Modeling

Machine learning

Machine Learning is:

"Field of study that gives computers the ability to learn without being explicitly programmed"

~ Arthur Samuel, 1959

- Machine Learning is subfield of Computer Science
- Objective: *Generalize from experience*

- Machine Learning is learning model from set of observations

ML task categories based on “feedback” available to learning system:

- Supervised learning
  - We know the right answers
  - Supervised learning algorithm is inferring decision function from labelled training data. The algorithm needs to generalize from training data to unseen data "reasonably".
- Unsupervised learning
  - We do not know right answers
  - Unsupervised learning algorithm is inferring function, which describes hidden structure of unlabelled data. We cannot estimate error of algorithm.
- Reinforcement Learning
  - Machine interacts with dynamic environment in which it needs to achieve certain goal without teacher telling it if it is close to the goal or not.

ML categories based on ”outputs” produced:

- Classification
- Regression
- Clustering
- Density estimation
- Dimensionality reduction

Supervised learning – Classification OR Regression

- Training dataset with N samples: $T = \{(\overrightarrow{x_1}, y_1), (\overrightarrow{x_2}, y_2), \ldots, (\overrightarrow{x_N}, y_N)\}$
- Machine learning algorithm tries to learn function, which maps features $\overrightarrow{x_i}$ to the corresponding value y_i:

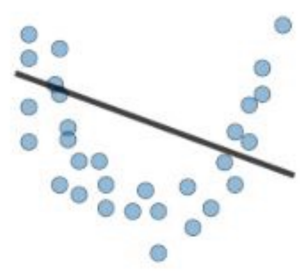
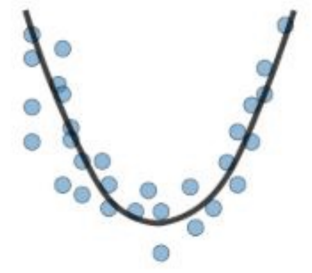
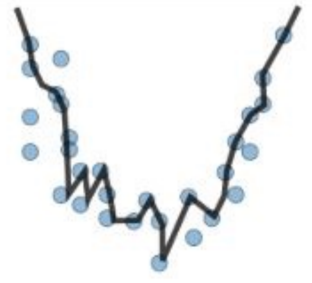
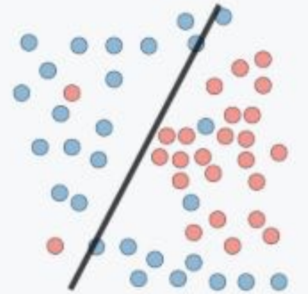
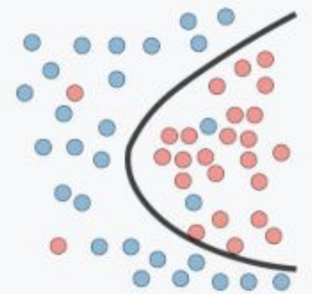
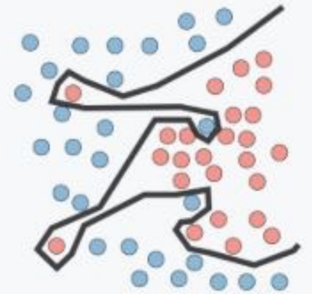
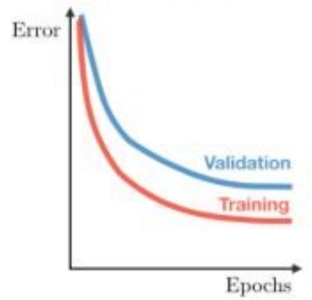
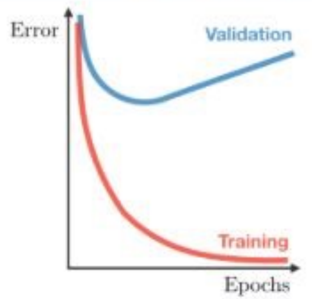
$$\widehat{y_i} = f(\overrightarrow{x_i})$$

- $\widehat{y_i}$ is the estimation of the reality
- Each ML algorithm works with different expectations and under different conditions -> there exist multiple solutions to each task and your task is to pick the best

