Machine Learning Landscape

Michael Claudius, Associate Professor, Roskilde

27.08.2021

What is Machine Learning

- ***Machine Learning is the science (and art) of programming computers so they can learn from data.***

> ***Here is a slightly more general definition****:*
> Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed. *(Arthur Samuel, 1959)*
>
> *And a more engineering-oriented one*:
> A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E. *(Tom Mitchell, 1997)*

- **Example Spam filter!: Task is to flag for spam.**
- **Experience is the training data (E-mails).**
- **Performance is the accuracy. E.G. 95% flagged correct.**

Machine Learning vs. Traditional Programming

- ***Traditional programming***
  - ***Write algorithm: Set up rules for words (For U, credit card, free)***
  - ***Many rules => Complex algorithm***
  - ***Spammer can work around rules (4U) => end-less number of rules***

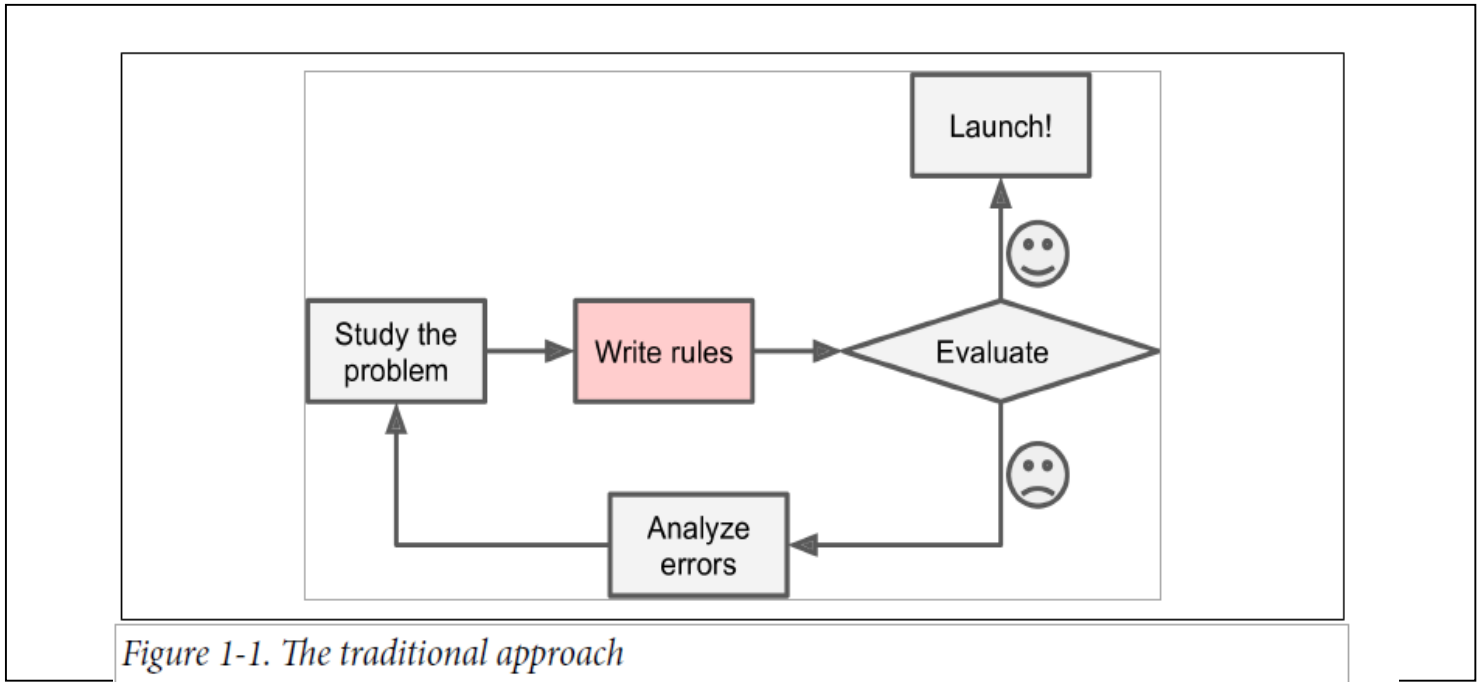


Figure 1-1. The traditional approach

Machine Learning vs. Traditional Programming

- ***Why Machine Learning***
  - ***Learns automatically (from users flagging e-mails w; e.g. For U)***
  - ***Short and easy to maintain***

