Hands on Machine Learning Chapter 1 - The Machine Learning Landscape

Python Book Clubs, Cohort 1
Sam Bryce-Smith - 29-03-2021

Chapter 1 - The Machine Learning Landscape

- What is Machine Learning?

- Types of Machine Learning algorithms

- Typical challenges with Machine Learning

What is Machine Learning?

- *“[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed”*
    - *Arthur Samuel, 1959*

- (We) program computers to learn from data and perform complex tasks

What is Machine Learning?

- *"[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed"*
  - *Arthur Samuel, 1959*

- E.g. **email spam filter**

✉️ ***Training instance/'sample'***
■ **Label - non-spam email**
■ **Label - spam email**

Training set

Spam emails flagged by users ✉️✉️✉️

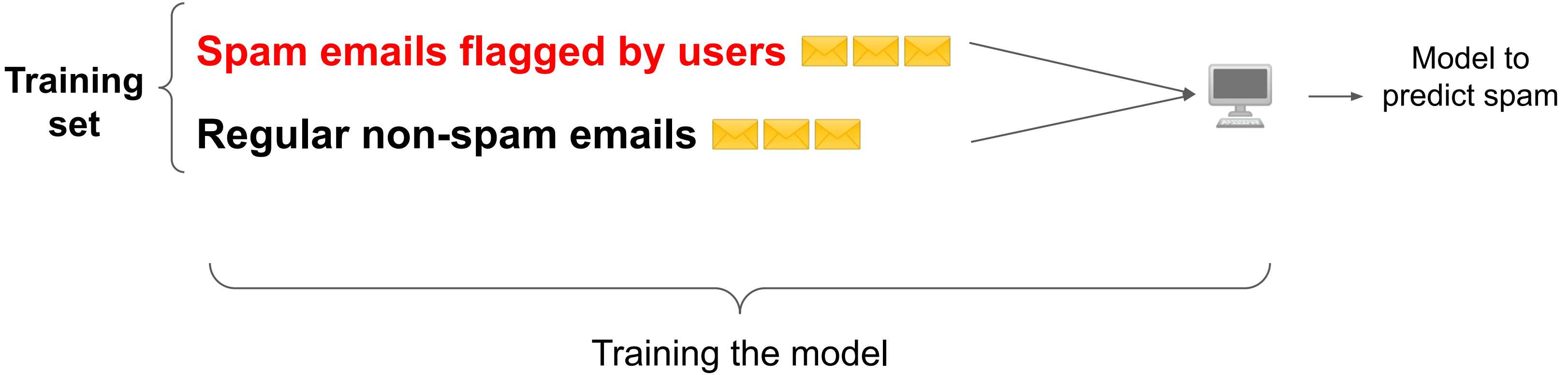
Regular non-spam emails ✉️✉️✉️

What is Machine Learning?

- *"[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed"*
  - *Arthur Samuel, 1959*

- E.g. **email spam filter**

✉ ***Training instance/'sample'***
■ **Label - non-spam email**
■ **Label - spam email**

