

dm25s1

Topic 06 : Data Modelling

Part 01 : Data Modelling - Introduction

Preparation

Data Handling

Exploring Data 1

Dr Bernard Butler
Department of Computing and Mathematics, WIT.
(bernard.butler@setu.ie)

Exploring Data 2

Building Models

Prediction

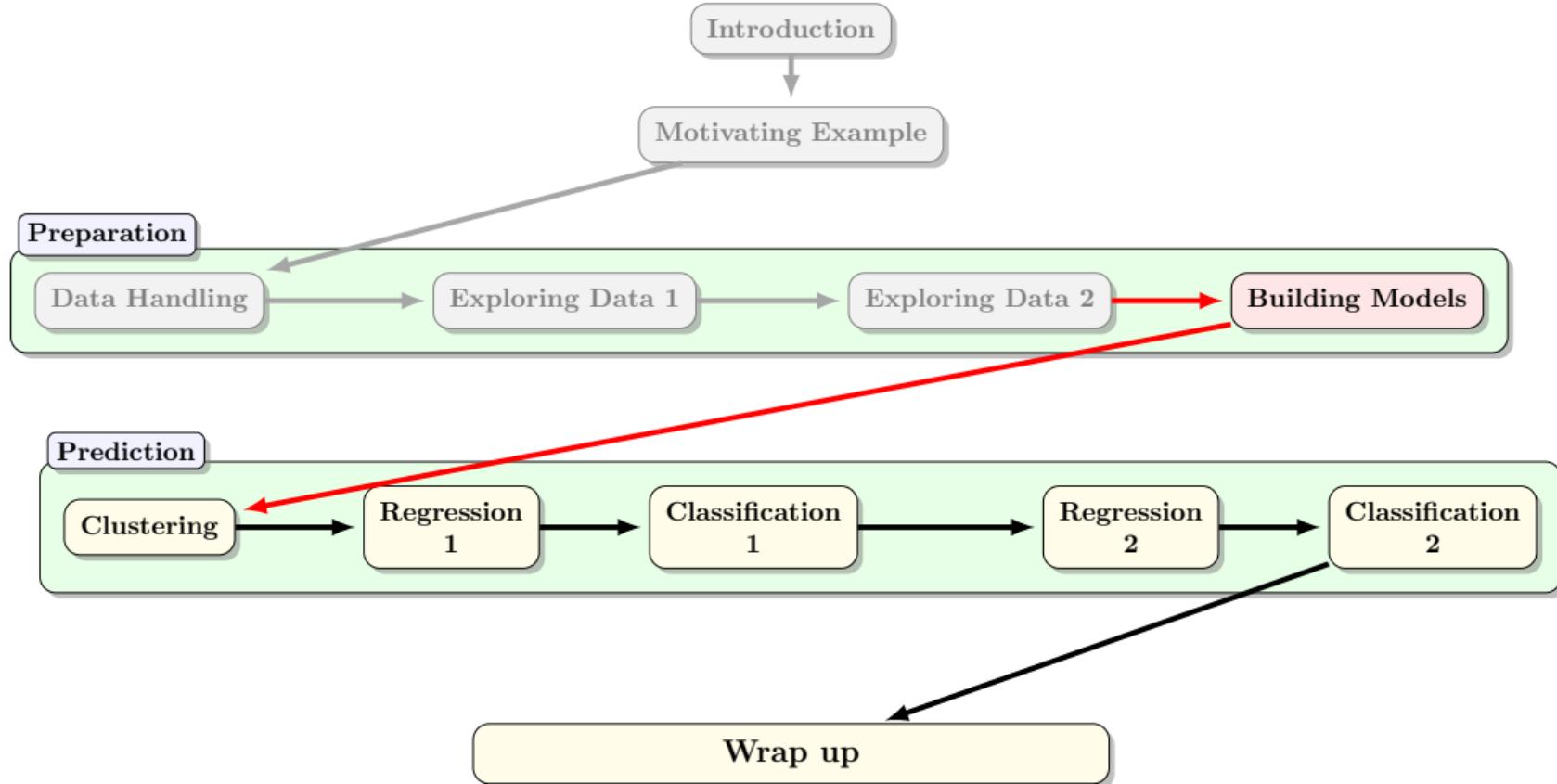
Autumn Semester, 2025

Outline

- Components of a machine learning problem
- Machine learning concepts and notation
- Bias vs variance

Wrap up

Data Mining (Week 6)



Outline

1. Machine Learning (ML) Overview	3
1.1. Components of a Machine Learning Problem	4
1.2. Problem–Task–Experience Perspective	9
1.3. Taxonomy of Machine Learning Methods	10
1.4. Statistical Models vs Machine Learning Models	12
2. Modelling Process	13
2.1. Models and error	15

Terminology / Notation

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Survived
1	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Nan	S	0
2	1	Cumings, Mrs. John Bradley (Florence Briggs Th... e)	female	38.0	1	0	PC 17599	71.2833	C85	C	1
3	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Nan	S	1
4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	1
5	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Nan	S	0
6	3	Moran, Mr. James	male	Nan	0	0	330877	8.4583	Nan	Q	0
7	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S	0
8	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	Nan	S	0
9	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	Nan	S	1
10	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	Nan	C	1
11	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S	1

Terminology / Notation

The diagram illustrates a labeled dataset with the following annotations:

- m observations / instances / cases / rows**: A bracket on the left side groups the 11 rows of data.
- n + 1 columns / variables**: A bracket at the top spans all columns except the last one.
- X**: A bracket below the first 10 columns is labeled X.
- n features / attributes / dimensions**: A bracket below the first 10 columns is labeled n features / attributes / dimensions.
- y target**: A bracket on the far right groups the final column, labeled y target.

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Survived
1	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Nan	S	0
2	1	Cumings, Mrs. John Bradley (Florence Briggs Th... ...	female	38.0	1	0	PC 17599	71.2833	C85	C	1
3	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Nan	S	1
4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	1
5	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Nan	S	0
6	3	Moran, Mr. James	male	Nan	0	0	330877	8.4583	Nan	Q	0
7	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S	0
8	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	Nan	S	0
9	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	Nan	S	1
10	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	Nan	C	1
11	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S	1

- A labeled dataset consists of m rows \times $(n + 1)$ columns / variables.
- Use bold to represent vectors and matrices.

Terminology / Notation

m observations / instances / cases / rows

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Survived
1	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Nan	S	0
2	1	Cumings, Mrs. John Bradley (Florence Briggs Th... ...	female	38.0	1	0	PC 17599	71.2833	C85	C	1
3	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Nan	S	1
4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	1
5	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Nan	S	0
6	3	Moran, Mr. James	male	Nan	0	0	330877	8.4583	Nan	Q	0
7	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S	0
8	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	Nan	S	0
9	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	Nan	S	1
10	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	Nan	C	1
11	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S	1

- A labeled dataset consists of m rows \times $(n + 1)$ columns / variables.
- Use bold to represent vectors and matrices.
- Use subscripts to indicate particular feature / attribute / column x_j

Terminology / Notation

m observations / instances / cases / rows

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Survived
1	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Nan	S	0
2	1	Cumings, Mrs. John Bradley (Florence Briggs Th... ...	female	38.0	1	0	PC 17599	71.2833	C85	C	1
3	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Nan	S	1
4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	1
5	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Nan	S	0
6	3	Moran, Mr. James	male	Nan	0	0	330877	8.4583	Nan	Q	0
7	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S	0
8	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	Nan	S	0
9	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	Nan	S	1
10	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	Nan	C	1
11	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S	1

x_j

$x^{(i)}$

- A labeled dataset consists of m rows \times $(n + 1)$ columns / variables.
- Use bold to represent vectors and matrices.
- Use subscripts to indicate particular **feature / attribute / column** x_j
- Use superscript in parenthesis to indicate particular **observation / instance/ case / row** $x^{(i)}$

Terminology / Notation

The diagram illustrates a labeled dataset with the following annotations:

- m observations / instances / cases / rows**: A bracket on the left side groups all 11 rows.
- n + 1 columns / variables**: A bracket at the top spans all columns except the last one.
- n features / attributes / dimensions**: A bracket below the first n columns.
- X**: A label for the matrix of features.
- y target**: A label for the target variable column.
- x⁽ⁱ⁾**: A label for the i-th observation.
- x_j**: A label for the j-th feature.

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Survived
1	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Nan	S	0
2	1	Cumings, Mrs. John Bradley (Florence Briggs Th... ...	female	38.0	1	0	PC 17599	71.2833	C85	C	1
3	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Nan	S	1
4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	1
5	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Nan	S	0
6	3	Moran, Mr. James	male	Nan	0	0	330877	8.4583	Nan	Q	0
7	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S	0
8	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	Nan	S	0
9	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	Nan	S	1
10	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	Nan	C	1
11	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S	1

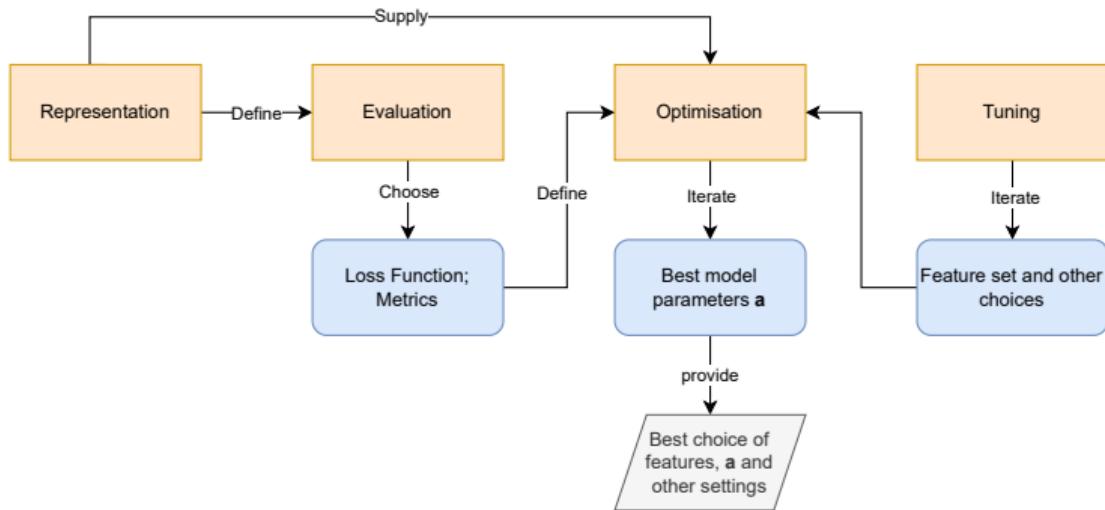
- A labeled dataset consists of **m** rows \times (**n + 1**) columns / variables.
- Use bold to represent vectors and matrices.
- Use subscripts to indicate particular **feature / attribute / column** **x_j**
- Use superscript in parenthesis to indicate particular **observation / instance/ case / row** **x⁽ⁱ⁾**
- So **x_j⁽ⁱ⁾** (or **x_{i,j}**) is the **i**-th observation in the **j**-th feature **x_j⁽ⁱ⁾**

Components of a Machine Learning Problem

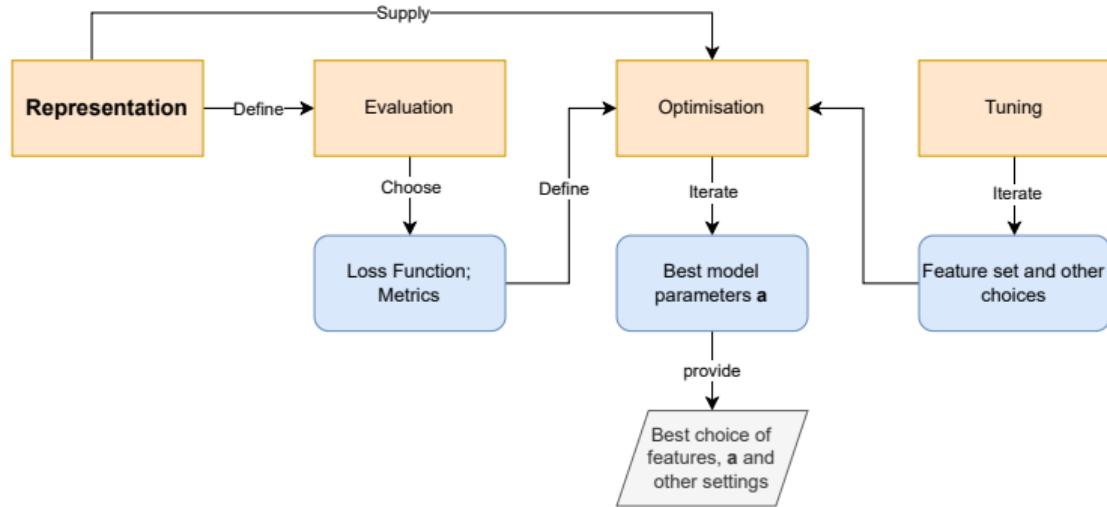
Where do we start????

Components of a Machine Learning Problem

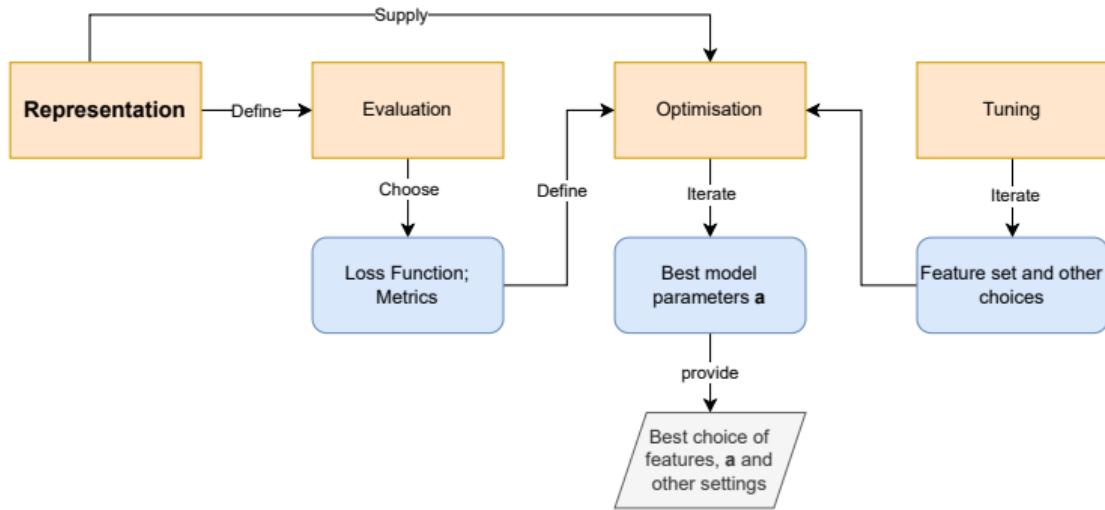
Where do we start????



