Learning theory and model validation

Lluís Talavera, 2022

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

ETSEIB

Generalization performance

Given a dataset of examples divided into classes (or with a numeric target), learn how to assign a class label (or predict the numeric value of the target) for **new, unseen examples**

We call generalization performance* to the performance of the model on new data, as opposed to the performance measured on the specific data used in training the model.

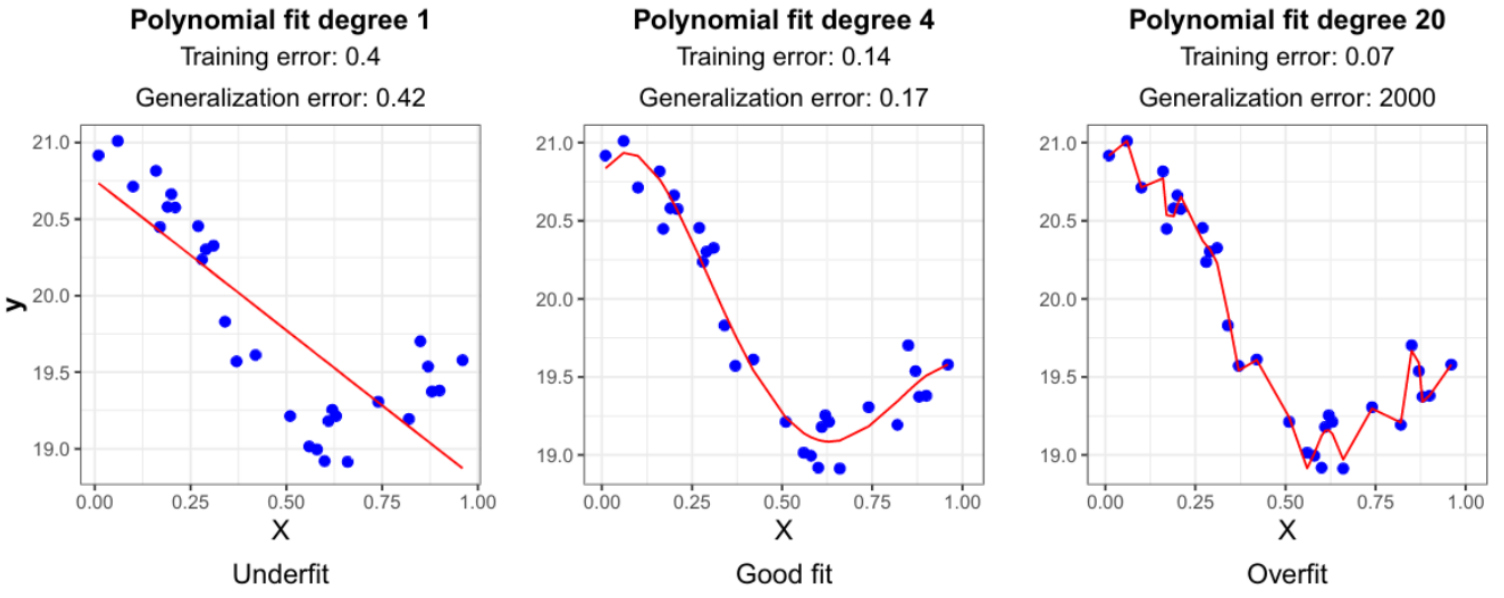
Performance on training data may be a poor estimator of generalization performance: too optimistic.

Evaluation is therefore performed splitting the data into:

- Training set: used to build the model.
- Test set: new data not used in training, only for estimating performance.

* or generalization error if we use error as a validation measure.