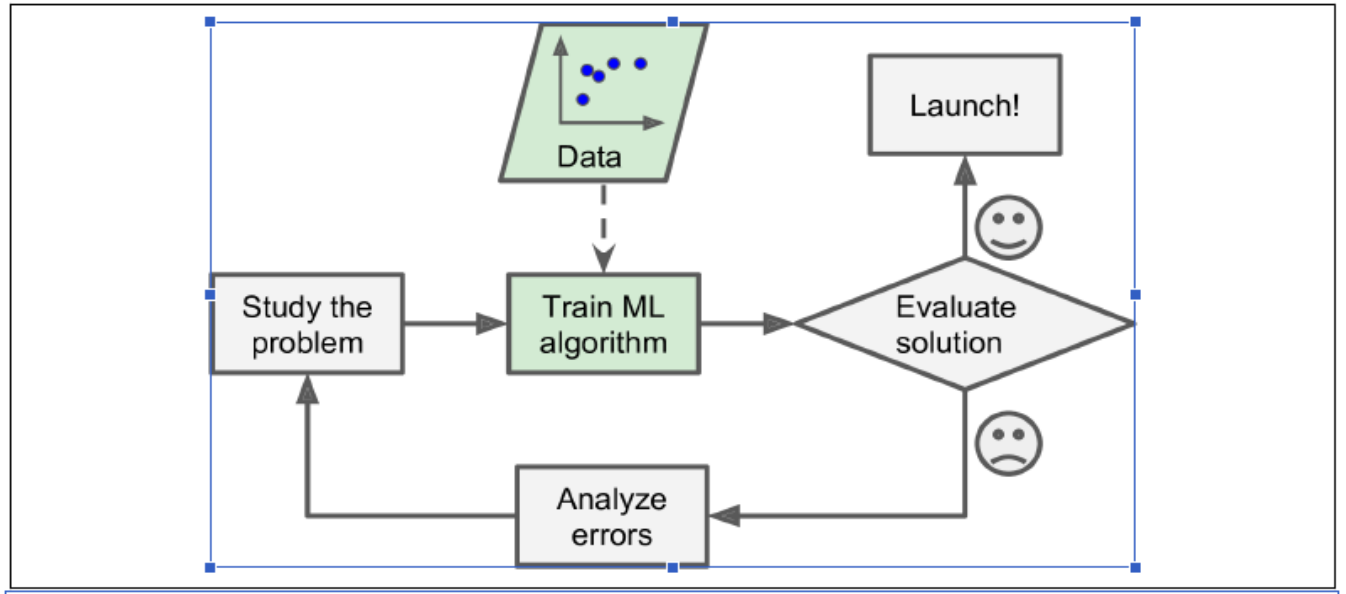


Figure 1-2. Machine Learning approach

When Machine Learning

- ***When to do Machine Learning***
  - ***Complex algorithms***
  - ***No known Algorithm (e.g. speech recognition)***
  - ***Huge data. Data mining discover disguised patterns***
  - ***Help humans to detect new trends or correlations***