Component 1: Representation

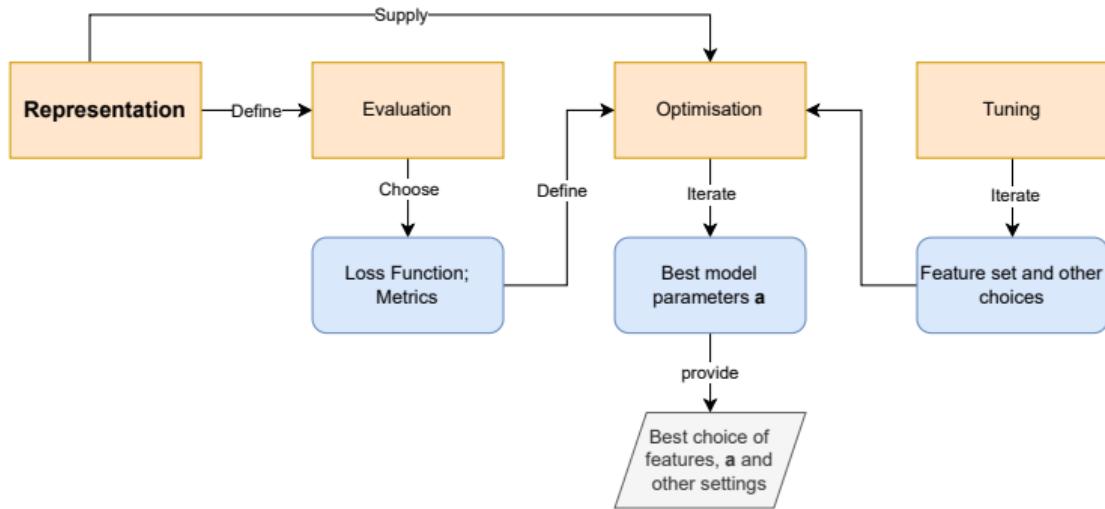


Component 1: Representation



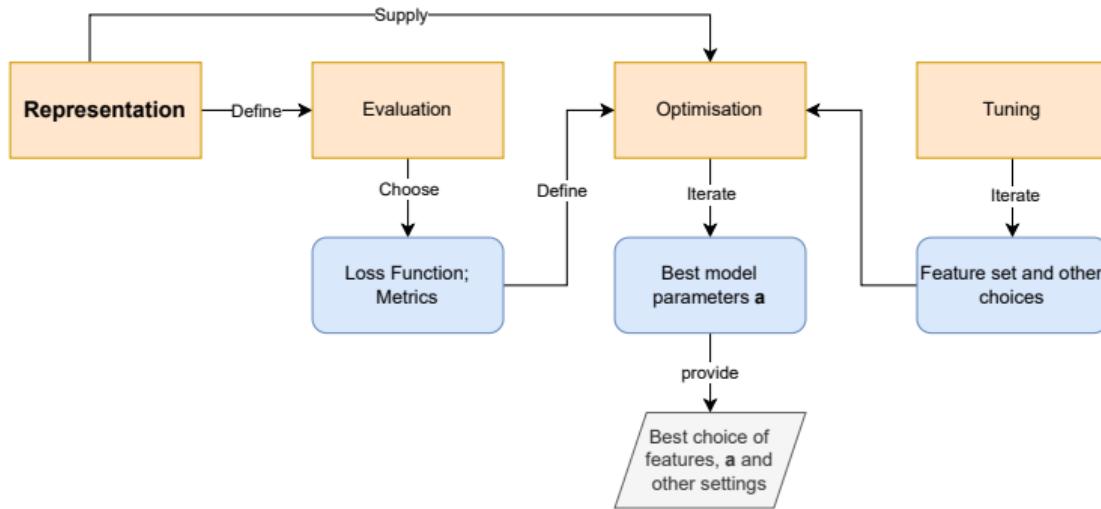
- Given the business objective and the data, choose what type of ML problem to solve: regression, classification, clustering, etc.

Component 1: Representation



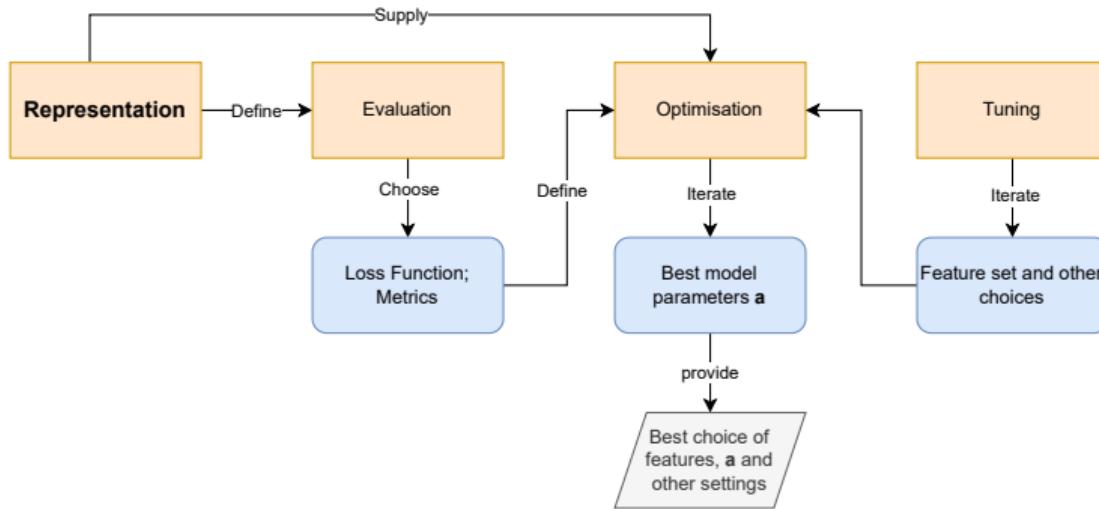
- Given the business objective and the data, choose what type of ML problem to solve: regression, classification, clustering, etc.
- Choose an algorithm for that problem. For example, if the ML problem is classification, we can choose KNN (other techniques will be seen later)

Component 1: Representation



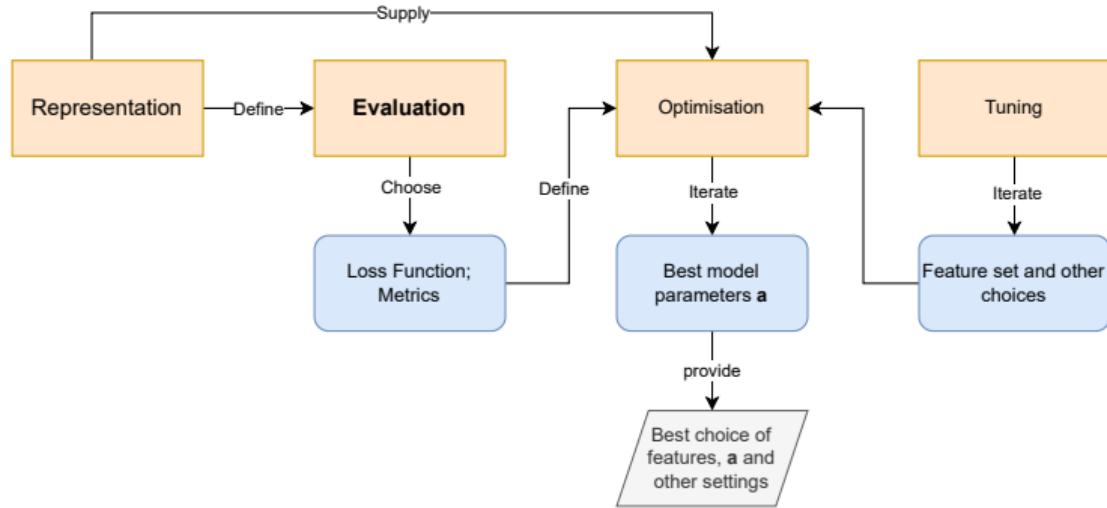
- Given the business objective and the data, choose what type of ML problem to solve: regression, classification, clustering, etc.
- Choose an algorithm for that problem. For example, if the ML problem is classification, we can choose KNN (other techniques will be seen later)
- What is the target and what features are available?

Component 1: Representation

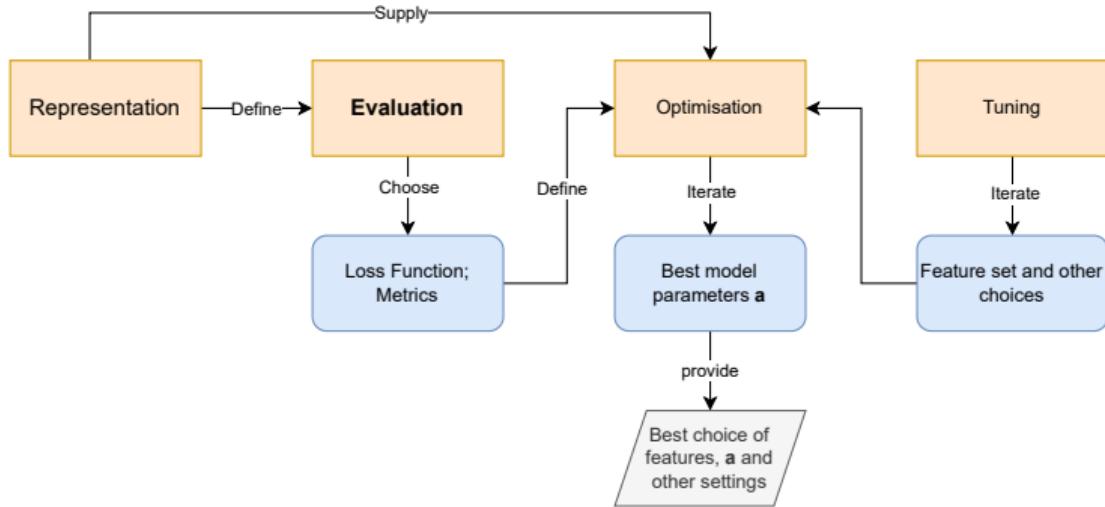


- Given the business objective and the data, choose what type of ML problem to solve: regression, classification, clustering, etc.
- Choose an algorithm for that problem. For example, if the ML problem is classification, we can choose KNN (other techniques will be seen later)
- What is the target and what features are available?
- These choices are made once - they are foundational for what follows.

Component 2: Evaluation

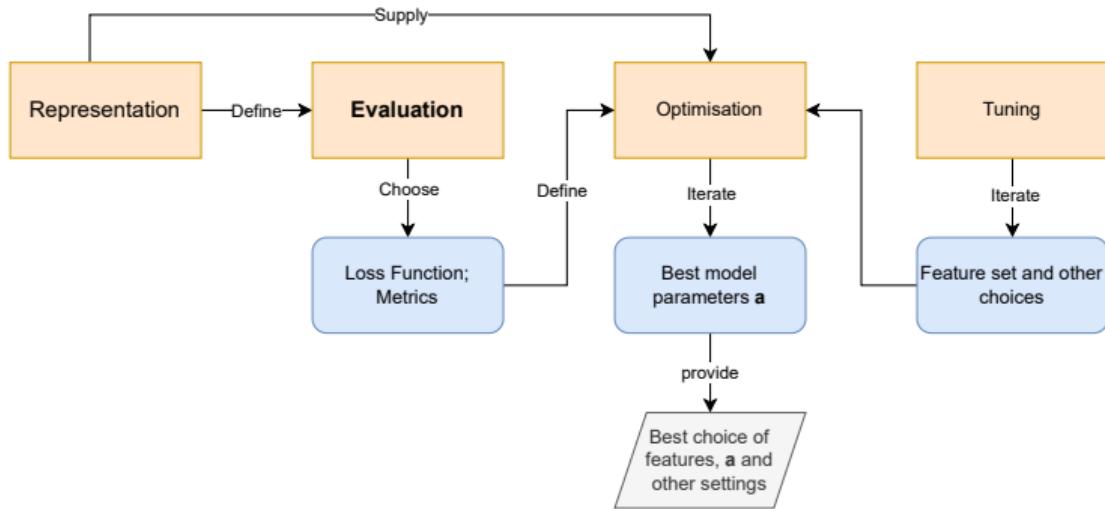


Component 2: Evaluation



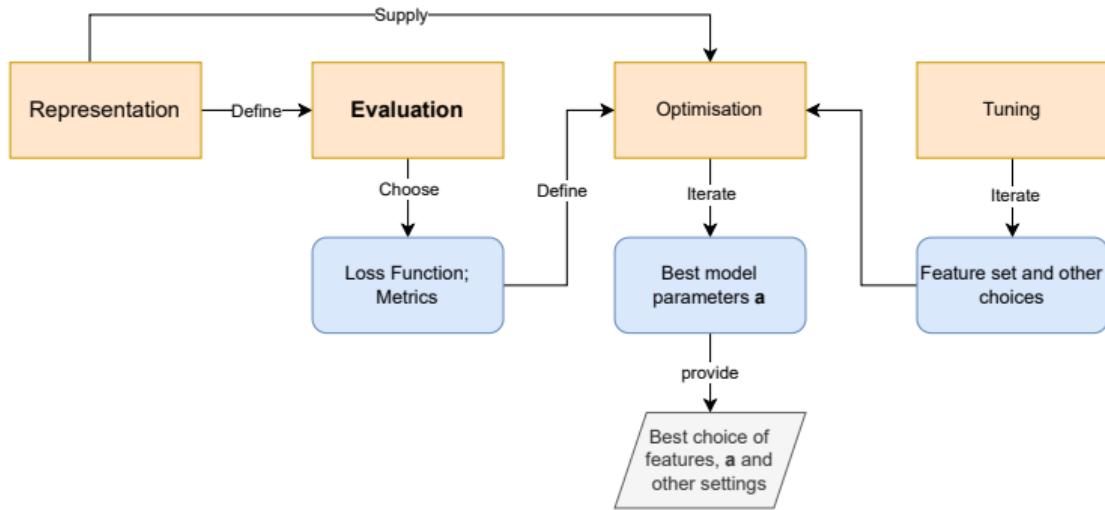
- Need an **Objective Function** that we can use to measure how well the model represents the data.

Component 2: Evaluation



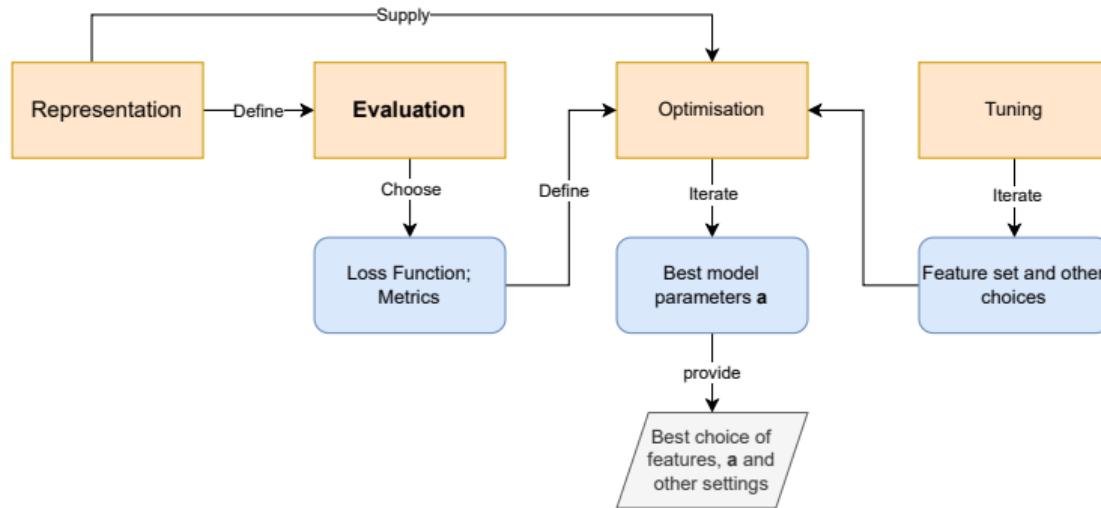
- Need an **Objective Function** that we can use to measure how well the model represents the data.
- For classification: how often does the model predict the right class, so that the predicted target and the actual target match each other?