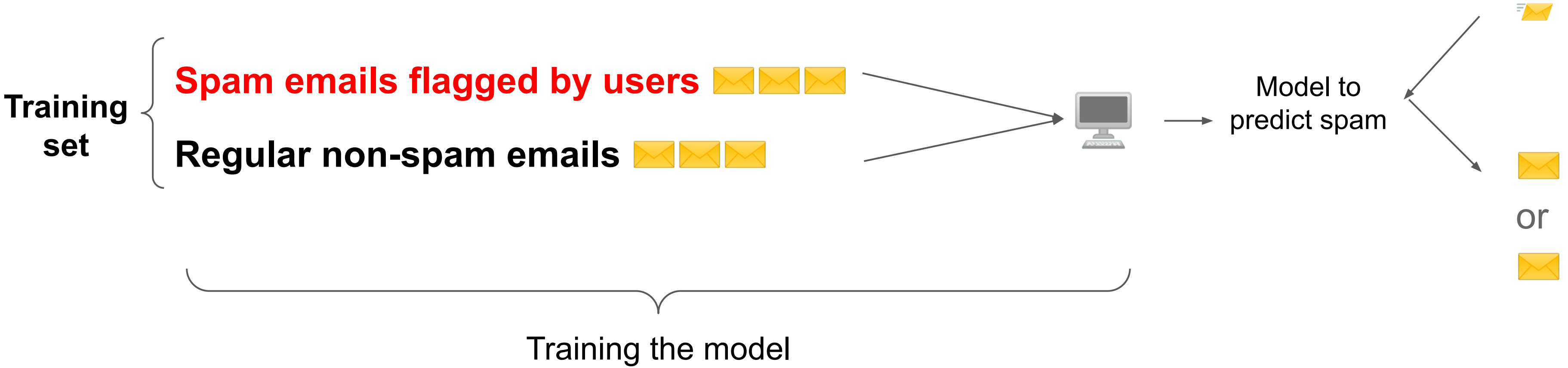


What is Machine Learning?

- *“[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed”*
    - *Arthur Samuel, 1959*

- E.g. **email spam filter**

✉ ***Training instance/’sample’***
■ **Label - non-spam email**
■ **Label - spam email**



Why use machine learning?

- Can adapt to new data (e.g. spammers trying new tricks)

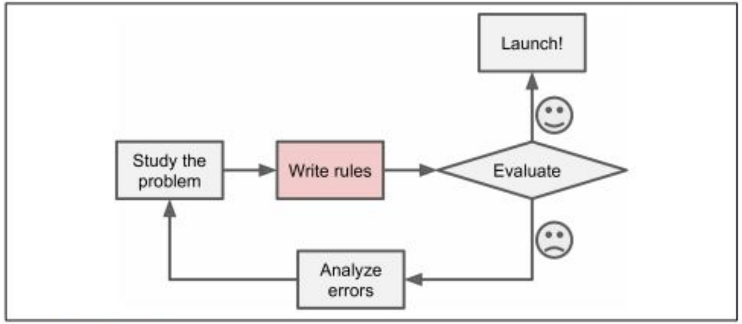


Figure 1-1. The traditional approach

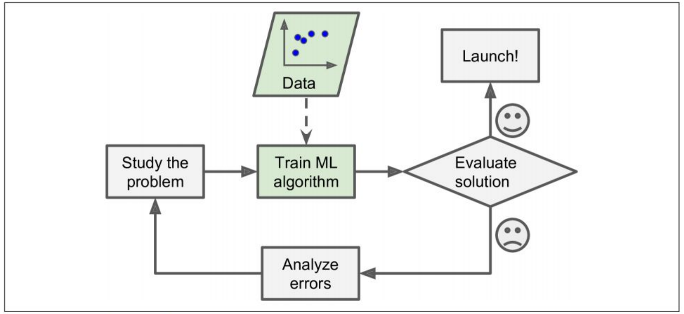


Figure 1-2. Machine Learning approach

Why use machine learning?

- Can adapt to new data (e.g. spammers trying new tricks)

- Find patterns in large datasets to help us learn (e.g. what words/phrases best predict a spam email?)