Underfitting and overfitting



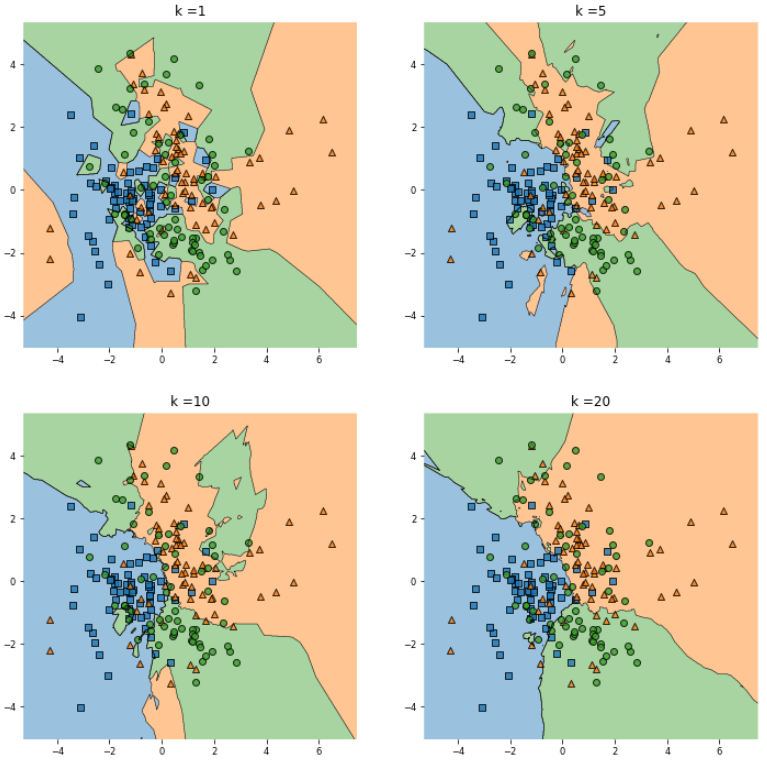
source

Underfitting: Both, training and test error are high.

Overfitting: Training error is lower than test error.

Simple and complex decision boundaries

In classification, we often can obtain models of different complexity by setting the parameters/hyperparameters of the learning algorithm.

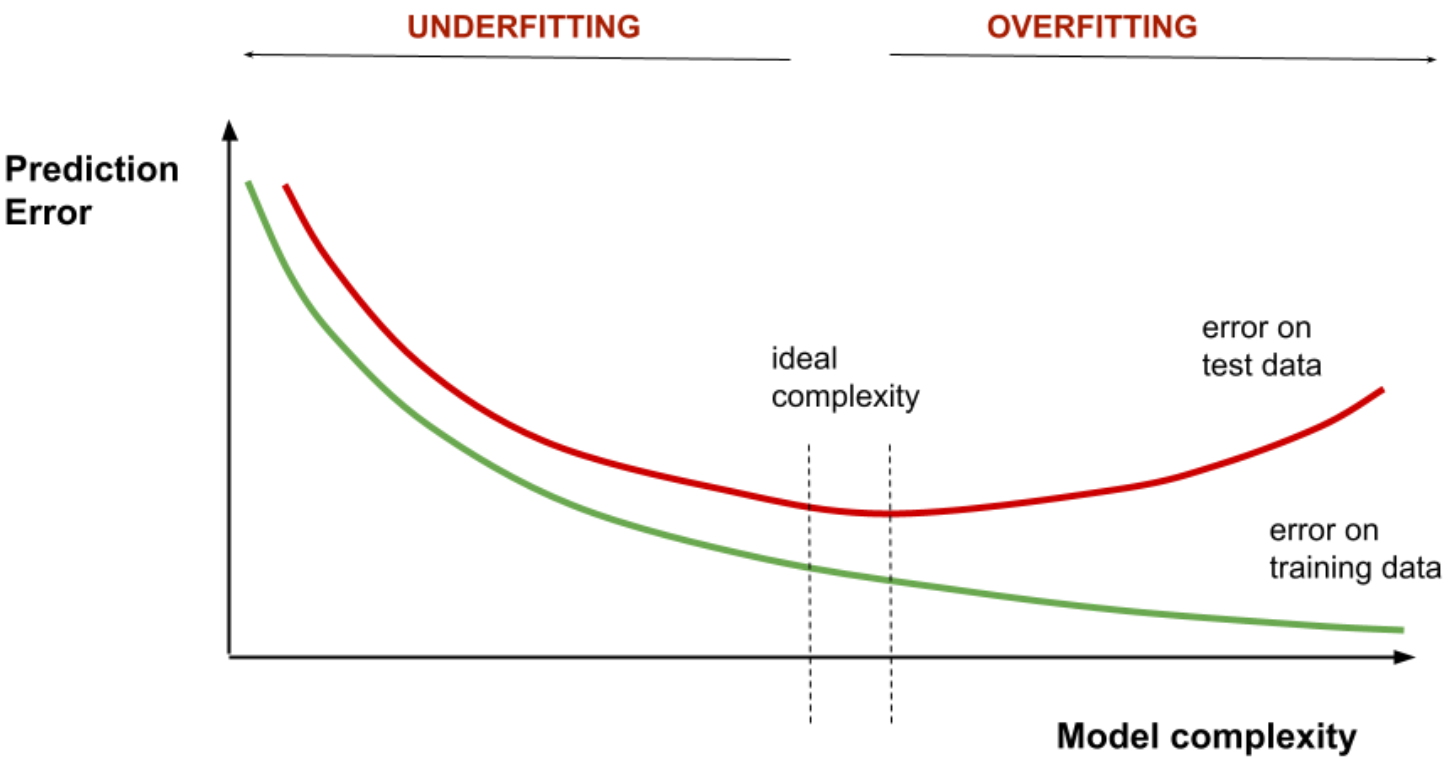


Example: k-NN

Lower values of k: more complex models, risk of overfitting

Big values of k: simpler models, risk of underfitting

Summary


UNDERFITTING
OVERFITTING
Prediction Error
ideal complexity
error on test data
error on training data
Model complexity

Bias and variance

Bias are the simplifying assumptions made by the algorithm to make the model easier to learn.