Component 2: Evaluation



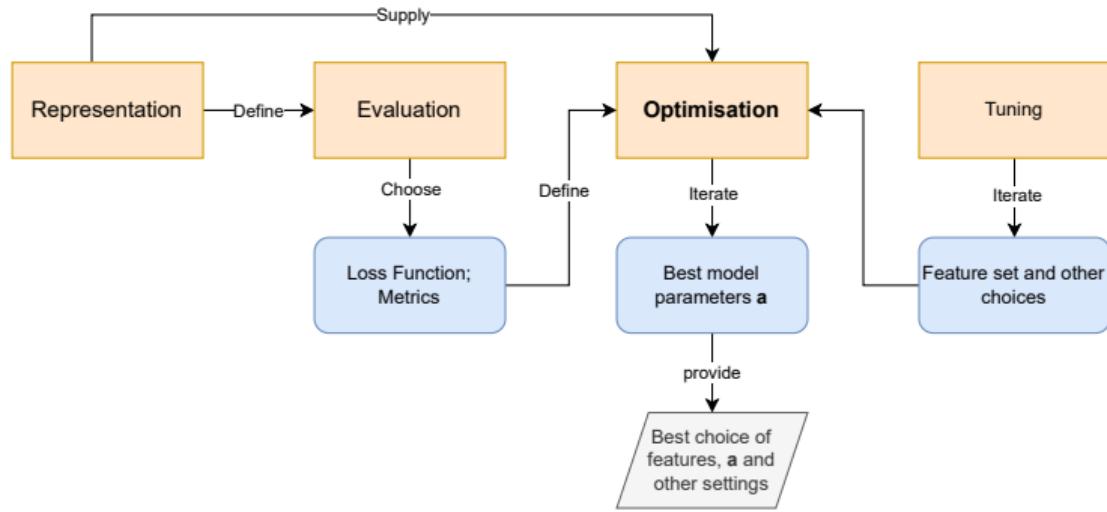
- Need an **Objective Function** that we can use to measure how well the model represents the data.
- For classification: how often does the model predict the right class, so that the predicted target and the actual target match each other?
- For regression: how close are the predicted target values to the actual target values?

Component 2: Evaluation

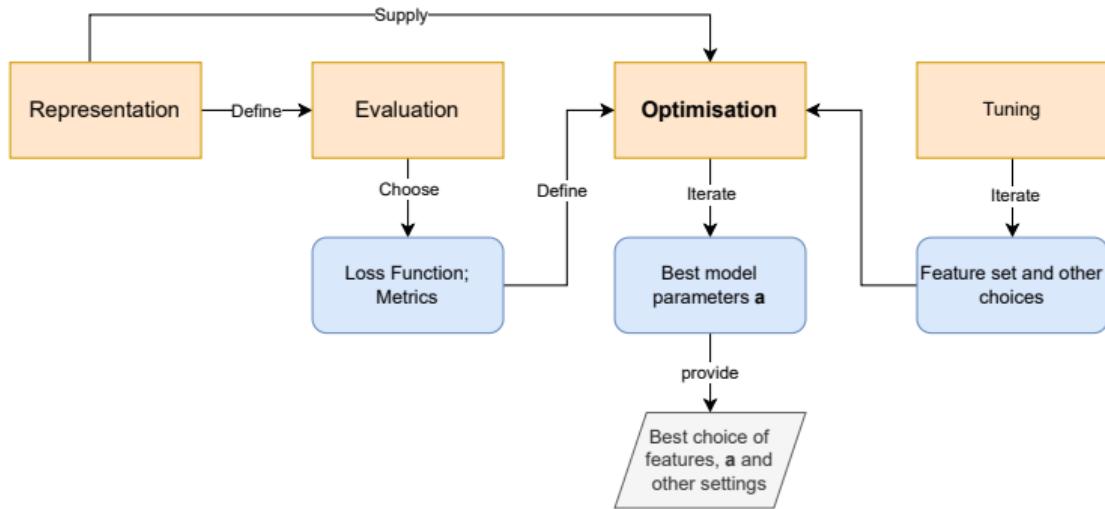


- Need an **Objective Function** that we can use to measure how well the model represents the data.
- For classification: how often does the model predict the right class, so that the predicted target and the actual target match each other?
- For regression: how close are the predicted target values to the actual target values?
- Generally the objective function takes actual and predicted target values and returns a single nonnegative number.

Component 3: Optimisation

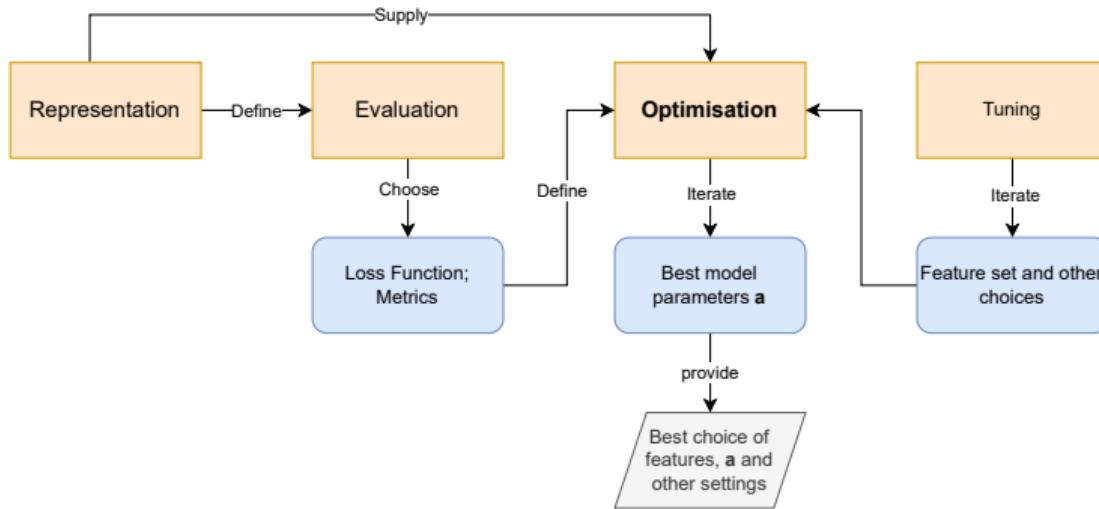


Component 3: Optimisation



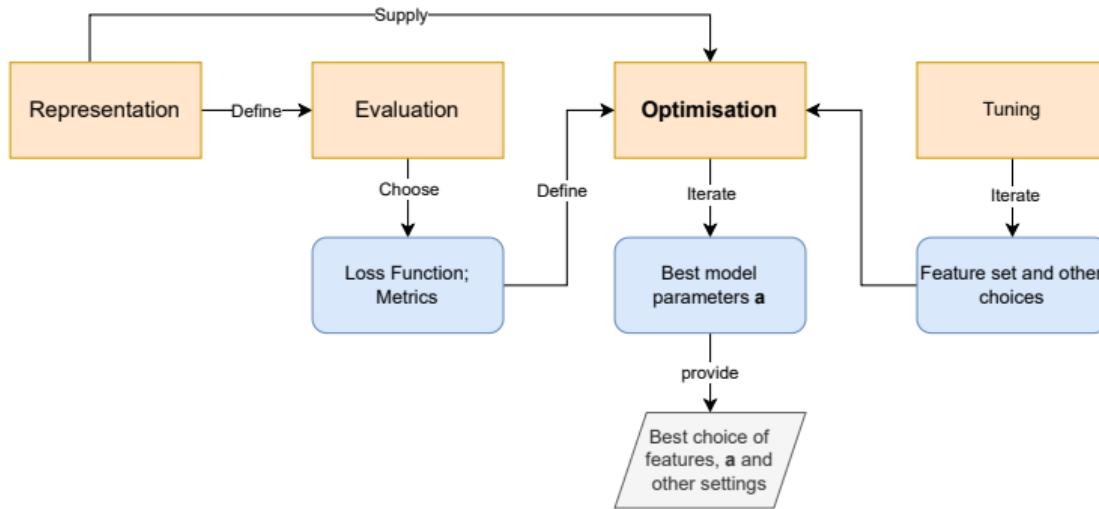
- Most ML algorithms are *iterative*: start with a guess at the model parameters $\textcolor{green}{a}$, and improve them until the process terminates.

Component 3: Optimisation



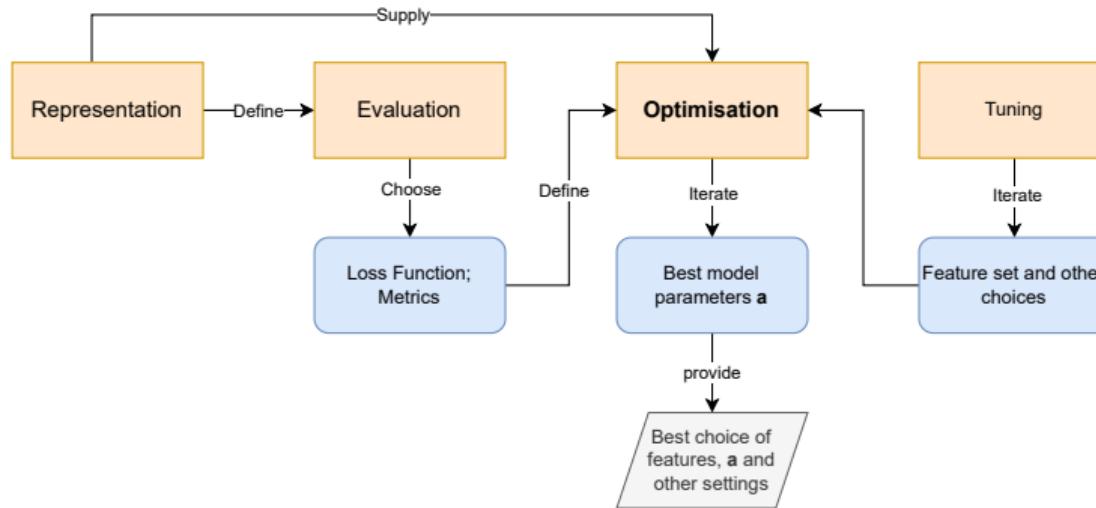
- Most ML algorithms are *iterative*: start with a guess at the model parameters $\textcolor{green}{a}$, and improve them until the process terminates.
- The nature of $\textcolor{green}{a}$ and how the code searches for better $\textcolor{green}{a}$ depends on the **Representation**, refined by the **Tuning** choices - we will see examples later

Component 3: Optimisation



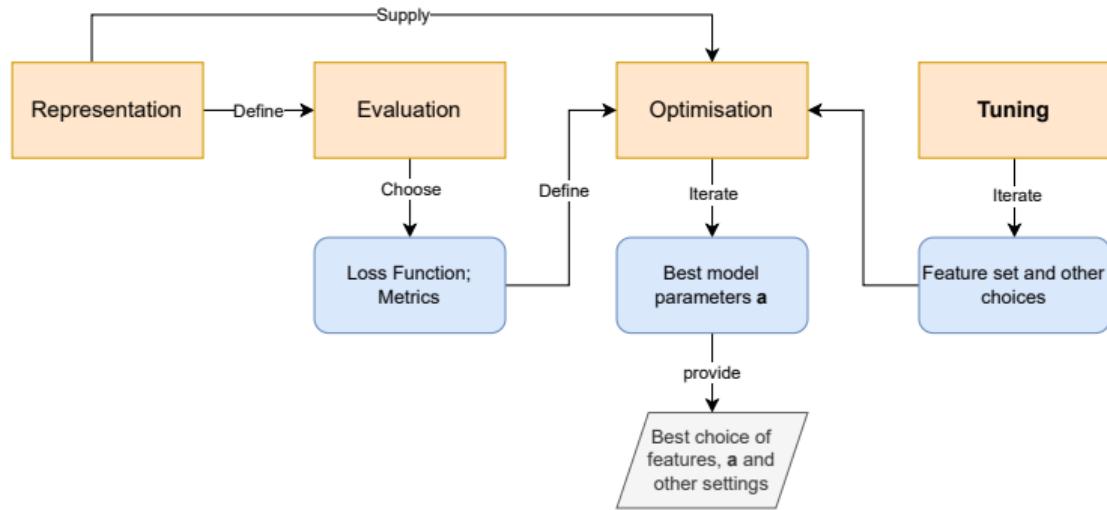
- Most ML algorithms are *iterative*: start with a guess at the model parameters $\textcolor{green}{a}$, and improve them until the process terminates.
- The nature of $\textcolor{green}{a}$ and how the code searches for better $\textcolor{green}{a}$ depends on the **Representation**, refined by the **Tuning** choices - we will see examples later
- The choice of when to terminate the optimisation procedure depends on the **Evaluation** choice.

Component 3: Optimisation

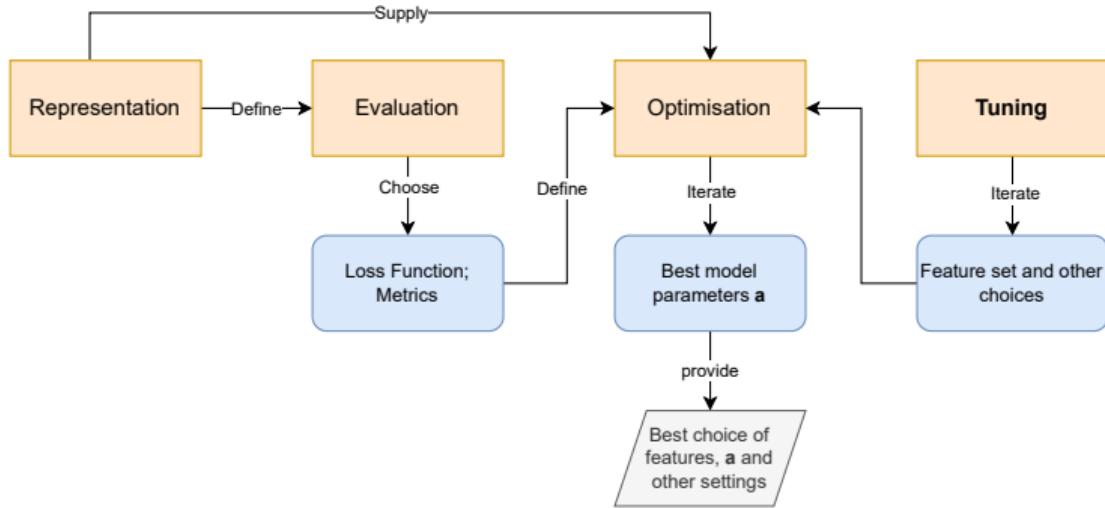


- Most ML algorithms are *iterative*: start with a guess at the model parameters α , and improve them until the process terminates.
- The nature of α and how the code searches for better α depends on the **Representation**, refined by the **Tuning** choices - we will see examples later
- The choice of when to terminate the optimisation procedure depends on the **Evaluation** choice.
- Most of the computational resources are needed for the **Optimisation** component.

