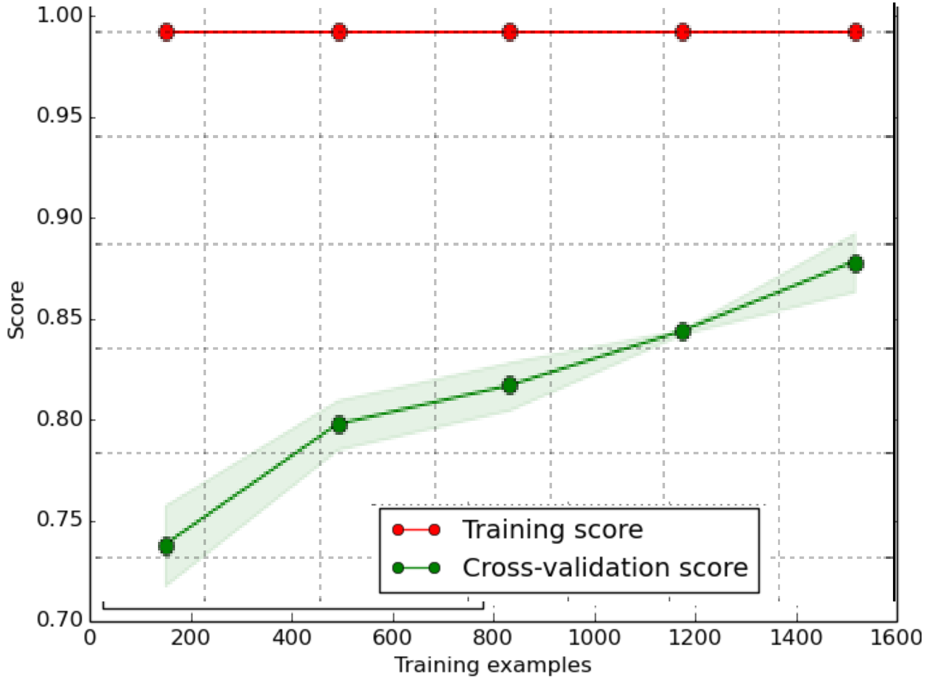
- High training score, stable
- Validation score still rising

Interpretation:

- More training examples will probably improve the classifier

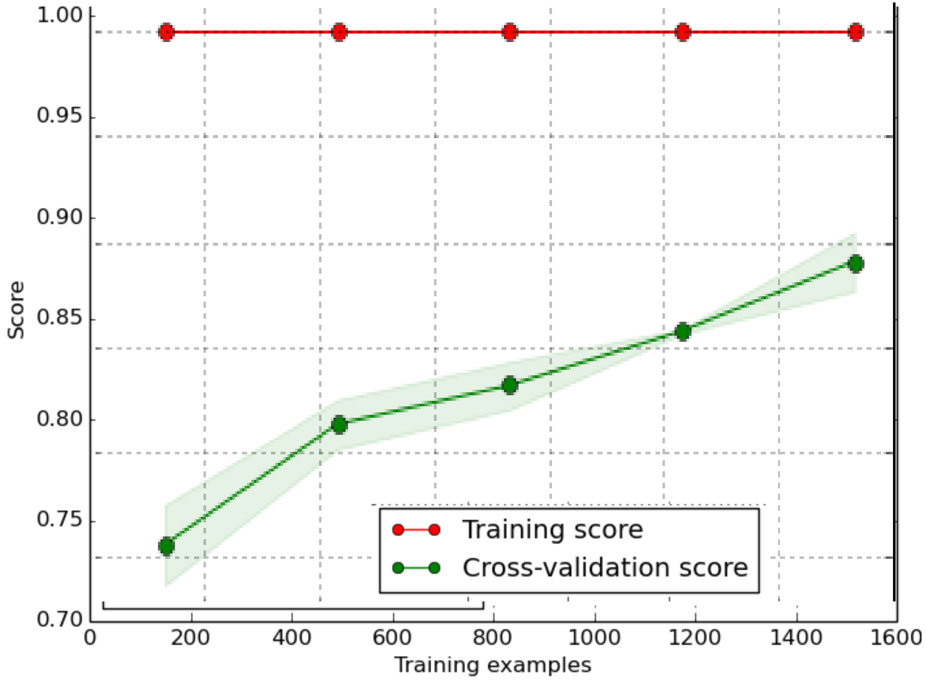
http://scikit-learn.org/0.15/auto_examples/plot_learning_curve.html

Exercise: Analyze the following Learning Curve



http://www.ritchieng.com/machinelearning-learning-

SOLUTION: Analyze the following Learning Curve



Observations:

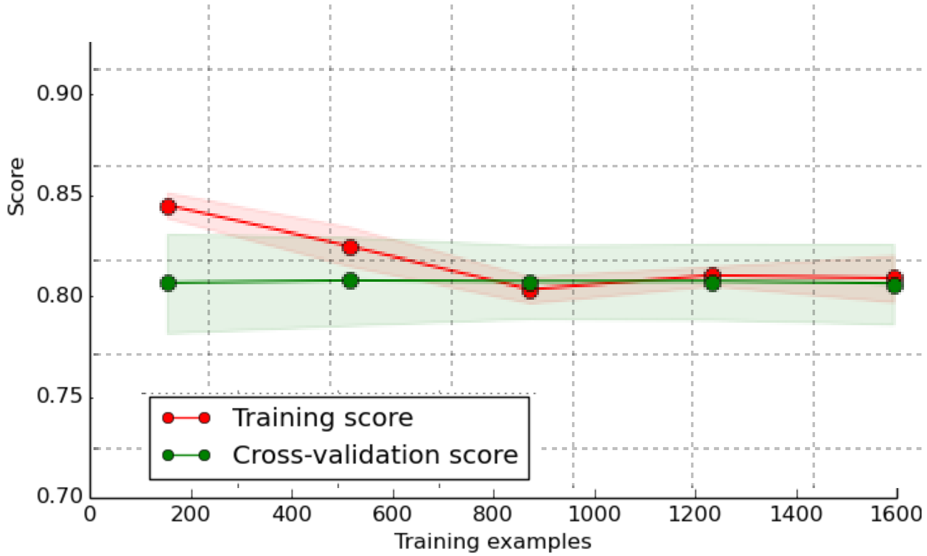
- Training score is at its maximum regardless of training examples
- Cross-validation score increases over time
- Huge gap between cross-validation score and training score

Interpretation:

- High training score shows severe overfitting
- High variance scenario
- Reduce complexity of the model or gather more data

http://www.ritchieng.com/machinelearning-learning-

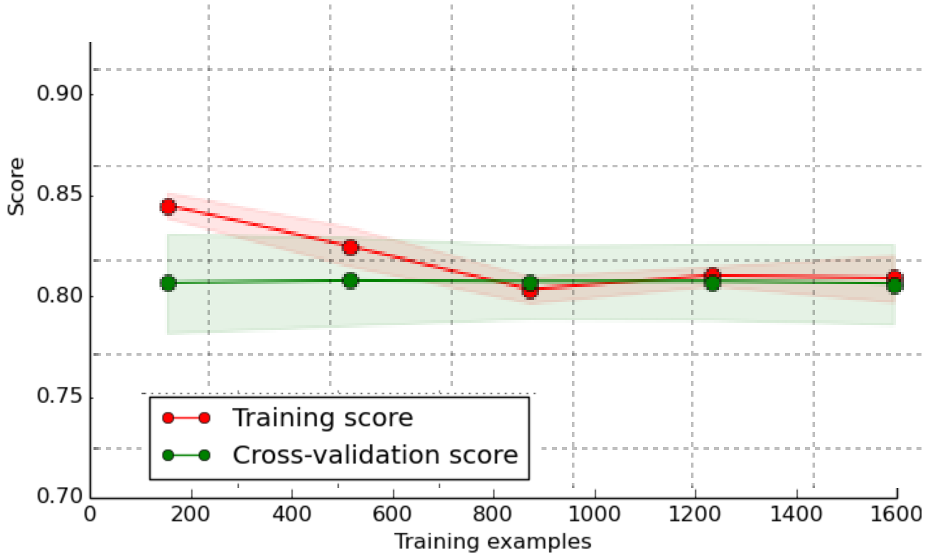
Exercise: Analyze the following Learning Curve



http://www.ritchieng.com/machinelearning-learning-curve/

SOLUTION: Analyze the following Learning Curve

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering



Observations:

- Training score decreases and plateau: indicates underfitting, High bias
- Cross-validation score stagnating throughout
- Low scores

Interpretation:

– Unable to learn from data

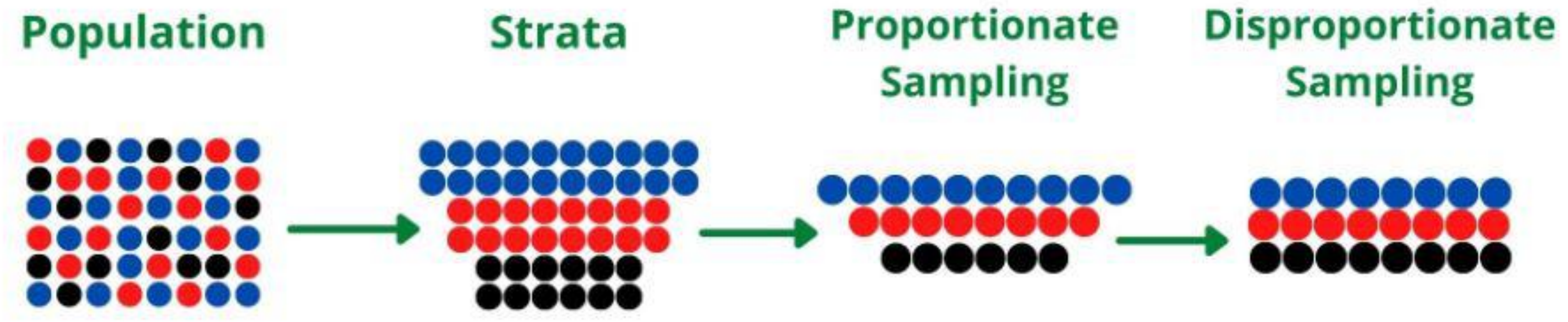
– Should tweak model (perhaps increase model complexity)

http://www.ritchieng.com/machinelearning-learning-curve/

Appendix

Stratified Sampling

Zürcher Hochschule
für Angewandte Wissenschaften
zh
aw
School of
Engineering

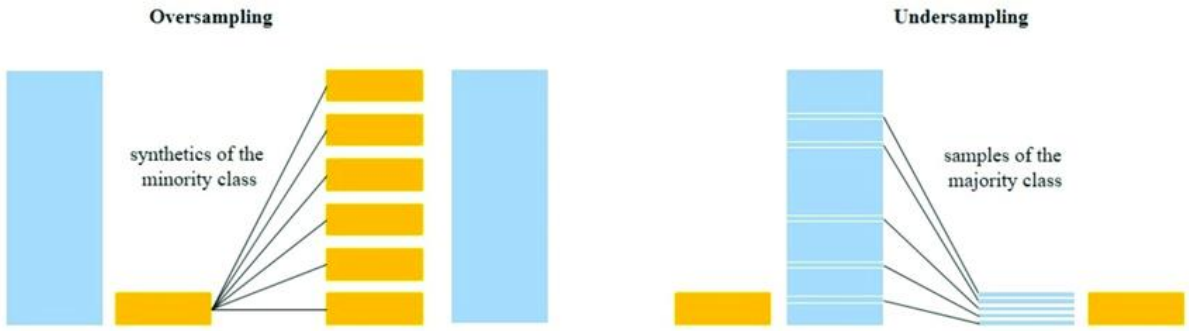


https://media.geeksforgeeks.org/wp-content/uploads/20210905231338/Stratified.JPG

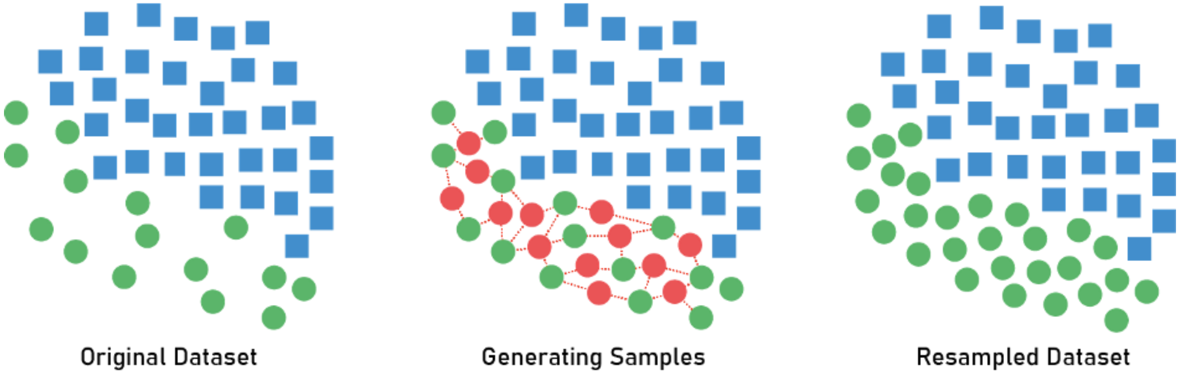
Oversampling and Undersampling

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering

- For highly imbalanced datasets
- **Oversampling:** repeat samples from underrepresented class, or create new synthetic samples (e.g. with SMOTE)
- **Untersampling:** only use a small fraction of samples from the overrepresented class, such that all classes have same amount of samples



SMOTE -



https://emilia-orellana44.medium.com/smote-2acd5dd09948