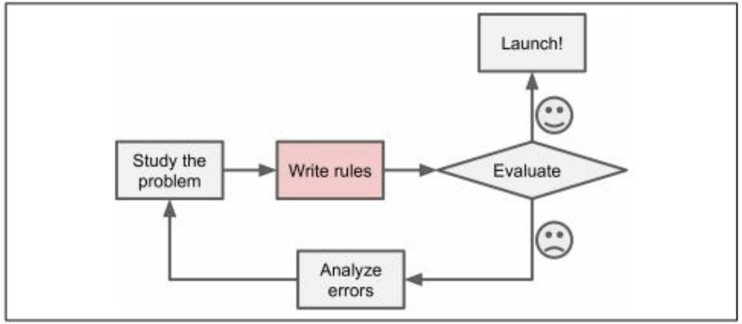


Figure 1-1. The traditional approach

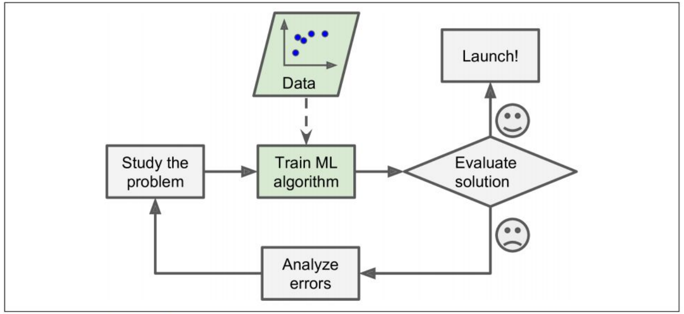


Figure 1-2. Machine Learning approach

Chapter 1 - The Machine Learning Landscape

- **What is Machine Learning?**
  - **Make computers perform tasks by learning from data**
  - **Find underlying patterns in large datasets to**
    - **Simplify complex problems & adapt to changing data**
    - **Help us understand complex systems/problems**

- Types of Machine Learning algorithms

- Typical challenges with Machine Learning

Chapter 1 - The Machine Learning Landscape

- What is Machine Learning?
    - Make computers perform tasks by learning from data
    - Find underlying patterns in large datasets to
        - Simplify complex problems & adapt to dynamic data
        - Help us understand complex systems/problems

- **Types of Machine Learning algorithms**

- Typical challenges with Machine Learning

Types of Machine Learning Systems

- How training data is provided
  - Labelled with desired solutions
- How the system trains
  - All data in one go
  - On the fly/as the data becomes available
- How it generalises to new datasets
  - Compare new data points to known data points
  - Detect patterns in training data and build a predictive model

Supervised/Unsupervised - is data labelled or not?

Supervised learning

- Training data contains desired solutions (spam/not spam) - ***labels***
- Classification (is new email spam or not?)
- Regression (predict a *target* numeric value given features)

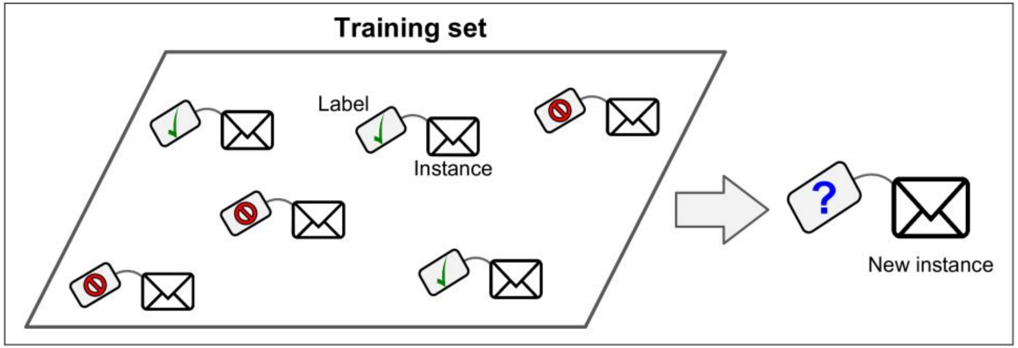


Figure 1-5. A labeled training set for supervised learning (e.g., spam classification)

Supervised/Unsupervised - is data labelled or not?

Unsupervised learning

- Training data does not contain desired solutions - ***unlabelled***
- Clustering (find groups of samples with similar features)
- Anomaly detection (which samples/instances look abnormal)