Component 4: Tuning

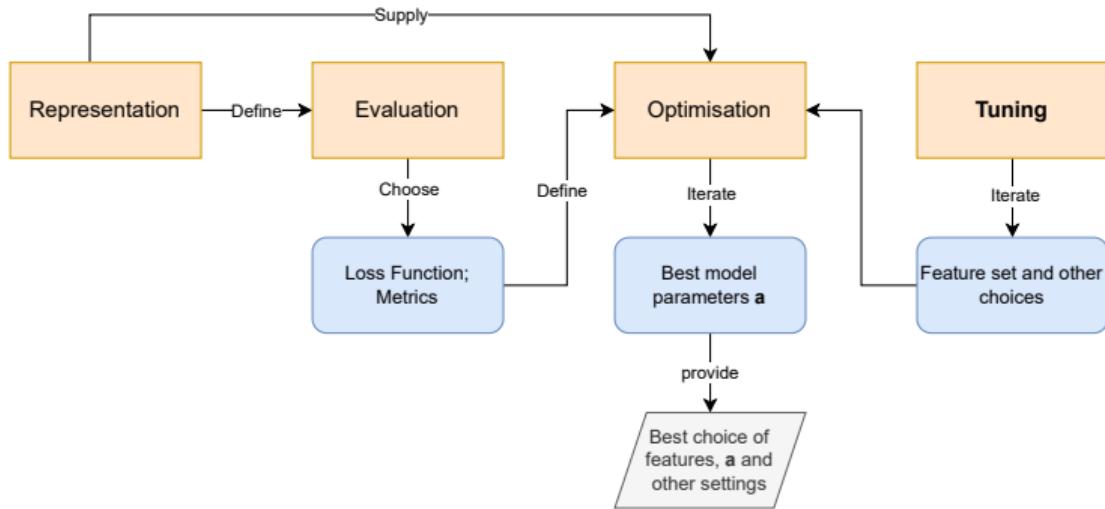


Component 4: Tuning



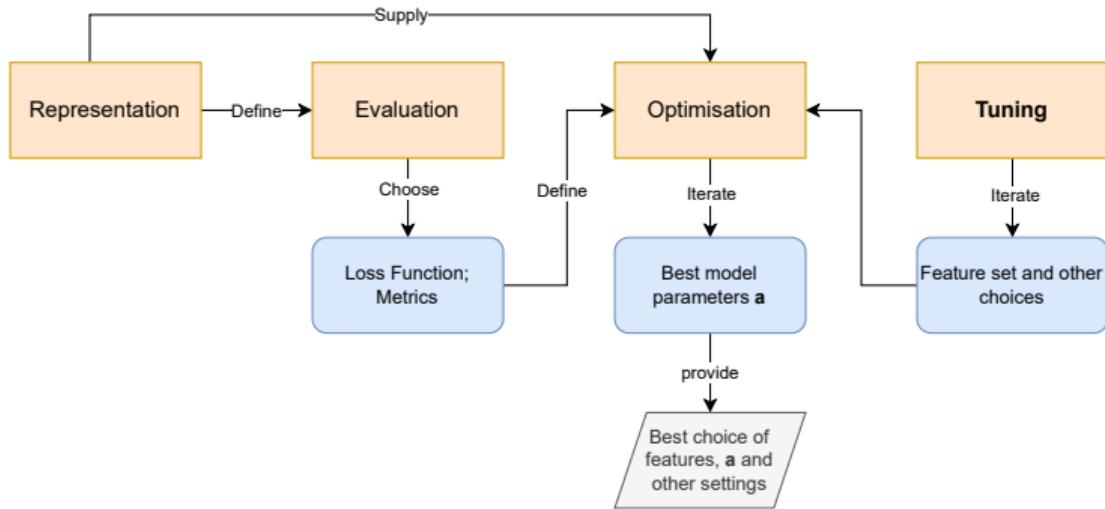
- Tuning is also iterative and is applied on top of Optimisation

Component 4: Tuning



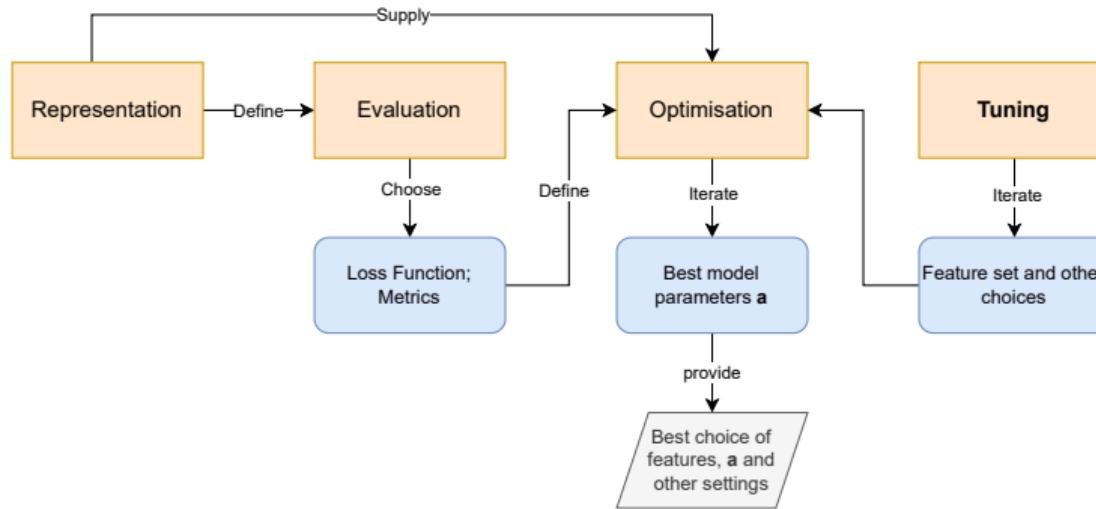
- Tuning is also iterative and is applied on top of Optimisation
- Tuning parameters (also known as *hyperparameters*) are kept constant during **Optimisation**

Component 4: Tuning



- Tuning is also iterative and is applied on top of Optimisation
- Tuning parameters (also known as *hyperparameters*) are kept constant during **Optimisation**
- Tuning parameters depend on **Representation**. For example, we have seen one hyperparameter already: ***k*** in the KNN classifier.

Component 4: Tuning



- Tuning is also iterative and is applied on top of Optimisation
- Tuning parameters (also known as *hyperparameters*) are kept constant during **Optimisation**
- Tuning parameters depend on **Representation**. For example, we have seen one hyperparameter already: *k* in the KNN classifier.
- For **Tuning**, GridSearch (exhaustive over a range) and RandomSearch (pick representative values) are popular ways to search for the best hyperparameter values.

Data Modelling (aka Machine Learning)

As alternative to the four component (Representation / Evaluation / Optimisation/ Tuning) viewpoint we can think of a machine learning problem as

Definition 1 (Machine Learning)

Study of algorithms that improve their performance P at some task T with experience E .

Well defined learning task: $\langle P, T, E \rangle$

Data Modelling (aka Machine Learning)

As alternative to the four component (Representation / Evaluation / Optimisation/ Tuning) viewpoint we can think of a machine learning problem as

Definition 1 (Machine Learning)

Study of algorithms that improve their performance P at some task T with experience E .

Well defined learning task: $\langle P, T, E \rangle$

- What metric should be used to measure performance?
- What cost function should be used?
- What is the cost of incorrect prediction?
- Computational cost?

Data Modelling (aka Machine Learning)

As alternative to the four component (Representation / Evaluation / Optimisation/ Tuning) viewpoint we can think of a machine learning problem as

Definition 1 (Machine Learning)

Study of algorithms that improve their performance P at some task T with experience E .

Well defined learning task: $\langle P, T, E \rangle$

- What metric should be used to measure performance?
- What cost function should be used?
- What is the cost of incorrect prediction?
- Computational cost?

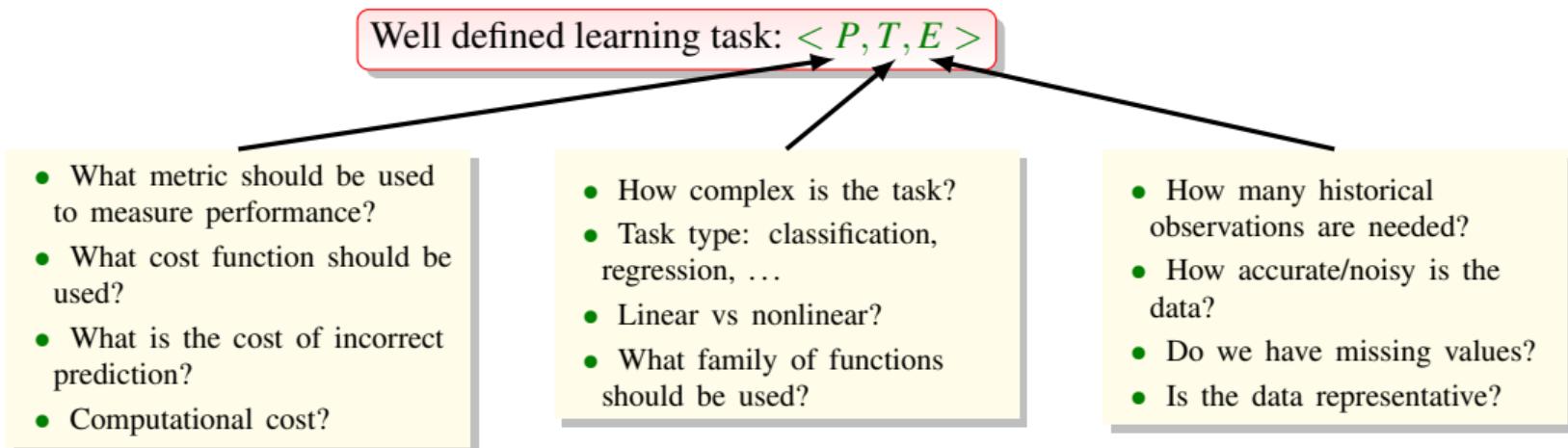
- How complex is the task?
- Task type: classification, regression, ...
- Linear vs nonlinear?
- What family of functions should be used?

Data Modelling (aka Machine Learning)

As alternative to the four component (Representation / Evaluation / Optimisation/ Tuning) viewpoint we can think of a machine learning problem as

Definition 1 (Machine Learning)

Study of algorithms that improve their performance P at some task T with experience E .



Taxonomy of Machine Learning Models . . .

... by Intuition/Motivation

... by Algorithmic Properties

... by Fixed/Variable Number of Parameters

Taxonomy of Machine Learning Models ...

... by Intuition/Motivation

- **Geometric models** use intuitions from geometry such as separating (hyper-)planes, linear transformations and distance metrics.
- **Probabilistic models** view learning as a process of reducing uncertainty, modelled by means of probability distributions.

... by Algorithmic Properties

... by Fixed/Variable Number of Parameters

Taxonomy of Machine Learning Models ...

... by Intuition/Motivation

- **Geometric models** use intuitions from geometry such as separating (hyper-)planes, linear transformations and distance metrics.
- **Probabilistic models** view learning as a process of reducing uncertainty, modelled by means of probability distributions.
- **Logical models** are defined in terms of easily interpretable logical expressions.

... by Algorithmic Properties

... by Fixed/Variable Number of Parameters

Taxonomy of Machine Learning Models ...

... by Intuition/Motivation

- **Geometric models** use intuitions from geometry such as separating (hyper-)planes, linear transformations and distance metrics.
- **Probabilistic models** view learning as a process of reducing uncertainty, modelled by means of probability distributions.
- **Logical models** are defined in terms of easily interpretable logical expressions.

... by Algorithmic Properties

- **Regression models** predict a numeric output.

... by Fixed/Variable Number of Parameters

Taxonomy of Machine Learning Models ...

... by Intuition/Motivation

- **Geometric models** use intuitions from geometry such as separating (hyper-)planes, linear transformations and distance metrics.
- **Probabilistic models** view learning as a process of reducing uncertainty, modelled by means of probability distributions.
- **Logical models** are defined in terms of easily interpretable logical expressions.

... by Algorithmic Properties

- **Regression models** predict a numeric output.
- **Classification models** predict a discrete class value.

... by Fixed/Variable Number of Parameters

Taxonomy of Machine Learning Models ...

... by Intuition/Motivation

- **Geometric models** use intuitions from geometry such as separating (hyper-)planes, linear transformations and distance metrics.
- **Probabilistic models** view learning as a process of reducing uncertainty, modelled by means of probability distributions.
- **Logical models** are defined in terms of easily interpretable logical expressions.

... by Algorithmic Properties

- **Regression models** predict a numeric output.
- **Classification models** predict a discrete class value.
- **Neural networks** learn based on a biological analogy

... by Fixed/Variable Number of Parameters

Taxonomy of Machine Learning Models ...

... by Intuition/Motivation

- **Geometric models** use intuitions from geometry such as separating (hyper-)planes, linear transformations and distance metrics.
- **Probabilistic models** view learning as a process of reducing uncertainty, modelled by means of probability distributions.
- **Logical models** are defined in terms of easily interpretable logical expressions.

... by Algorithmic Properties

- **Regression models** predict a numeric output.
- **Classification models** predict a discrete class value.
- **Neural networks** learn based on a biological analogy
- **Local models predict** in the local region of a query instance.

... by Fixed/Variable Number of Parameters

Taxonomy of Machine Learning Models ...

... by Intuition/Motivation

- **Geometric models** use intuitions from geometry such as separating (hyper-)planes, linear transformations and distance metrics.
- **Probabilistic models** view learning as a process of reducing uncertainty, modelled by means of probability distributions.
- **Logical models** are defined in terms of easily interpretable logical expressions.

... by Algorithmic Properties

- **Regression models** predict a numeric output.
- **Classification models** predict a discrete class value.
- **Neural networks** learn based on a biological analogy
- **Local models predict** in the local region of a query instance.
- **Tree-based models** (recursively) partition the data to make predictions.

... by Fixed/Variable Number of Parameters

Taxonomy of Machine Learning Models ...

... by Intuition/Motivation

- **Geometric models** use intuitions from geometry such as separating (hyper-)planes, linear transformations and distance metrics.
- **Probabilistic models** view learning as a process of reducing uncertainty, modelled by means of probability distributions.
- **Logical models** are defined in terms of easily interpretable logical expressions.

... by Algorithmic Properties

- **Regression models** predict a numeric output.
- **Classification models** predict a discrete class value.
- **Neural networks** learn based on a biological analogy
- **Local models predict** in the local region of a query instance.
- **Tree-based models** (recursively) partition the data to make predictions.
- **Ensembles** learn multiple models and combine their predictions.

... by Fixed/Variable Number of Parameters

Taxonomy of Machine Learning Models ...

... by Intuition/Motivation

- **Geometric models** use intuitions from geometry such as separating (hyper-)planes, linear transformations and distance metrics.
- **Probabilistic models** view learning as a process of reducing uncertainty, modelled by means of probability distributions.
- **Logical models** are defined in terms of easily interpretable logical expressions.

... by Algorithmic Properties

- **Regression models** predict a numeric output.
- **Classification models** predict a discrete class value.
- **Neural networks** learn based on a biological analogy
- **Local models predict** in the local region of a query instance.
- **Tree-based models** (recursively) partition the data to make predictions.
- **Ensembles** learn multiple models and combine their predictions.

... by Fixed/Variable Number of Parameters

- **Parametric models** have a fixed number of parameters.

Taxonomy of Machine Learning Models ...

... by Intuition/Motivation

- **Geometric models** use intuitions from geometry such as separating (hyper-)planes, linear transformations and distance metrics.
- **Probabilistic models** view learning as a process of reducing uncertainty, modelled by means of probability distributions.
- **Logical models** are defined in terms of easily interpretable logical expressions.