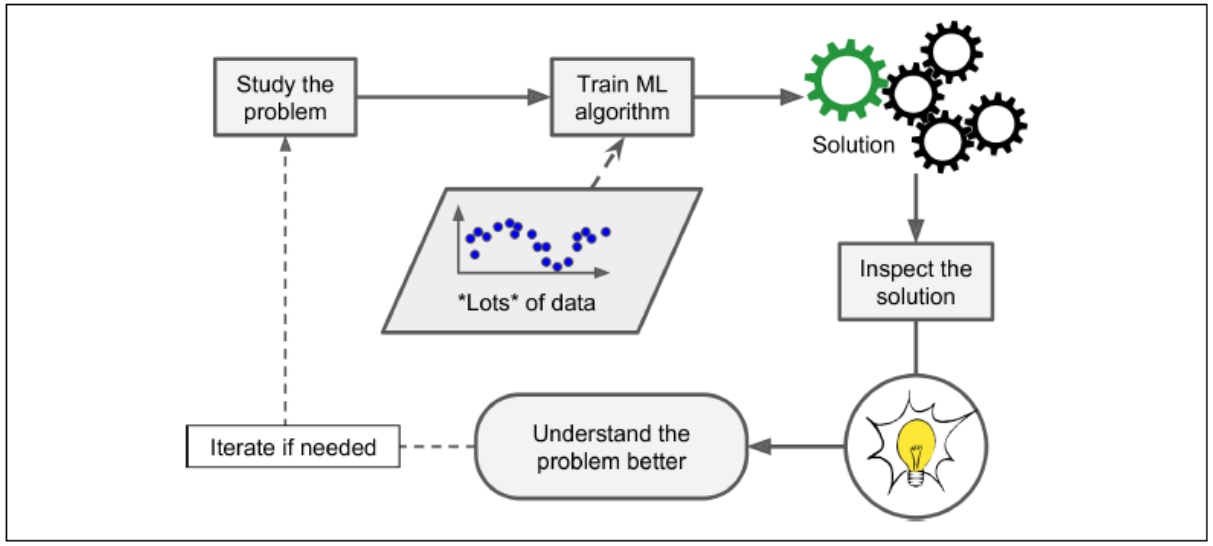


Figure 1-4. Machine Learning can help humans learn

Examples of Machine Learning

- ***Now discuss and find more areas where ML can be used !***
- ***Students answers:***
- ***Spam filters***

Examples of Machine Learning

- ***Now discuss and find more areas where ML can be used !***
- ***Students answers:***
- ***Spam***
- ***Traffic bottlenecks***
- ***Adverts based on customer patterns***
- ***Supermarkets shopping pattern***
- ***Amazon***
- ***Tesla self drivng cars***
- ***Speech recognition***
- ***Surveillance of persons***
- ***Gaming AI-cheating optimizing the shooting***
- ***Google-photos: face pictures***
- ***Diagnose of illness***
- ***YouTube choice selection***
- ***Searching algorithms***
- ***Robots, reinforcement learning***
- ***Statistical data, Clustering by K-means***

Types of Machine Learning Systems

- ***Human supervision or NOT***
  - ***Supervised***
  - ***Unsupervised***
  - ***Semisupervised***
  - ***Reinforcement learning***

- ***Online training (learn on the fly) vs. batch (NOT online)***

- ***Instance based (compare new data with old data) vs. Model based (build a predicative model)***

ML Supervised

- ***Human supervision***
- ***Features have a label.***
- ***Classification (Spam or not spam)***
- ***(Linear) Regression (Car with the label price). Features of a car ??***