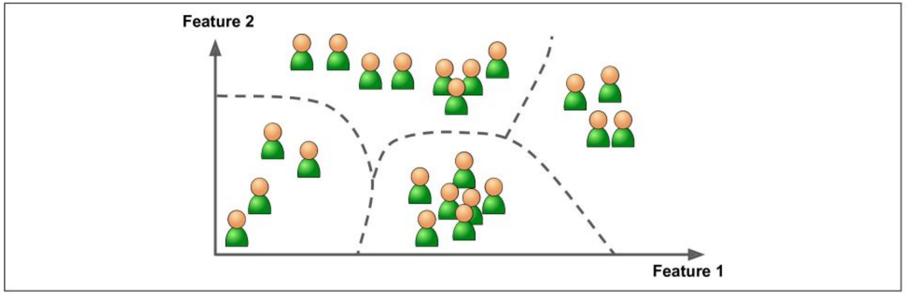


Figure 1-8. Clustering

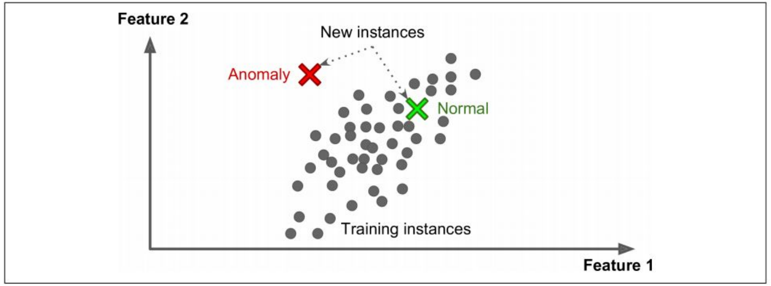


Figure 1-10. Anomaly detection

Supervised/Unsupervised - is data labelled or not?

Semi-supervised Learning

- Some instances in training data are labelled, some are not
- Combinations of supervised & unsupervised algorithms

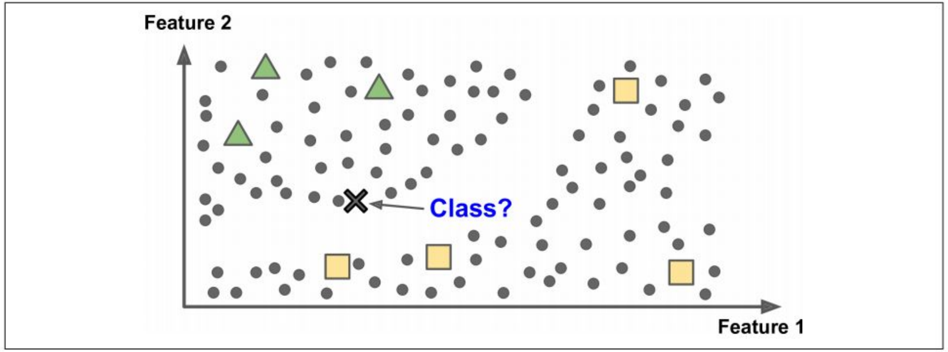


Figure 1-11. Semisupervised learning

Types of Machine Learning Systems

- **How training data is provided**
  - **Supervised - training data labelled with solutions you want**
  - **Unsupervised - training data is not labelled with solutions**

- How the system trains
  - All data in one go
  - On the fly/as the data becomes available

- How it generalises to new datasets
  - Compare new data points to known data points
  - Detect patterns in training data and build a predictive model

Types of Machine Learning Systems

- How training data is provided
  - Supervised - training data labelled with solutions you want
  - Unsupervised - training data is not labelled with solutions

- **How the system trains**
  - **All data in one go**
  - **On the fly/as the data becomes available**

- How it generalises to new datasets
  - Compare new data points to known data points
  - Detect patterns in training data and build a predictive model

Batch/Online Learning - learn continuously from flow of data or not?

Batch/offline learning

- Cannot learn instance by instance
- Train on all available data, then doesn’t learn from future incoming data

Batch/Online Learning - learn continuously from flow of data or not?

Batch/offline learning

- Cannot learn instance by instance
- Train on all available data, then doesn't learn from future incoming data

Incremental/Online learning

- Can learn from data one/ a few instances at a time
- ***Learning rate*** controls how fast a system incorporates new data into its model

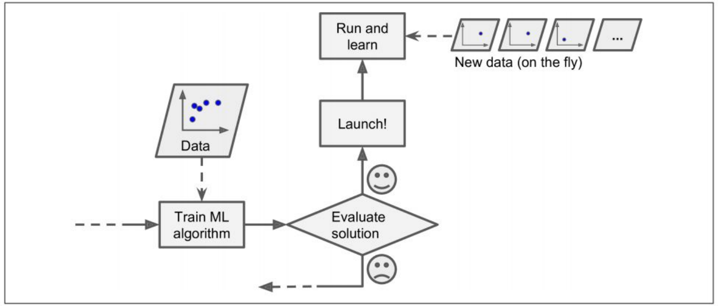


Figure 1-13. Online learning

Types of Machine Learning Systems

- How training data is provided
  - Supervised - training data labelled with solutions you want
  - Unsupervised - training data is not labelled with solutions