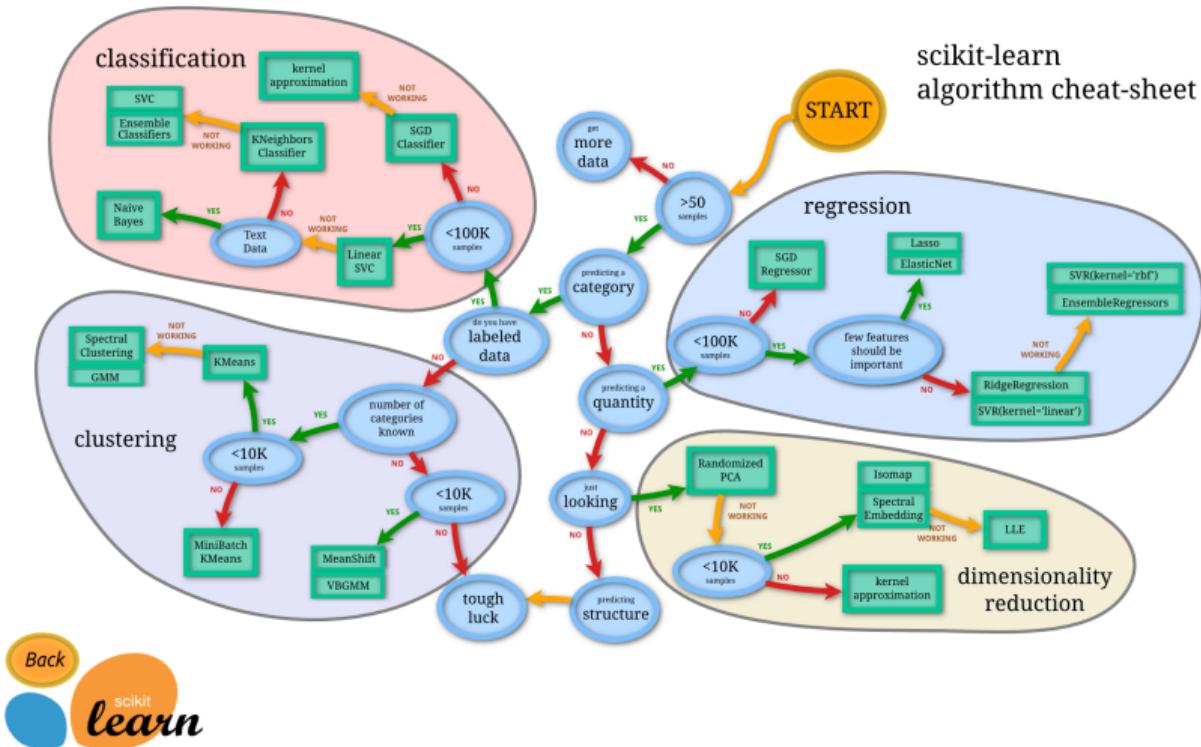
... by Algorithmic Properties

- **Regression models** predict a numeric output.
- **Classification models** predict a discrete class value.
- **Neural networks** learn based on a biological analogy
- **Local models predict** in the local region of a query instance.
- **Tree-based models** (recursively) partition the data to make predictions.
- **Ensembles** learn multiple models and combine their predictions.

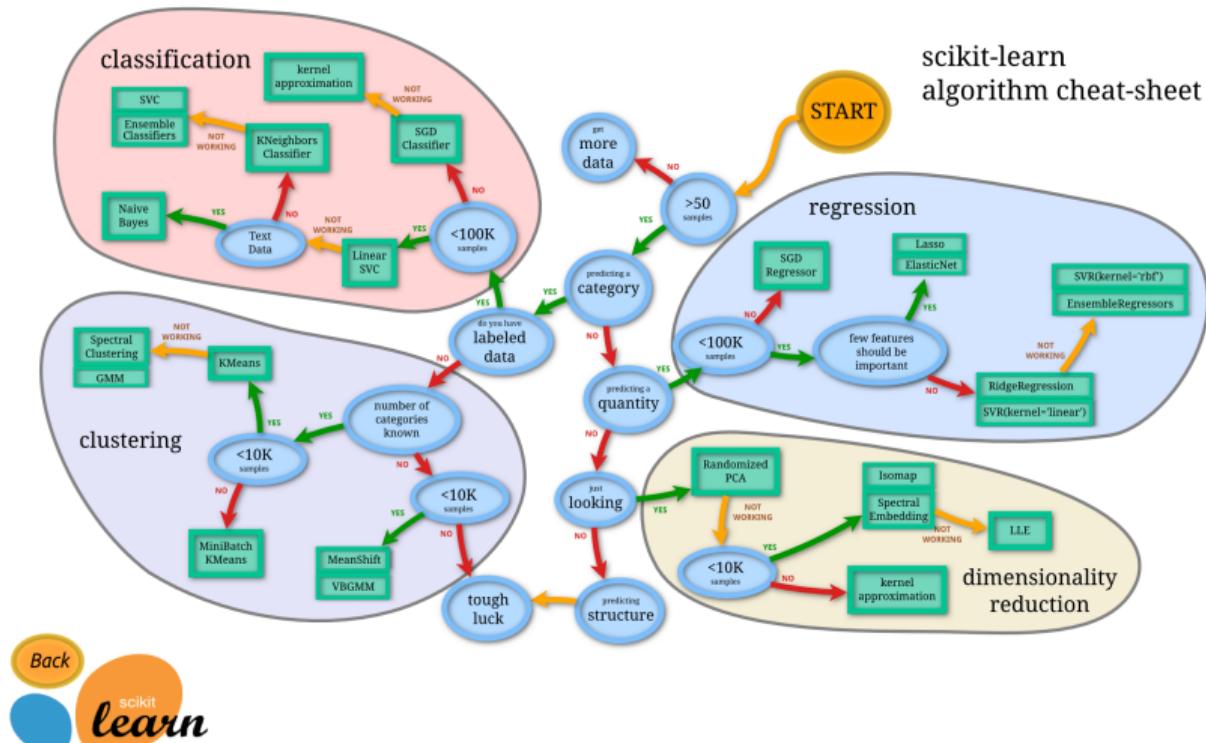
... by Fixed/Variable Number of Parameters

- **Parametric models** have a fixed number of parameters.
- In **non-parametric models** the number of parameters grows with the amount of training data.

Aside: Scikit-learn Flowchart of Models (Shallow Learners)

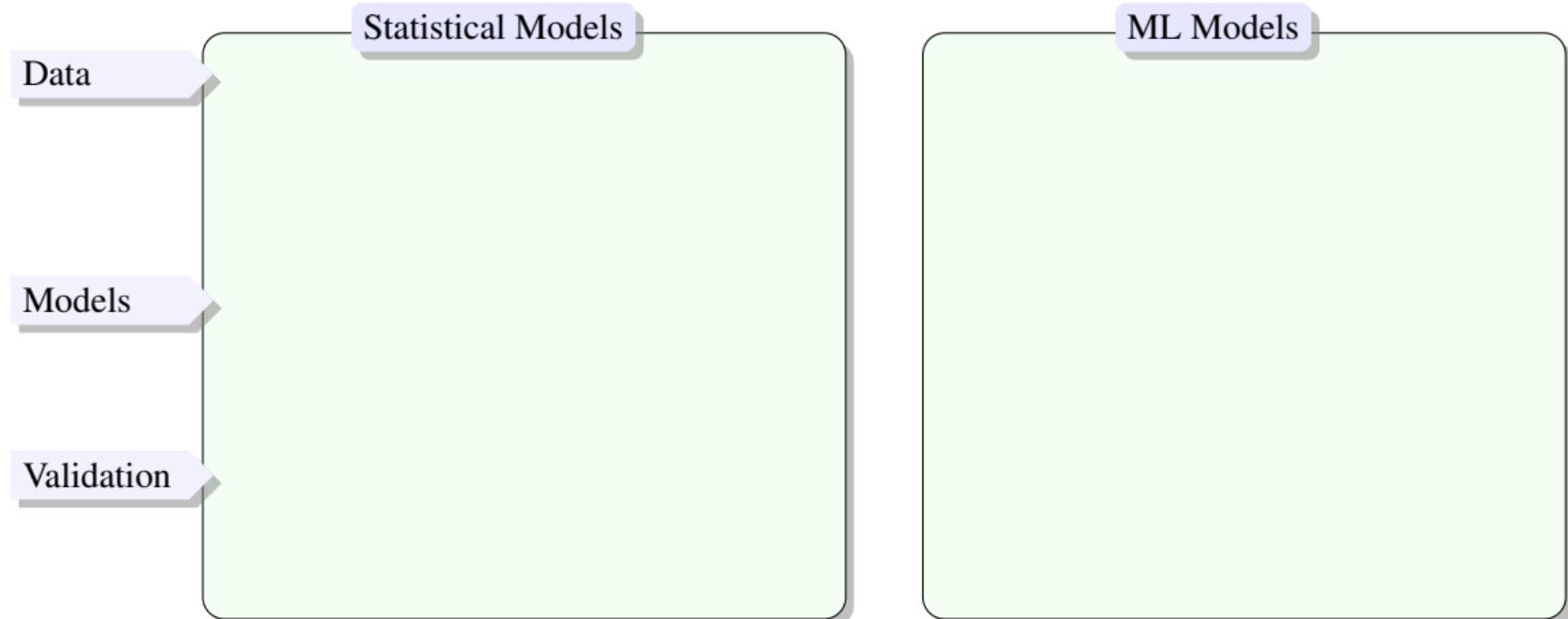


Aside: Scikit-learn Flowchart of Models (Shallow Learners)



A neural network with more than one hidden layer is called a **deep learner**, all other learners are **shallow learners**.

Statistical Models vs Machine Learning Models



Statistical Models vs Machine Learning Models

Data

Statistical Models

- Usually small (< 1000 observations)
- Low dimension (< 10 variables)
- Can have detailed understanding of data
- Data is clean — human has looked at each data point

Models

ML Models

- Can be huge (million+ observations)
- Large dimension (1000+, more for vision)
- Too large for human to parse / understand
- Data not clean — humans can't afford to understand/fix each point

Validation

Statistical Models vs Machine Learning Models

Data

Statistical Models

- Usually small (< 1000 observations)
- Low dimension (< 10 variables)
- Can have detailed understanding of data
- Data is clean — human has looked at each data point

Models

- Simple models — complexity limited by theory
- Detailed/complex statistical assumptions re data
- Model known, and data is carefully examined to verify assumptions.

Validation

ML Models

- Can be huge (million+ observations)
- Large dimension (1000+, more for vision)
- Too large for human to parse / understand
- Data not clean — humans can't afford to understand/fix each point

- “No” upper limit on model complexity
- Fewer statistical assumptions re data
- Don’t know right model? No problem! have multiple models and vote/weight results

Statistical Models vs Machine Learning Models

Data

Statistical Models

- Usually small (< 1000 observations)
- Low dimension (< 10 variables)
- Can have detailed understanding of data
- Data is clean — human has looked at each data point

Models

ML Models

- Can be huge (million+ observations)
- Large dimension (1000+, more for vision)
- Too large for human to parse / understand
- Data not clean — humans can't afford to understand/fix each point

Validation

- Simple models — complexity limited by theory
- Detailed/complex statistical assumptions re data
- Model known, and data is carefully examined to verify assumptions.
- Evaluation based on theoretical estimates under stated statistical assumptions
- Analysis of errors using theoretical distributions

- “No” upper limit on model complexity
- Fewer statistical assumptions re data
- Don’t know right model? No problem! have multiple models and vote/weight results
- Empirical evaluation methods instead of theory — how well does it work on **unseen** data?
- Don’t calculate expected error, measure it from **unseen** data.

Statistical Models vs Machine Learning Models

Data

Statistical Models

- Usually small (< 1000 observations)
- Low dimension (< 10 variables)
- Can have detailed understanding of data
- Data is clean — human has looked at each data point

Models

- Simple models — complexity limited by theory
 - Detailed/complex statistical assumptions re data
 - Model known, and data is carefully examined to verify assumptions.
-
- Evaluation based on theoretical estimates under stated statistical assumptions
 - Analysis of errors using theoretical distributions

Validation

ML Models

- Can be huge (million+ observations)
- Large dimension (1000+, more for vision)
- Too large for human to parse / understand
- Data not clean — humans can't afford to understand/fix each point

- “No” upper limit on model complexity
 - Fewer statistical assumptions re data
 - Don’t know right model? No problem! have multiple models and vote/weight results
-
- Empirical evaluation methods instead of theory — how well does it work on **unseen** data?
 - Don’t calculate expected error, measure it from **unseen** data.

Statistics would be very different if it had been born after the computer instead of 100 years before

Statistical Models vs Machine Learning Models

Data

Statistical Models

- Usually small (< 1000 observations)
- Low dimension (< 10 variables)
- Can have detailed understanding of data
- Data is clean — human has looked at each data point

Models

- Simple models — complexity limited by theory
- Detailed/complex statistical assumptions re data
- Model known, and data is carefully examined to verify assumptions.

Validation

- Evaluation based on theoretical estimates under stated statistical assumptions
- Analysis of errors using theoretical distributions

Statistics would be very different if it had been born after the computer instead of 100 years before

ML Models

- Can be huge (million+ observations)
- Large dimension (1000+, more for vision)
- Too large for human to parse / understand
- Data not clean — humans can't afford to understand/fix each point

- “No” upper limit on model complexity
- Fewer statistical assumptions re data
- Don’t know right model? No problem! have multiple models and vote/weight results

- Empirical evaluation methods instead of theory — how well does it work on **unseen** data?
- Don’t calculate expected error, measure it from **unseen** data.

Splitting data into train+test(+validation) is vital

Outline

1. Machine Learning (ML) Overview	3
1.1. Components of a Machine Learning Problem	4
1.2. Problem–Task–Experience Perspective	9
1.3. Taxonomy of Machine Learning Methods	10
1.4. Statistical Models vs Machine Learning Models	12
2. Modelling Process	13
2.1. Models and error	15

The Pipeline Metaphor

Model Building Pipeline



Defining the Goal



Building the Model



Interpreting the Model



Preparing the Dataset for ML



Evaluating the Model

Source: Dataiku

The Pipeline Metaphor

Model Building Pipeline



Defining the Goal



Building the Model



Interpreting the Model



Preparing the Dataset for ML



Evaluating the Model

Source: Dataiku

Comments

- We saw the first two stages in previous weeks
- This week we look at the remaining stages
- Of course this pipeline is a simplification. In reality it is iterative.

What does a (supervised learning) model look like?

Definition 2 (Linear Model)

General form of linear model used in this module looks like

$$y_i \sim f_i^{(1)} + f_i^{(2)} + \dots + f_i^{(n)}$$

where y_i is the value of the response variable (target) for row (observation) i , and $f_i^{(j)}; j = 1, \dots, n$ is the value of the j^{th} feature for that observation.

In words: each target value y_i in the data is modelled as a linear combination of the model parameters a and the features, plus some error.

What does a (supervised learning) model look like?

Definition 2 (Linear Model)

General form of linear model used in this module looks like

$$y_i \sim f_i^{(1)} + f_i^{(2)} + \dots + f_i^{(n)}$$

where y_i is the value of the response variable (target) for row (observation) i , and $f_i^{(j)}; j = 1, \dots, n$ is the value of the j^{th} feature for that observation.

The model is linear in the sense that it can be turned into the following linear *equation*:

$$y_i = a_0 + a_1 f_i^{(1)} + a_2 f_i^{(2)} + \dots + a_n f_i^{(n)} + \varepsilon_i$$

In words: each target value y_i in the data is modelled as a linear combination of the model parameters a and the features, plus some error.

What does a (supervised learning) model look like?

Definition 2 (Linear Model)

General form of linear model used in this module looks like

$$y_i \sim f_i^{(1)} + f_i^{(2)} + \dots + f_i^{(n)}$$

where y_i is the value of the response variable (target) for row (observation) i , and $f_i^{(j)}; j = 1, \dots, n$ is the value of the j^{th} feature for that observation.

The model is linear in the sense that it can be turned into the following linear *equation*:

$$y_i = a_0 + a_1 f_i^{(1)} + a_2 f_i^{(2)} + \dots + a_n f_i^{(n)} + \varepsilon_i$$

In words: each target value y_i in the data is modelled as a linear combination of the model parameters a and the features, plus some error.

➤ Note that the features f can be nonlinear (such as $\sin x$) but the model parameters a must appear linearly.

What does a (supervised learning) model look like?

Definition 2 (Linear Model)

General form of linear model used in this module looks like

$$y_i \sim f_i^{(1)} + f_i^{(2)} + \dots + f_i^{(n)}$$

where y_i is the value of the response variable (target) for row (observation) i , and $f_i^{(j)}; j = 1, \dots, n$ is the value of the j^{th} feature for that observation.