> High bias: Linear Regression, Logistic Regression
> Low bias: k-NN

Variance is the sensivity of the model to different training sets, how much the predictions change with small modifications of the data.

> High variance: k-NN
> Low variance: Linear Regression, Logistic Regression

Three types of error:

- Irreducible error (cannot be reduced regardless of the algorithm)
- Bias error
- Variance error

Bias-variance tradeoff

Reducing bias:

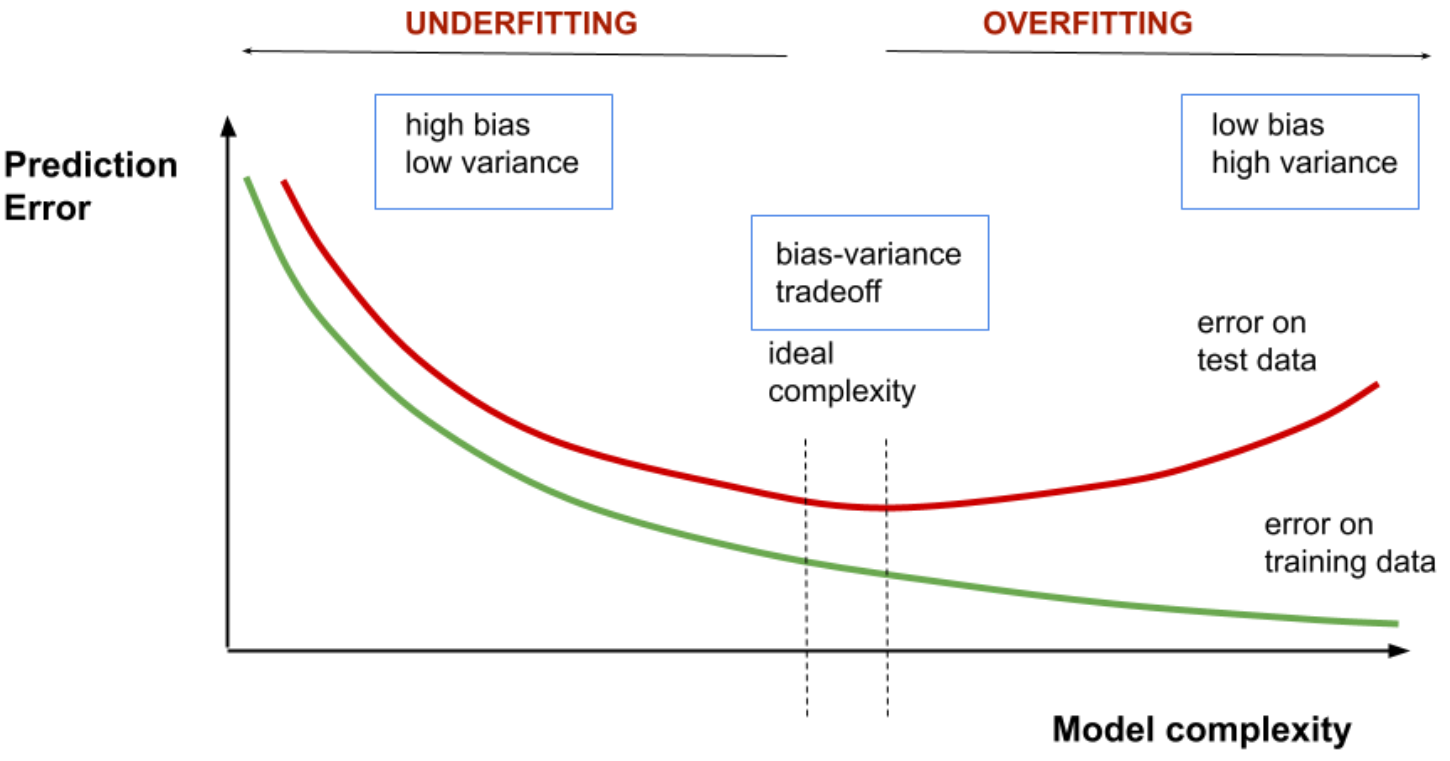
- Choose a different algorithm
- Add new features
- Tune hyperparameters to increase complexity of the model

Reducing variance:

- Train with more data
- Reduce the number of features
- Tune hyperparameters to decrease complexity of the model
- Ensemble learning

bias-variance tradeoff: Increasing the bias will decrease the variance. Increasing the variance will decrease the bias.