- **How the system trains**
  - **Batch/offline learning - learns from a single batch of data**
  - **Incremental/online learning - can learn continuously from new data**

- How it generalises to new datasets
  - Compare new data points to known data points
  - Detect patterns in training data and build a predictive model

Types of Machine Learning Systems

- How training data is provided
  - Supervised - training data labelled with solutions you want
  - Unsupervised - training data is not labelled with solutions

- How the system trains
  - Batch/offline learning - learns from a single batch of data
  - Incremental/online learning - can learn continuously from new data

- **How it generalises to new datasets**
  - **Compare new data points to known data points**
  - **Detect patterns in training data and build a predictive model**

Instance/Model based - compare to known points or build a predictive model?

Instance-based learning

- Learn by heart from training data
- Use a ***similarity measure*** to compare new data to known data points

Model-based learning

- Build a model of instances from training data
- Use model to make predictions on incoming new data

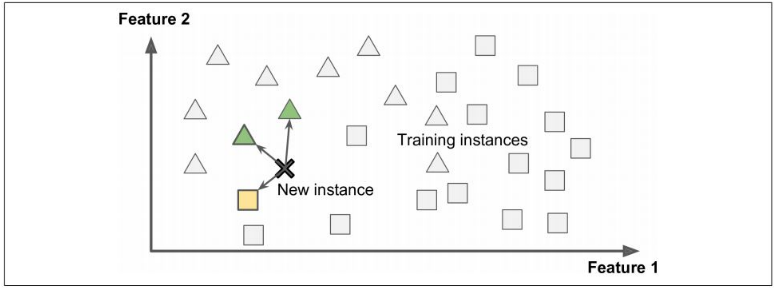


Figure 1-15. Instance-based learning

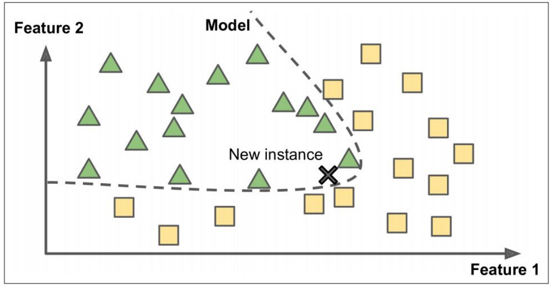


Figure 1-16. Model-based learning

Types of Machine Learning Systems

- **How training data is provided**
  - **Supervised - training data labelled with solutions you want**
  - **Unsupervised - training data is not labelled with solutions**

- **How the system trains**
  - **Batch/offline learning - learns from a single batch of data**
  - **Incremental/online learning - can learn continuously from new data**

- **How it generalises to new datasets**
  - **Instance-based - compare new data points to known data points**
  - **Model-based - detect patterns in training data and build a predictive model**

Chapter 1 - The Machine Learning Landscape

- What is Machine Learning?
  - Make computers perform tasks by learning from data
  - Find underlying patterns in large datasets to
    - Simplify complex problems & adapt to dynamic data
    - Help us understand complex systems/problems

- Types of Machine Learning algorithms
  - Training data labelled or not labelled
  - Model learns from batch of data or on the fly
  - Model compares to known data points or builds predictive model from training data

- **Typical challenges with Machine Learning**
  - Bad / insufficient quality data
  - Bad / insufficient performing algorithms

Bad input - Insufficient quantity & quality of training data

Not enough training data

- ML algorithms tend to need at least 1000s of training instances
- Algorithm choice still important for small datasets