The model is linear in the sense that it can be turned into the following linear *equation*:

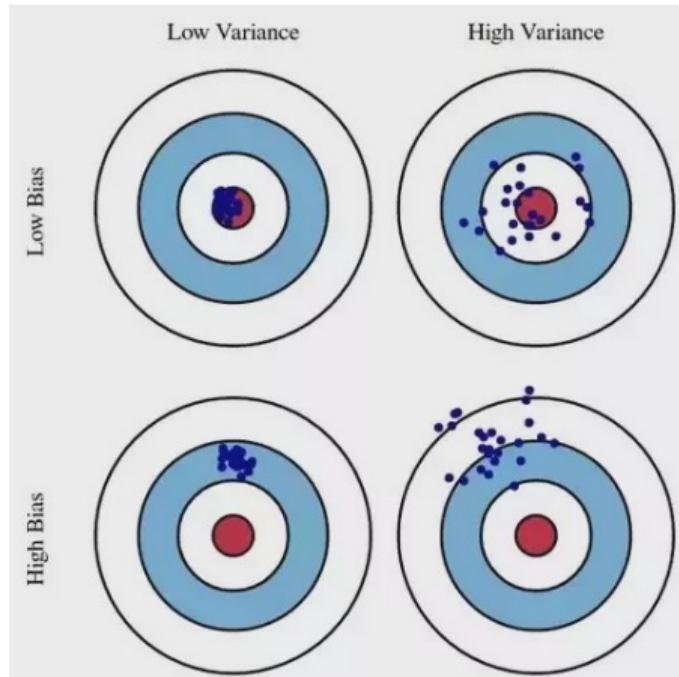
$$y_i = a_0 + a_1 f_i^{(1)} + a_2 f_i^{(2)} + \dots + a_n f_i^{(n)} + \varepsilon_i$$

In words: each target value y_i in the data is modelled as a linear combination of the model parameters a and the features, plus some error.

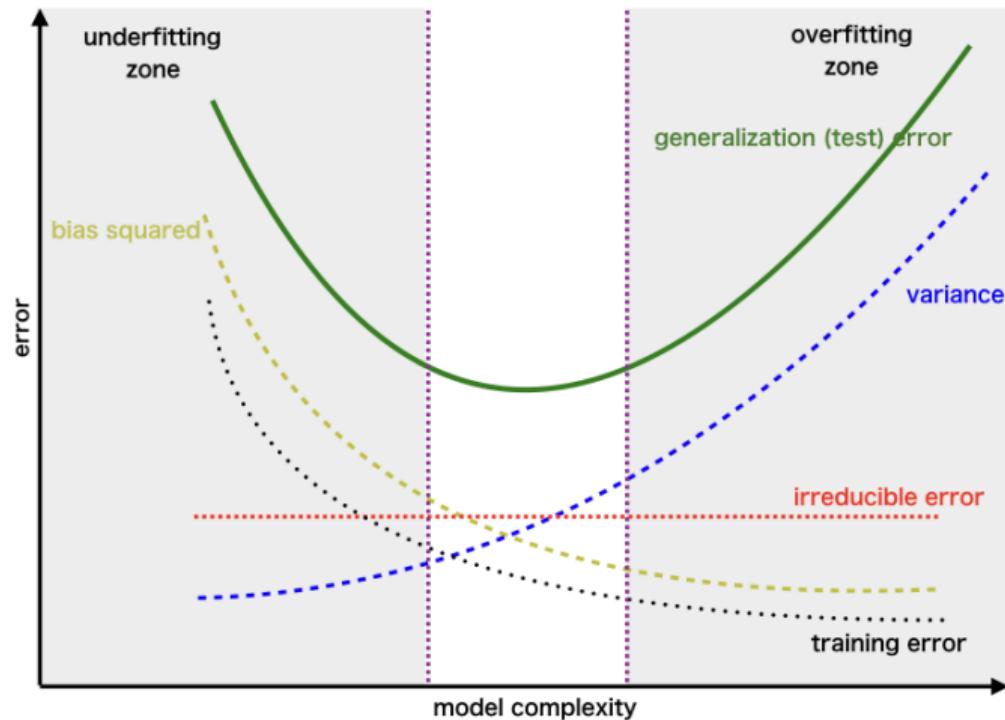
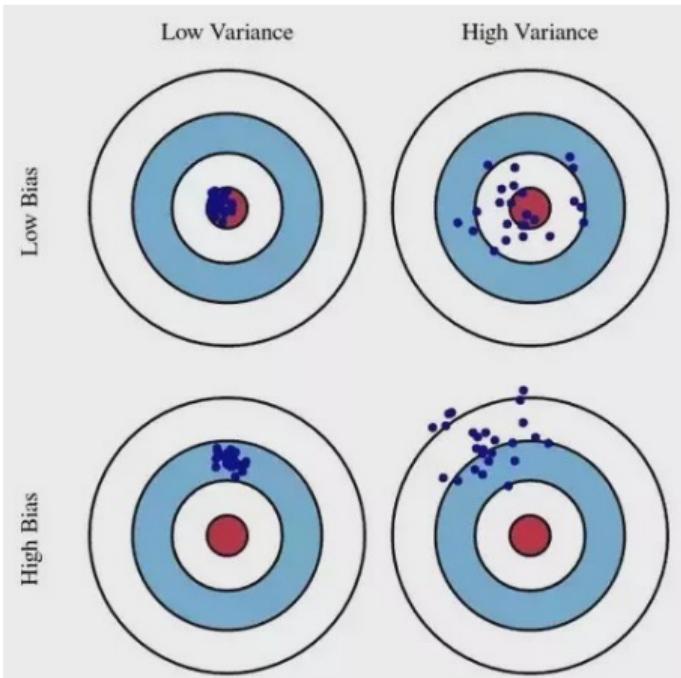
➤ Note that the features f can be nonlinear (such as $\sin x$) but the model parameters a must appear linearly.

➤ The goal of modelling is to find a so that the *prediction error* (loss function $\sim \|\varepsilon\|$) is a minimum.

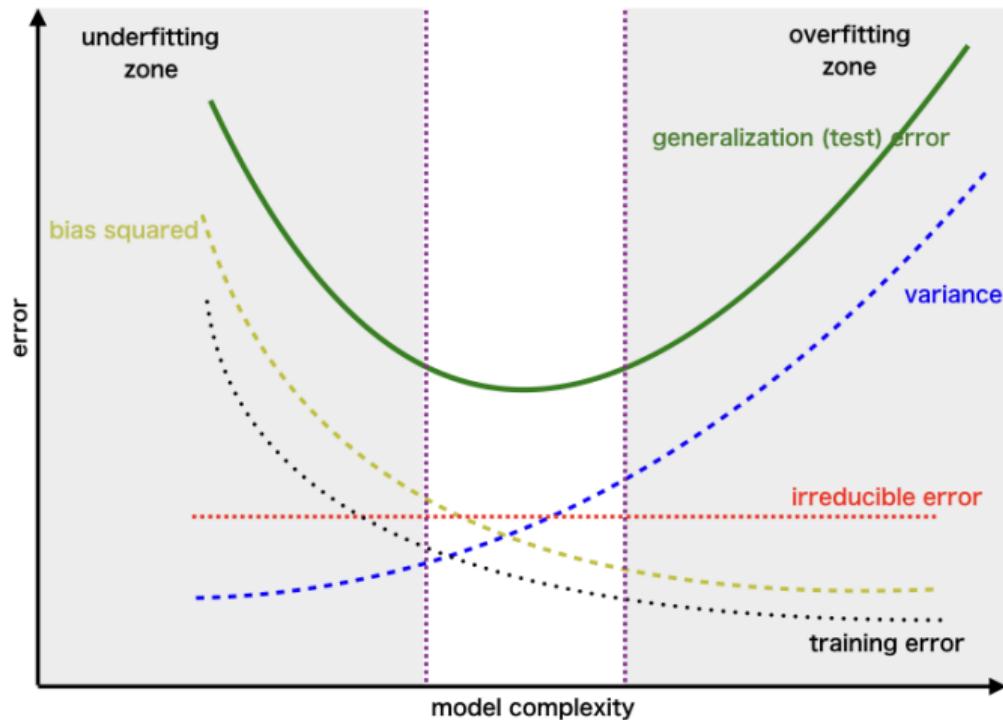
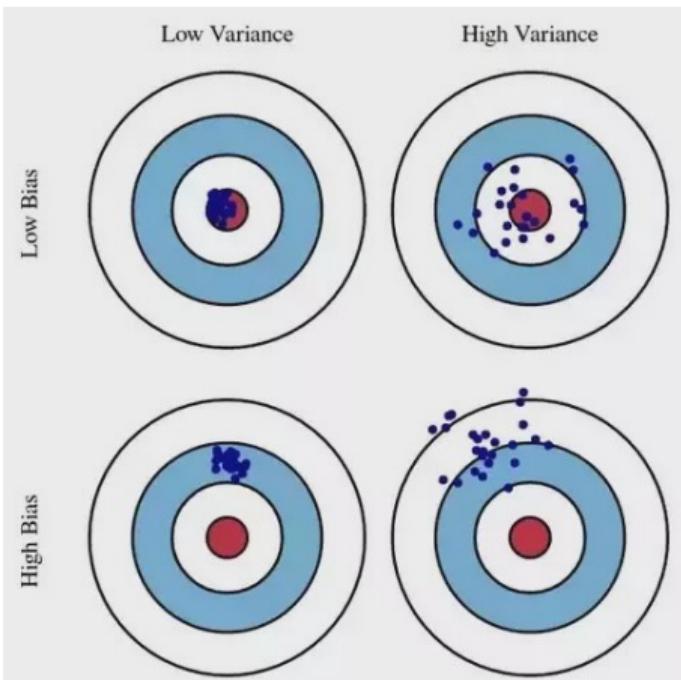
Bias-Variance and Total Error



Bias-Variance and Total Error

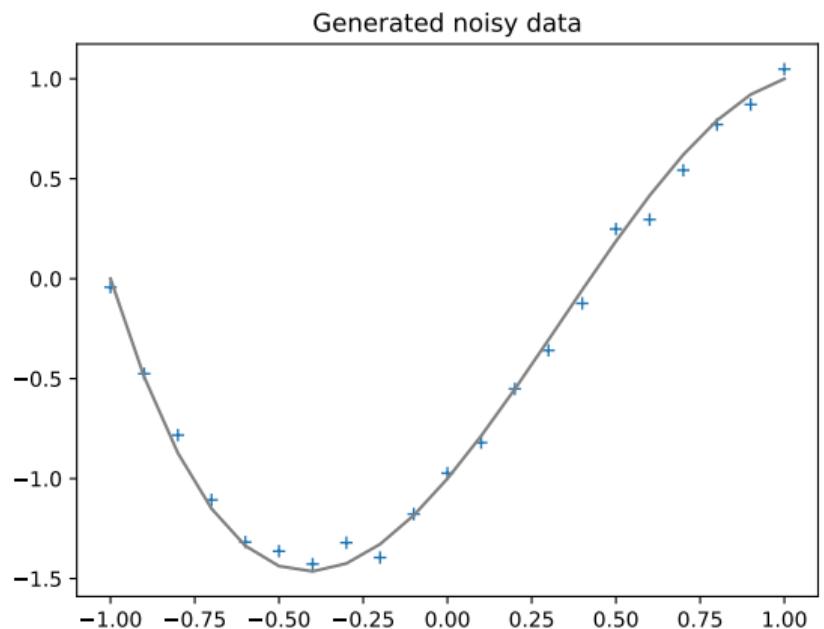


Bias-Variance and Total Error

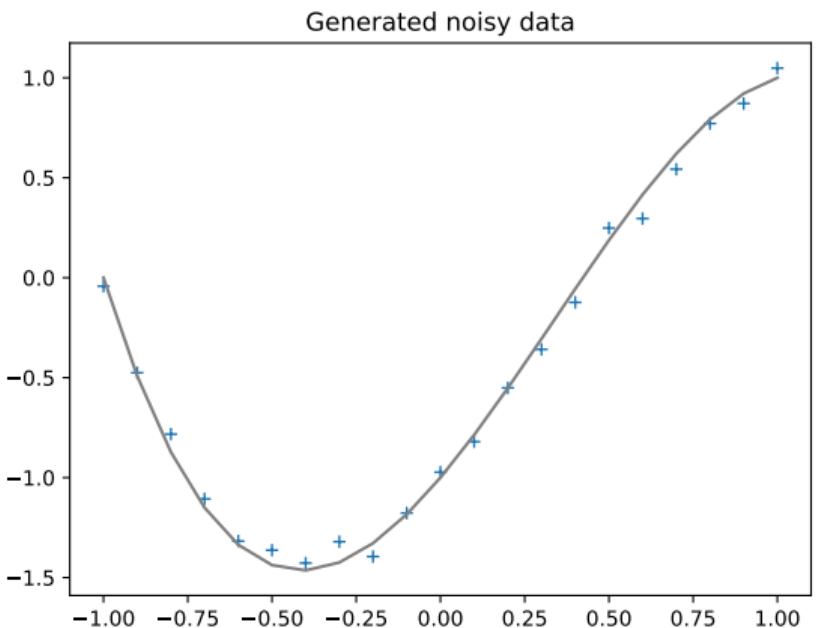


Look for parameters α that minimise the generalization error (estimated using the test set that was not used during training)

Example: Noisy data



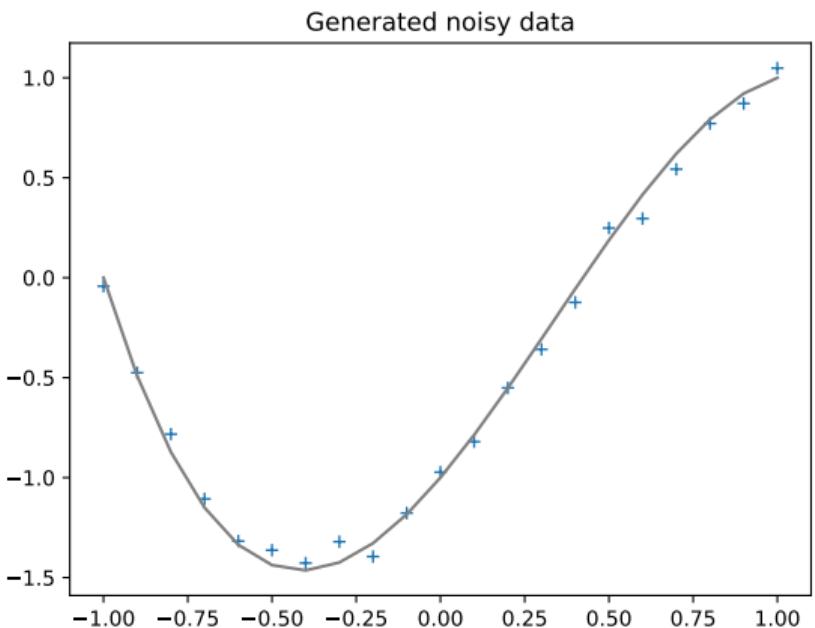
Example: Noisy data



Comments

- Given data with some error (noise)
- Expected underlying model is indicated by the grey curve