Summary


UNDERFITTING
OVERFITTING
high bias
low variance
low bias
high variance
Prediction
Error
bias-variance
tradeoff
ideal
complexity
error on
test data
error on
training data
Model complexity

Model validation

We want:

- An estimate of the performance of the model on unseen data.

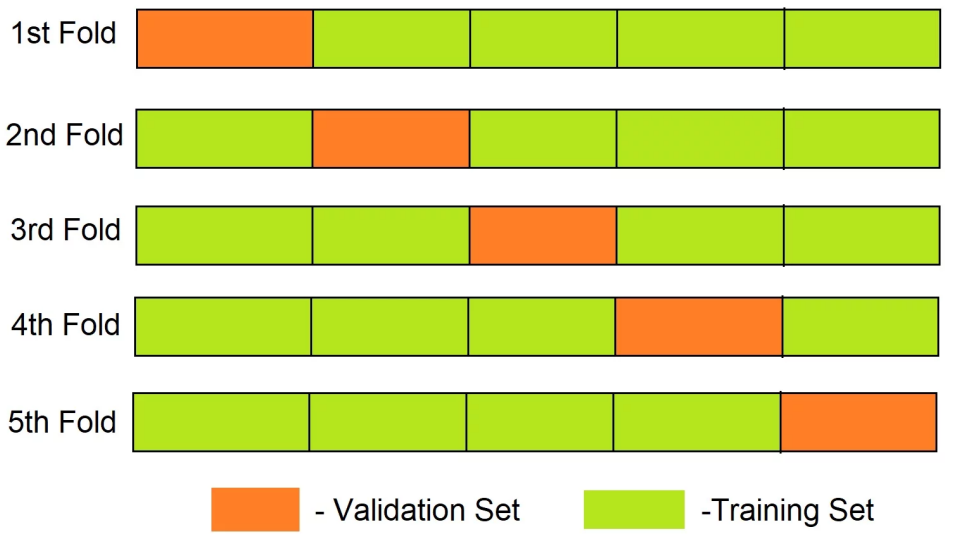
We need:

- A validation strategy: How are the data divided into the training and test subsets?
    - Hold-out (what we have been using)
    - Repeated hold-out (average over several hold-out runs)
    - k-fold Cross Validation
- A validation metric: How do we measure model performance?
    - Accuracy/error
    - Confusion matrix
    - Precision, Recall, F1, precision-recall curves
    - Specificity, Sensitivity, ROC curves

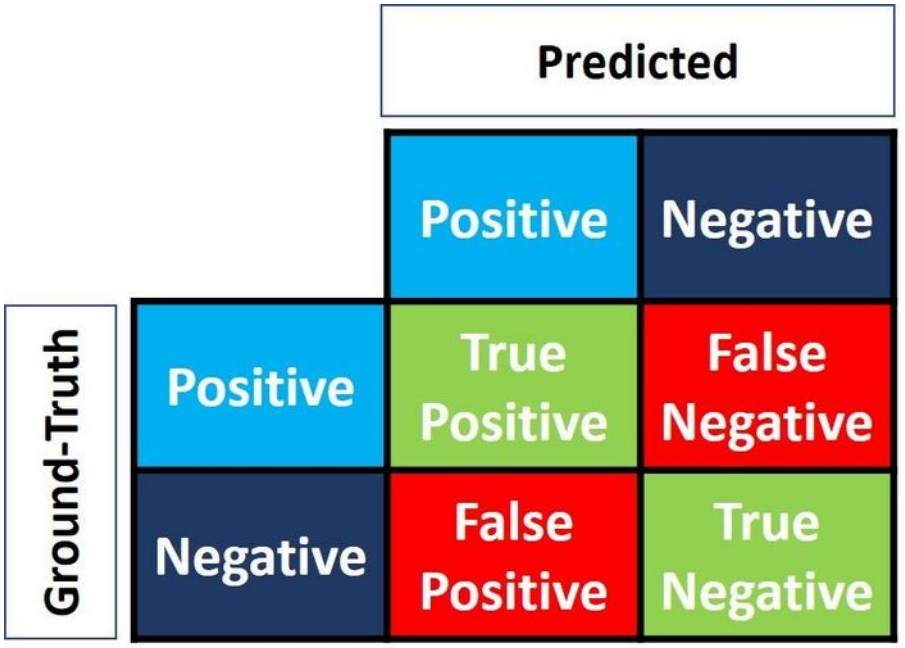
k-fold Cross Validation

Data is split into k subsets (folds) and, at each step, $k-1$ subsets are used to train the model and the remaining one is used for testing. The performance of the model is the average of the scores.