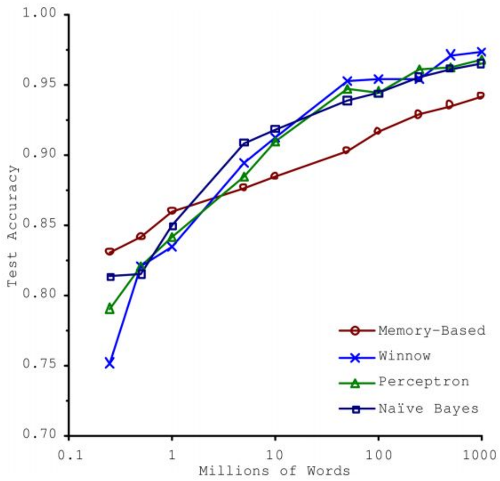


Figure 1-20. The importance of data versus algorithms[9]

Bad input - Insufficient quantity & quality of training data

Non-representative training data

- Can’t make accurate predictions if training data doesn’t reflect data want to generalise
- Sampling bias affects large samples sizes/training sets
  - I.e. flawed sampling methodology
- Sampling noise affects small sample sizes/training sets
  - I.e. Sample is unrepresentative due to random chance

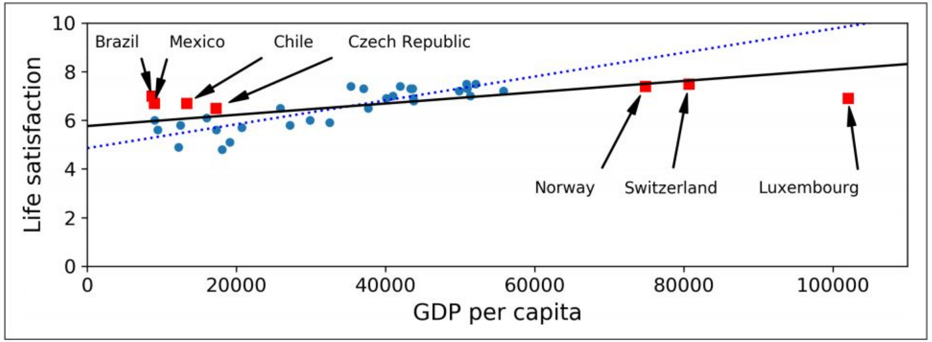


Figure 1-21. A more representative training sample

Bad input - Insufficient quantity & quality of training data

Outliers or noise in measurements

- Introduce unwanted variation - harder to detect underlying patterns in data
- Clear outliers should be removed

Bad input - Insufficient quantity & quality of training data

Outliers or noise in measurements

- Introduce unwanted variation - harder to detect underlying patterns in data
- Clear outliers should be removed

Missing values for some instances in training data

- Remove instances with missing values
- Fill in missing values (e.g. with median of training set)
- Train a model with feature containing missing values & one without

- (Hopefully these decisions will be discussed later in book)

Bad input - Insufficient quantity & quality of training data

Providing irrelevant features for the model to train on

- A model is only as good as the training data/features it is provided

- ***Feature engineering*** - finding good quality features to train your model on
    - *Feature selection* - pick out most useful features among features to train on
    - *Feature extraction* - creating new features from existing features
    - Gather more data & create new features

Chapter 1 - The Machine Learning Landscape

- What is Machine Learning?
    - Make computers perform tasks by learning from data
    - Find underlying patterns in large datasets to
        - Simplify complex problems & adapt to dynamic data
        - Help us understand complex systems/problems

- Types of Machine Learning algorithms
    - Training data labelled or not labelled
    - Model learns from batch of data or on the fly
    - Model compares to known data points or builds predictive model from training data

- **Typical challenges with Machine Learning**
    - Not enough training data
    - Outliers, missing features in training data
    - Irrelevant features (don’t help to find underlying patterns in data)
    - **Bad / insufficiently performing algorithms**

Bad ‘model behaviour’ - Overfitting training data

- Model fits the training data (too) well but does not generalise to new data
- ***“Overfitting happens when the model is too complex relative to the amount & robustness of the training data”***

- ***Regularisation*** - constrain/limit values of a parameter in model
    - **Hyperparameter** - how much algorithm applies regularisation when training