Example: Noisy data

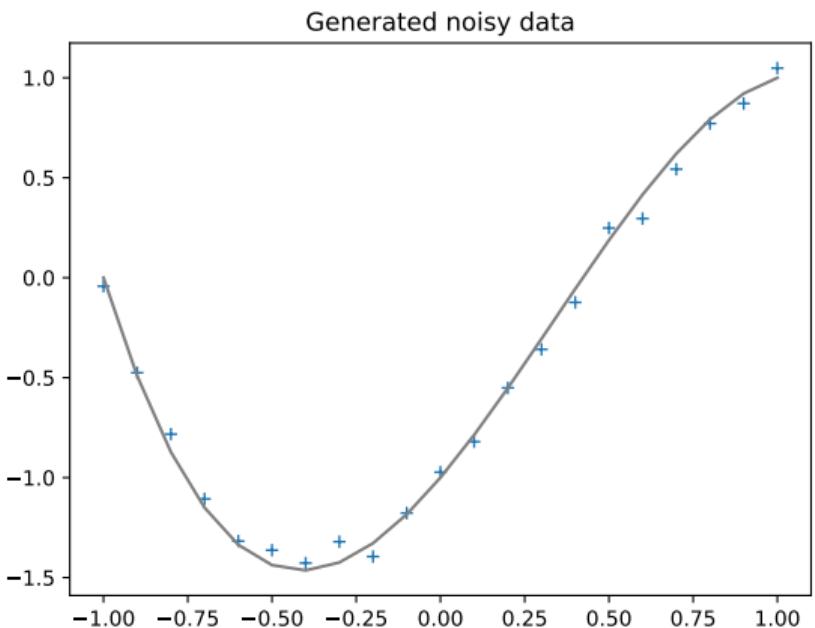


Comments

- Given data with some error (noise)
- Expected underlying model is indicated by the grey curve
- In the next slides we will compare different models, indicated by red curves

Look for the number of features that minimise the loss function

Example: Noisy data

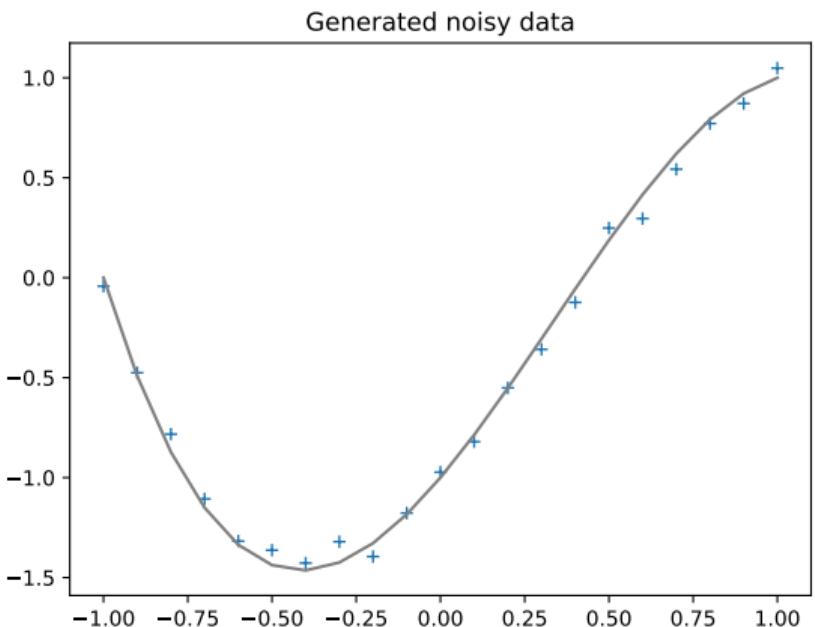


Comments

- Given data with some error (noise)
- Expected underlying model is indicated by the grey curve
- In the next slides we will compare different models, indicated by red curves
- The models have different numbers of *features*

Look for the number of features that minimise the loss function

Example: Noisy data

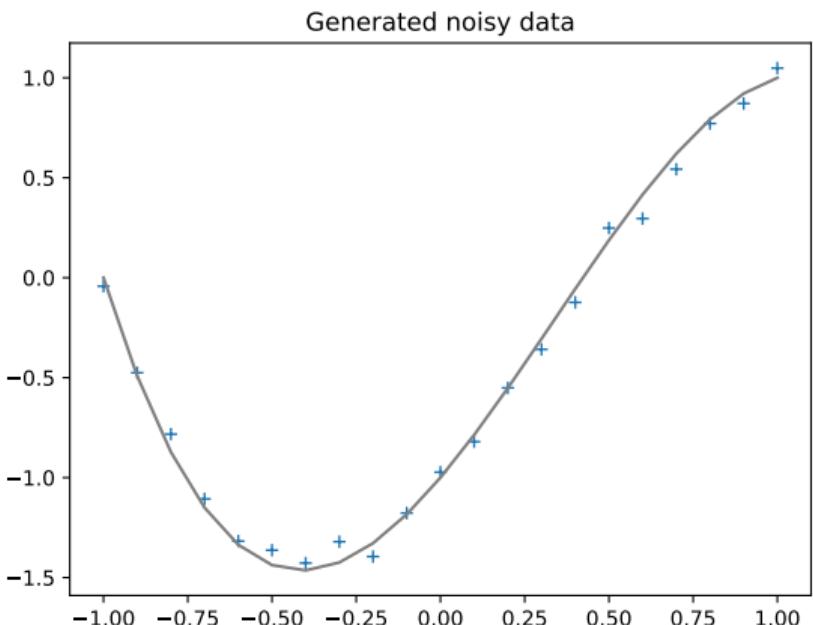


Comments

- Given data with some error (noise)
- Expected underlying model is indicated by the grey curve
- In the next slides we will compare different models, indicated by red curves
- The models have different numbers of *features*
- The values predicted by each model lie on the red curve

Look for the number of features that minimise the loss function

Example: Noisy data

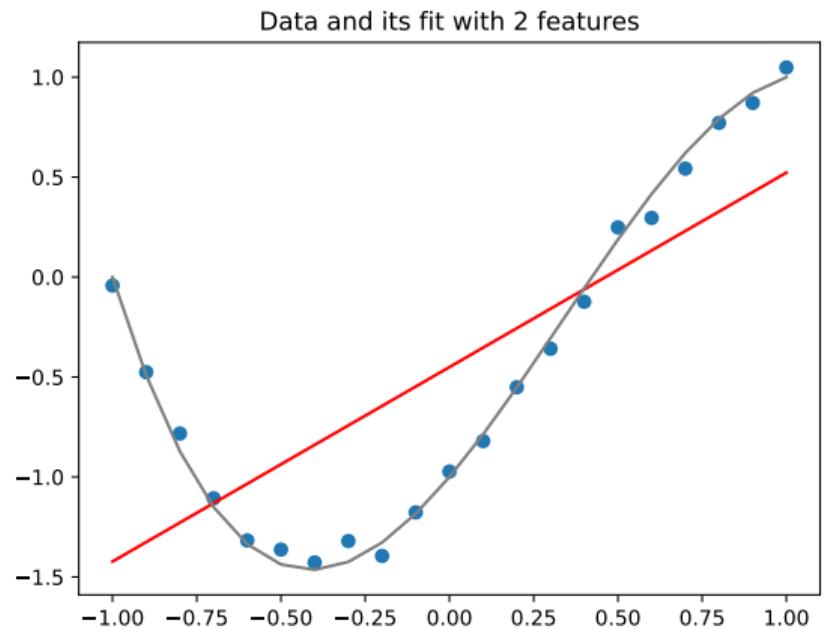


Comments

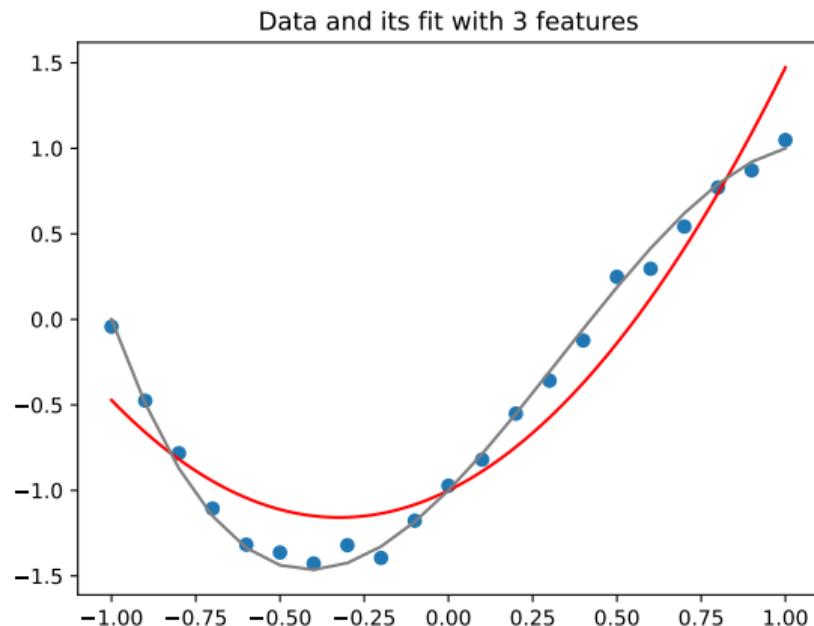
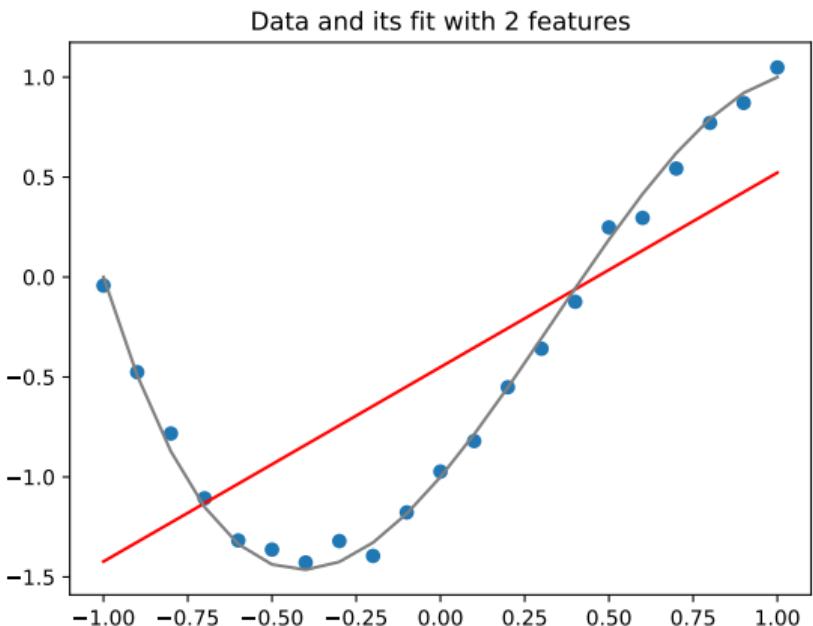
- Given data with some error (noise)
- Expected underlying model is indicated by the grey curve
- In the next slides we will compare different models, indicated by red curves
- The models have different numbers of *features*
- The values predicted by each model lie on the red curve
- The **loss function** is an estimate of how much the grey and red curves differ