Preprocessing steps that estimate some parameter from data (scaling, replacing missing values, etc.) must be repeated for each different training subset.



Confusion matrix



Precision, Recall and F1

$Precision = \Large\frac{TP}{TP + FP}$

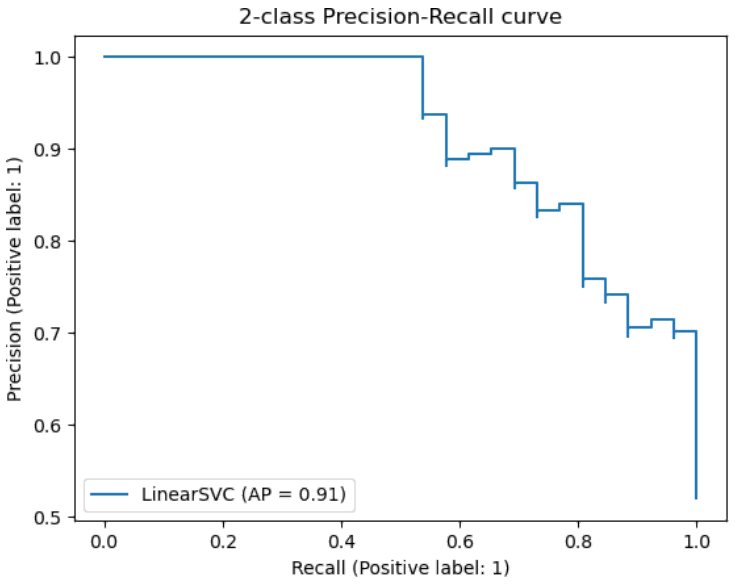
ratio of correctly predicted positive to all items predicted to be positive

$Recall = \Large\frac{TP}{FN + TP}$

ratio of correctly predicted positive to all items that are actually positive

$F1 = 2 \cdot \Large\frac{Precision \cdot Recall}{Precision + Recall}$

harmonic mean of precision and recall



The precision-recall curve shows the relationship between precision and recall for different prediction thresholds (for classifiers that produce probabilistic scores)

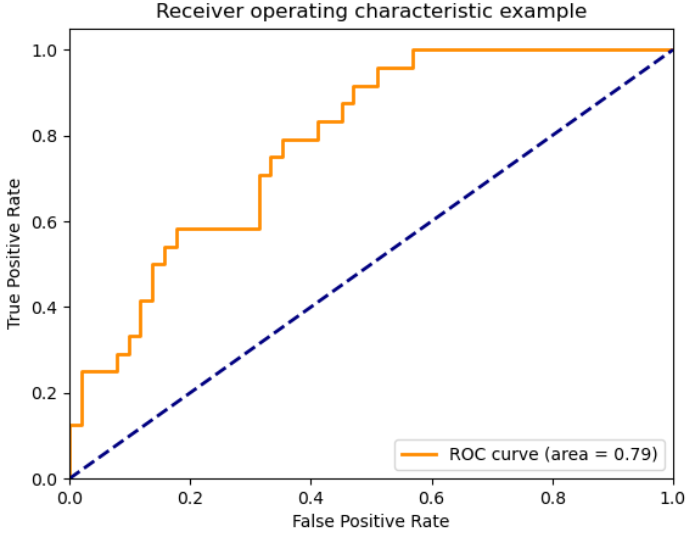
Specificity, sensitivity and ROC curve

$Specifity = \Large\frac{TN}{TN + FP}$

ratio of correctly predicted negative to all items that are actually negative

$Sensitivity = \Large\frac{TP}{FN + TP}$ (same as Recall)

ratio of correctly predicted positive to all items that are actually positive



The ROC curve shows the relationship between specificity and sensitivity for different prediction thresholds (for classifiers that produce probabilistic scores)

The Area Under the Curve AUC can be used as a summary.

Metrics for regression

- Mean Absolute Error: Not sensitive to outliers.

$$MAE = \Large\frac{1}{n} \sum_{i}^{n}|y_i - \hat{y_i}|$$

- Mean Squared Error: Sensitive to outliers.

$$MSE = \Large\frac{1}{n} \sum_{i}^{n}(y_i - \hat{y_i})^2$$

- Coefficient of determination or R^2: How well the model fits the data.