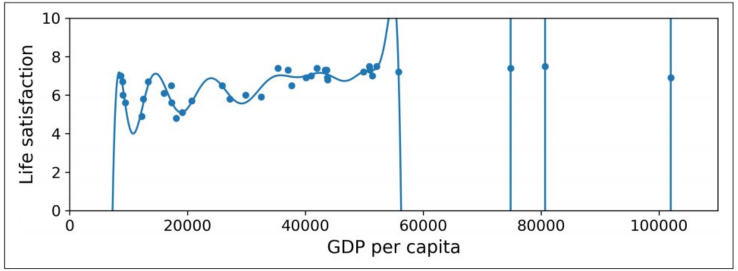


Figure 1-22. Overfitting the training data

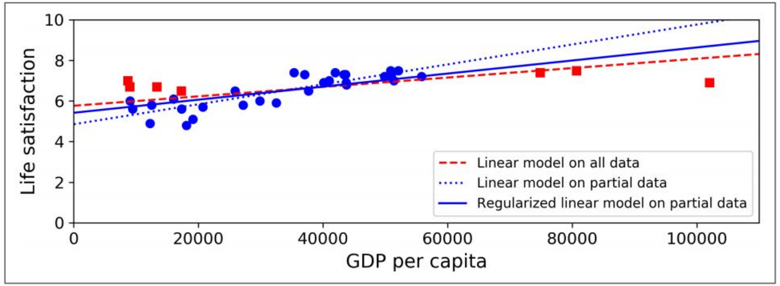


Figure 1-23. Regularization reduces the risk of overfitting

Bad ‘model behaviour’ - Underfitting training data

- Model too simple to identify underlying patterns in the data
  - Pick a more complex model (more parameters)
  - Find better features for the model (feature engineering)
  - Reduce the regularisation hyperparameter (more flexibility in values a parameter can take)

Bad ‘model behaviour’ - Underfitting training data

- Model too simple to identify underlying patterns in the data
  - Pick a more complex model (more parameters)
  - Find better features for the model (feature engineering)
  - Reduce the regularisation hyperparameter (more flexibility in values a parameter can take)

- How to evaluate performance of a model?

Splitting your input data into training and test sets aids evaluation before using model on new data

Evaluating performance before running new data through model

- Split input data into *training* and *test* sets
  - *Training error* - error rate on cases in training set
  - *Generalisation error* - error rate on ‘new’ cases in test set

Splitting your input data into training and test sets aids evaluation before using model on new data

Evaluating performance before running new data through model

- Split input data into *training* and *test* sets
  - *Training error -* error rate on cases in training set
  - *Generalisation error* - error rate on ‘new’ cases in test set

Choosing between different models (and hyperparameter values)

- ***Holdout validation*** - set aside part of training data for evaluation (*validation set, dev set)*
  - Train all potential models on reduced training data
  - Select model that performs best on validation set

Splitting your input data into training and test sets aids evaluation before using model on new data

Repetition is key

- One split = have adapted model for one particular training & test set

- ***Cross-validation*** *-* repeat training & testing with multiple validation sets
  - Pick model with best average performance

Chapter 1 - The Machine Learning Landscape

- What is Machine Learning?
  - Make computers perform tasks by learning from data
  - Find underlying patterns in large datasets to
    - Simplify complex problems & adapt to dynamic data
    - Help us understand complex systems/problems

- Types of Machine Learning algorithms
  - Training data labelled or not labelled
  - Model learns from batch of data or on the fly
  - Model compares to known data points or builds predictive model from training data

- **Typical challenges with Machine Learning**
  - Not enough and/or poor quality training data
  - Overfitting training data - (poor predictions on new data)
  - Underfitting training data - (poor predictions of both training & new data)
  - Not enough training data
  - Outliers, missing features in training data
  - Irrelevant features (don’t help to find underlying patterns in data)

The typical workflow of a ML project (Chapter 2 👀)

1. **Gather training data for your task of choice & select an algorithm**

2. **Train & evaluate algorithm on training data**

3. **Apply model to new data (to make inferences)**

- Thoughts on chapter?
- How to split workload for next few sessions?