Look for the number of features that minimise the loss function

High Bias, Low variance

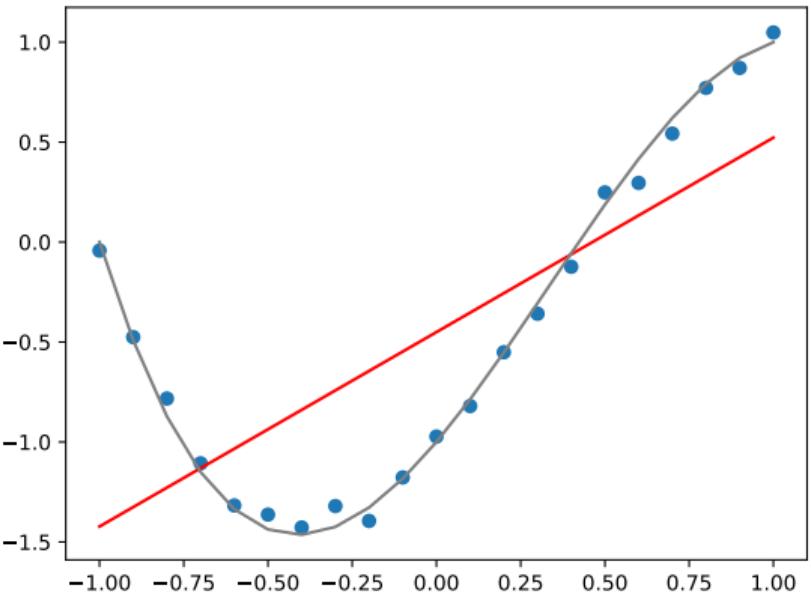


High Bias, Low variance

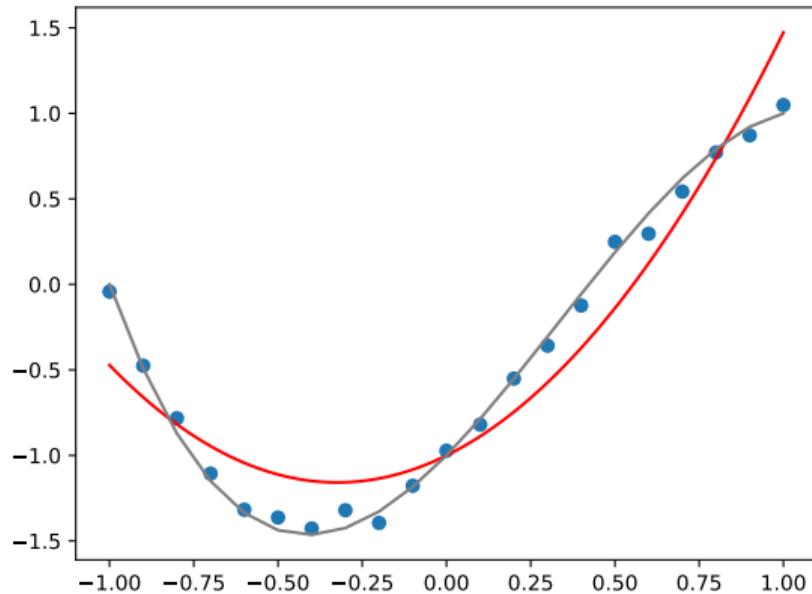


High Bias, Low variance

Data and its fit with 2 features

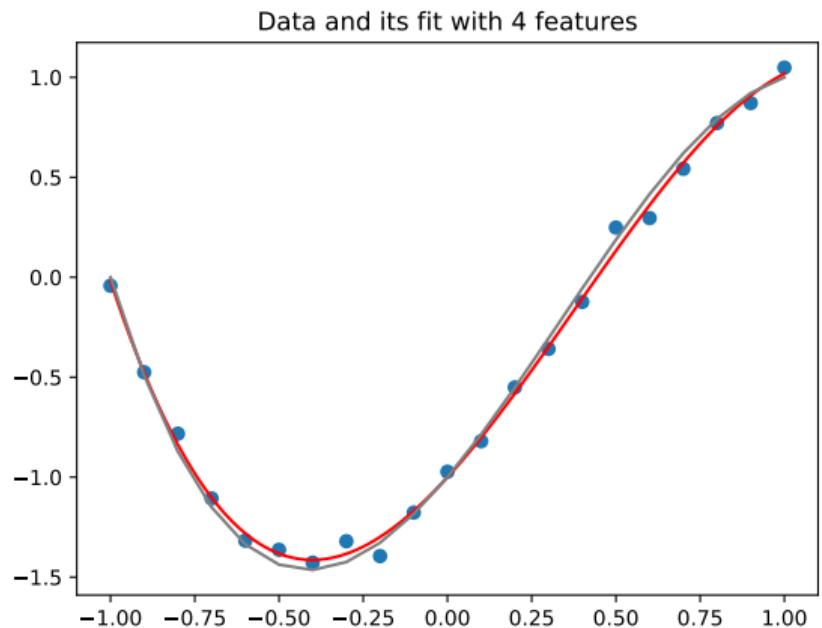


Data and its fit with 3 features

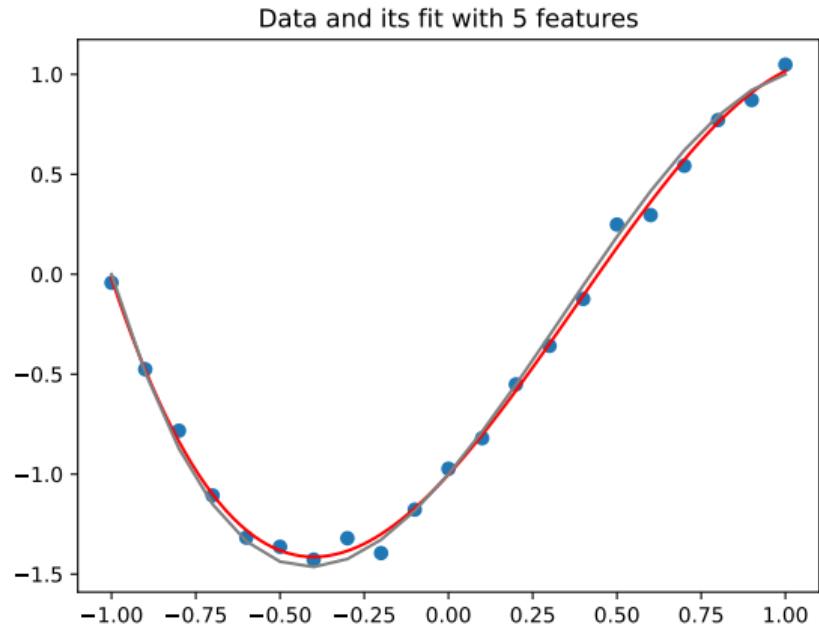
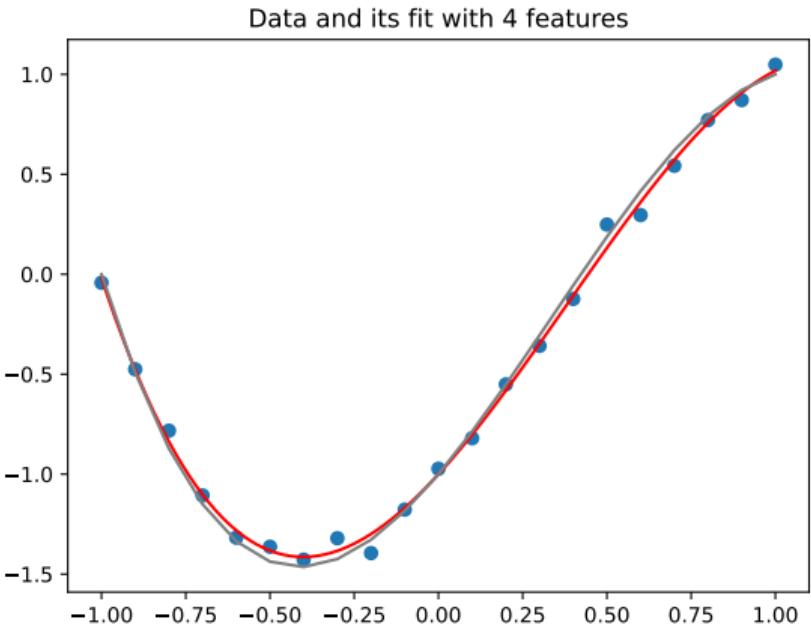


➤ Need more features: add more - but which ones?...

Low Bias, Low variance

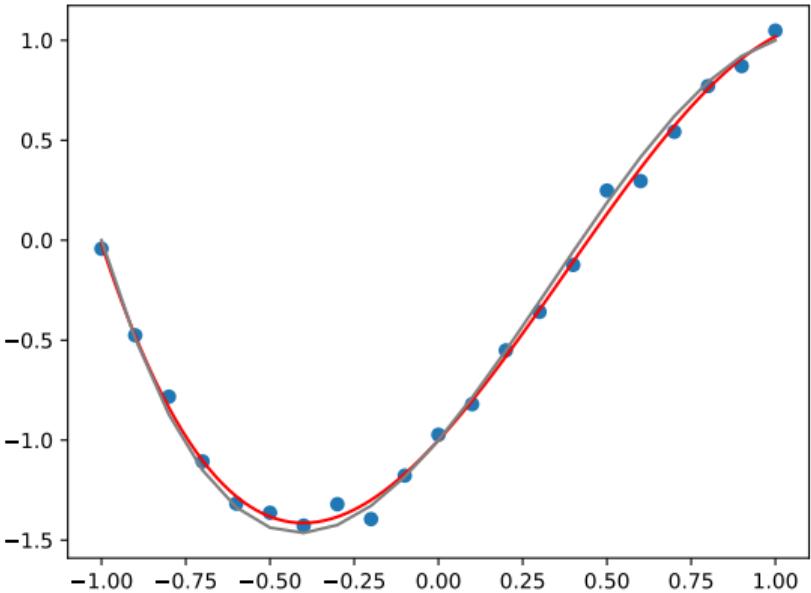


Low Bias, Low variance

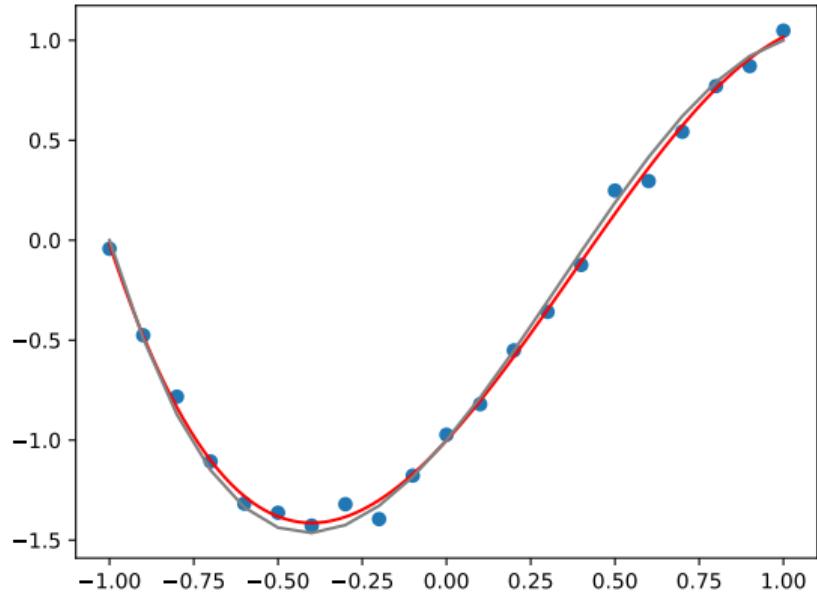


Low Bias, Low variance

Data and its fit with 4 features



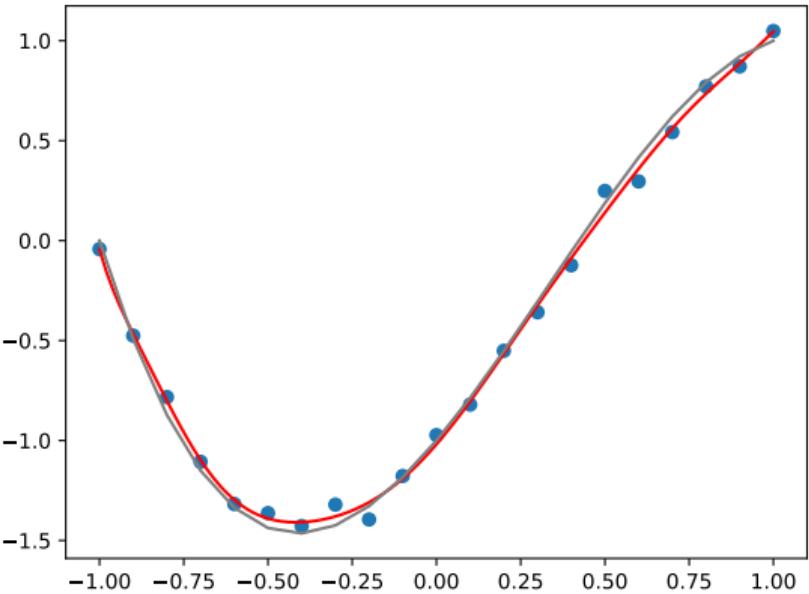
Data and its fit with 5 features



>About the right number of features...

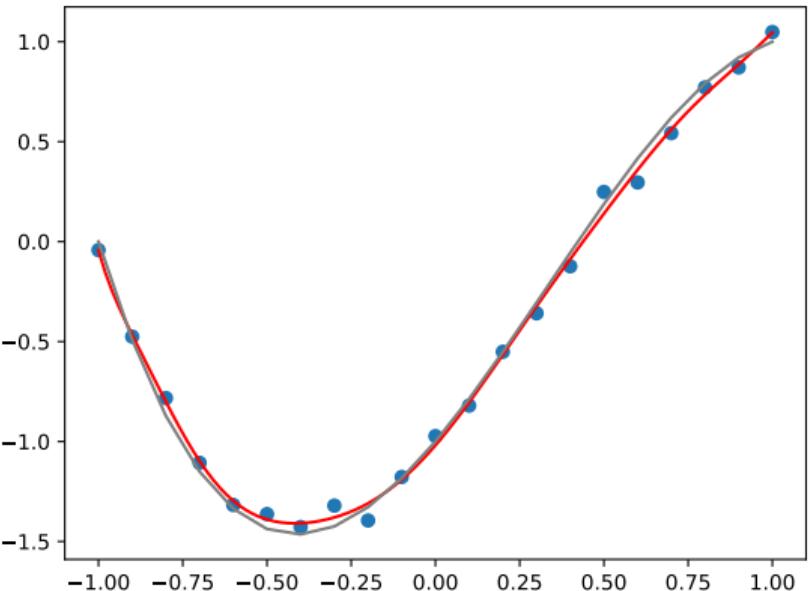
Low Bias, High variance

Data and its fit with 12 features

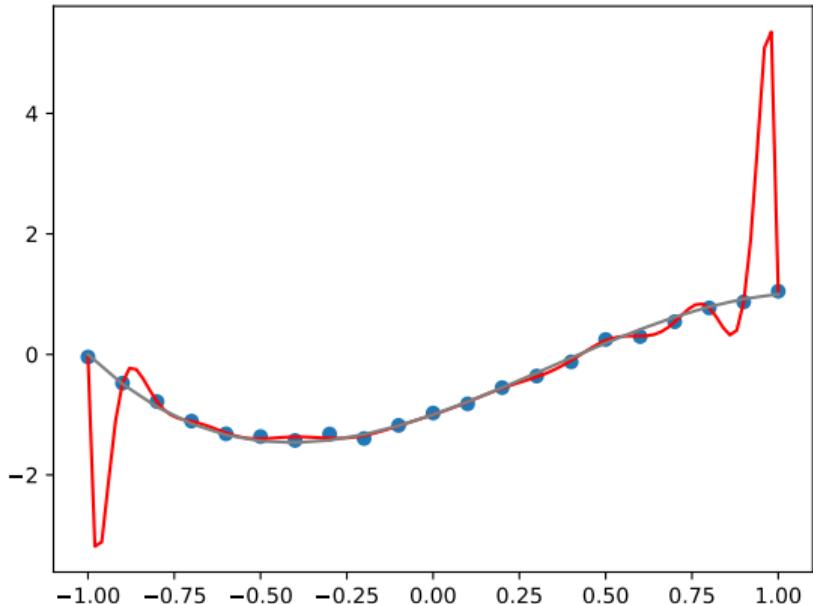


Low Bias, High variance

Data and its fit with 12 features

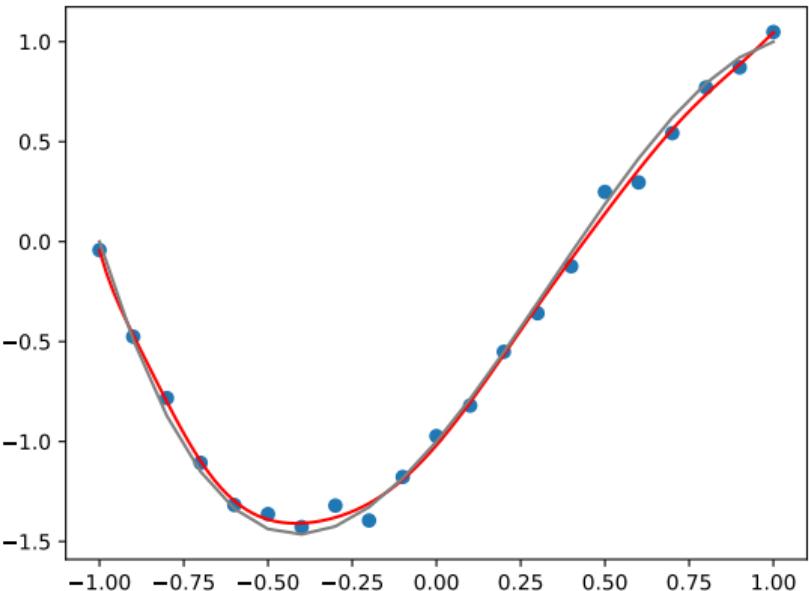


Data and its fit with 18 features

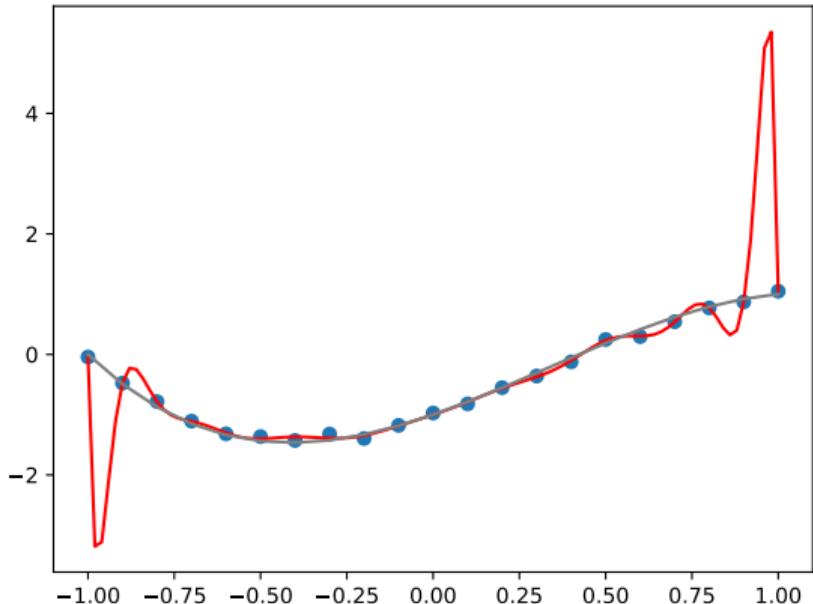


Low Bias, High variance

Data and its fit with 12 features



Data and its fit with 18 features



Too many features: remove some, but which ones?...

Example Model Types

Model	Applications	Concerns
Logistic Regression	X-ray classification	Regression with transformed variable
Fully connected networks	Classification	Classical ANN: choose encoding and size
Convolutional Neural Networks	Image processing	deep learning - choose segmentation
Recurrent Neural Networks	Voice recognition	ANN with feedback - how much?
Random Forest	Fraud Detection	Ensemble method - how many?
Reinforcement Learning	Learning by trial and error	Choose goal and penalties
Generative Models	Text, Image creation	Choose parameters
K-means	Segmentation	Choose distance function and k
k-Nearest Neighbors	Recommendation systems	Choose distance function and k
Bayesian Classifiers	Spam and noise filtering	Deal with imbalances