Model training

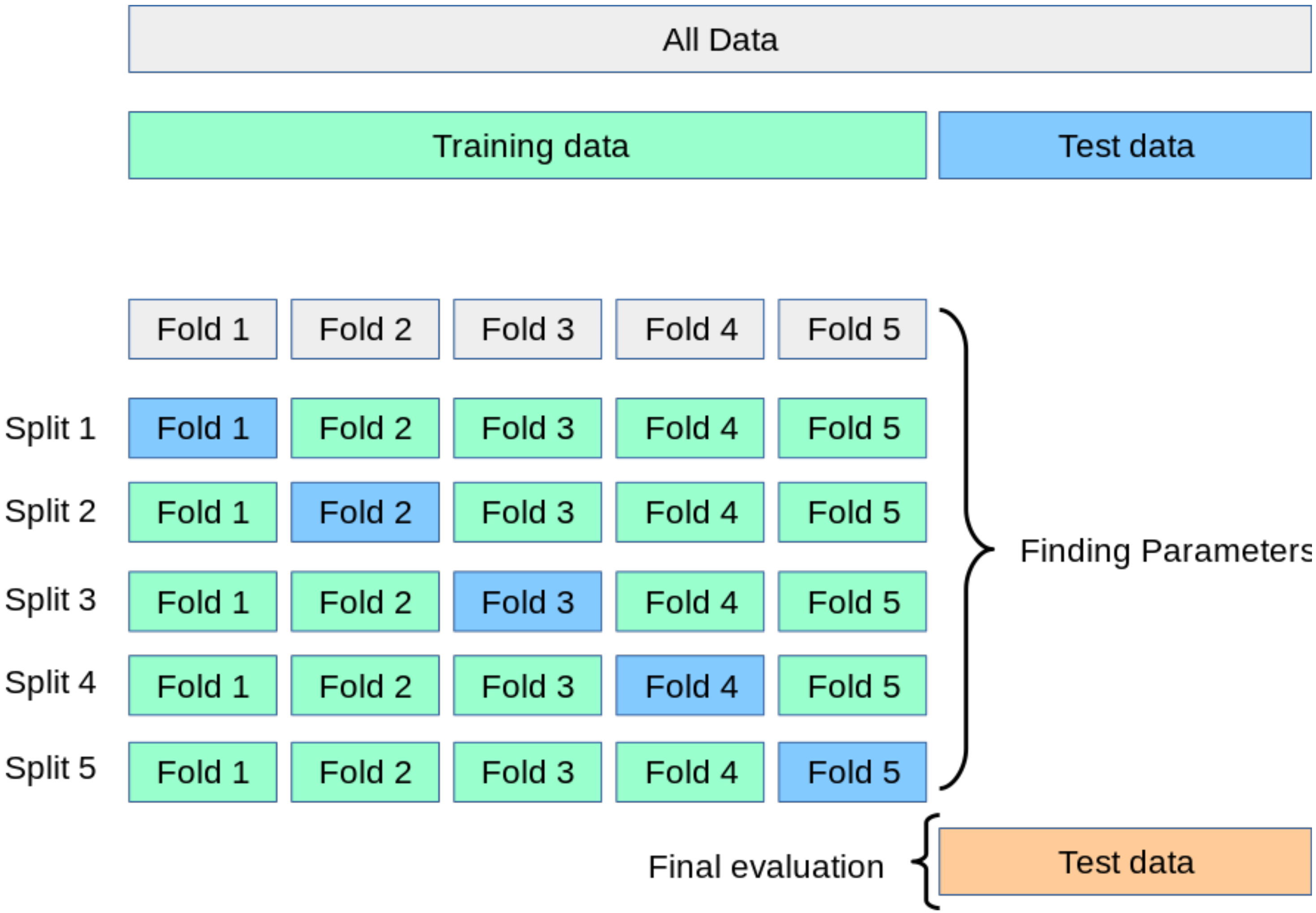
- The ML model is trained (learned) on the sample of the population -> you do not have the whole reality captured
- It is critical to estimate the model error correctly
- To train correctly model training data needs to divided:
  - Training data – used for estimation of model parameters
  - Validation data – used for evaluation (estimation of error) of model on "unseen" data
- Usually you are training model in several cycles (this holds especially for Neural Networks) and stops when the model error reach acceptable value on both validation and training data

Model bias/variance

	Underfitting	Just right	Overfitting
Symptoms	• High training error • Training error close to test error • High bias	• Training error slightly lower than test error	• Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			
Deep learning illustration			
Possible remedies	• Complexify model • Add more features • Train longer		• Perform regularization • Get more data

https://www.kaggle.com/getting-started/166897

K-fold cross-validation



https://scikit-learn.org/stable/modules/cross_validation.html

Evaluation metrics

- For supervised learning
- Classification

		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

- Regression
  - Mean Squared Error: $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
  - Root Mean Squared Error $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$
  - Mean Absolute Error $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$

Accuracy paradox

Predicted class / Actual class	Terrorist	Not terrorist	Sum
Terrorist	0	1	1
Not terrorist	0	99	99
Sum	0	100	100

- Accuracy = 99%
- Use other measures such as Sensitivity (Recall) and Precision (Positive predictive value).
- With highly imbalanced data
- There exists techniques for balancing the dataset (i. e. SMOTE)

https://en.wikipedia.org/wiki/Accuracy_paradox

Conclusion

- Not every model is suitable for given task
- Check model conditions
- Always check the error on both training and validation dataset
- Use cross-validation, especially with small sample size
- Choose proper evaluation metrics