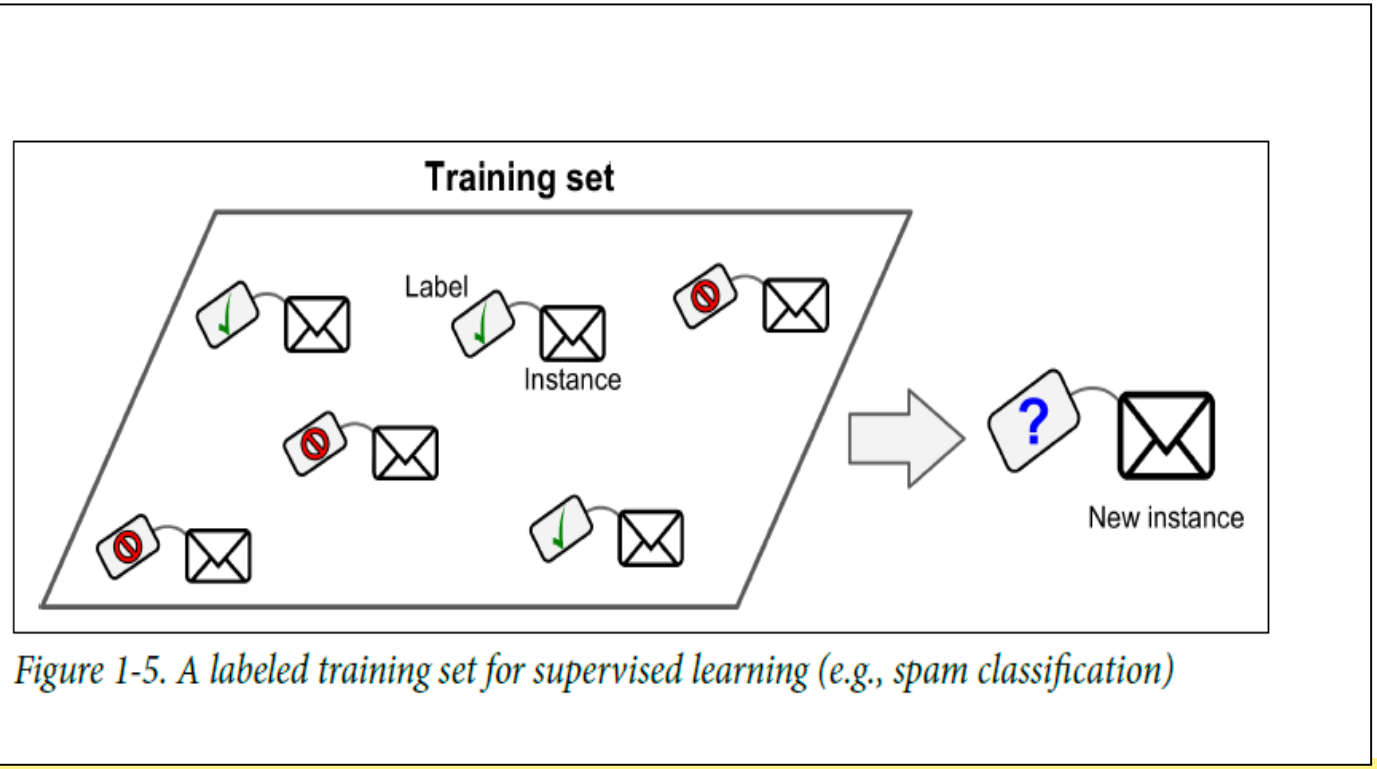


Figure 1-5. A labeled training set for supervised learning (e.g., spam classification)

ML UnSupervised

- ***No Human supervision***
- ***No labels.***
- ***Clustering (similar visitors to your blog/supermarket)***
- ***Not like Regression (Car with the label price)***

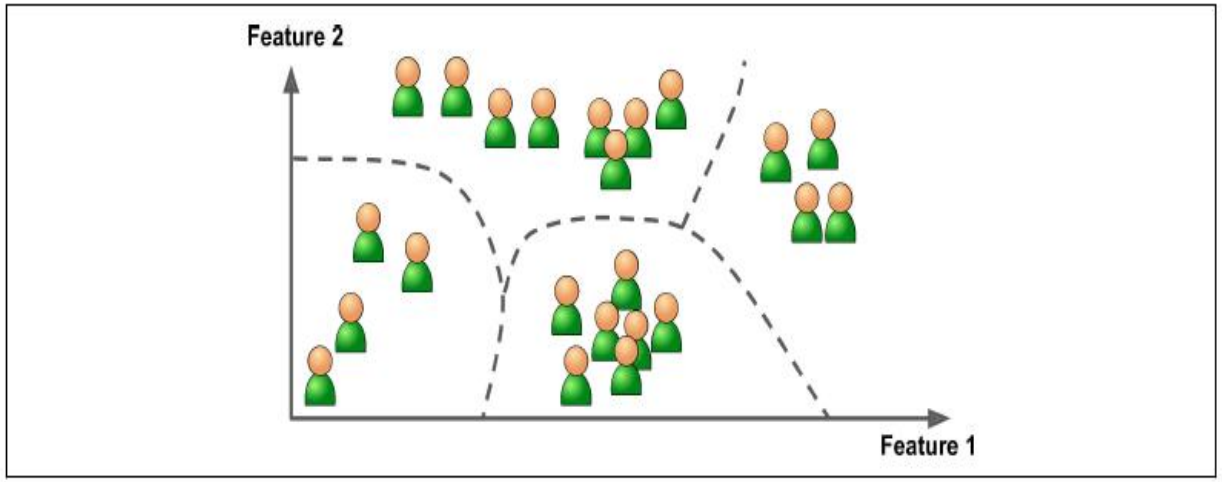


Figure 1-8. Clustering

ML UnSupervised Annomaly detection

- ***Credit transactions***
- ***Mobile call patterns***

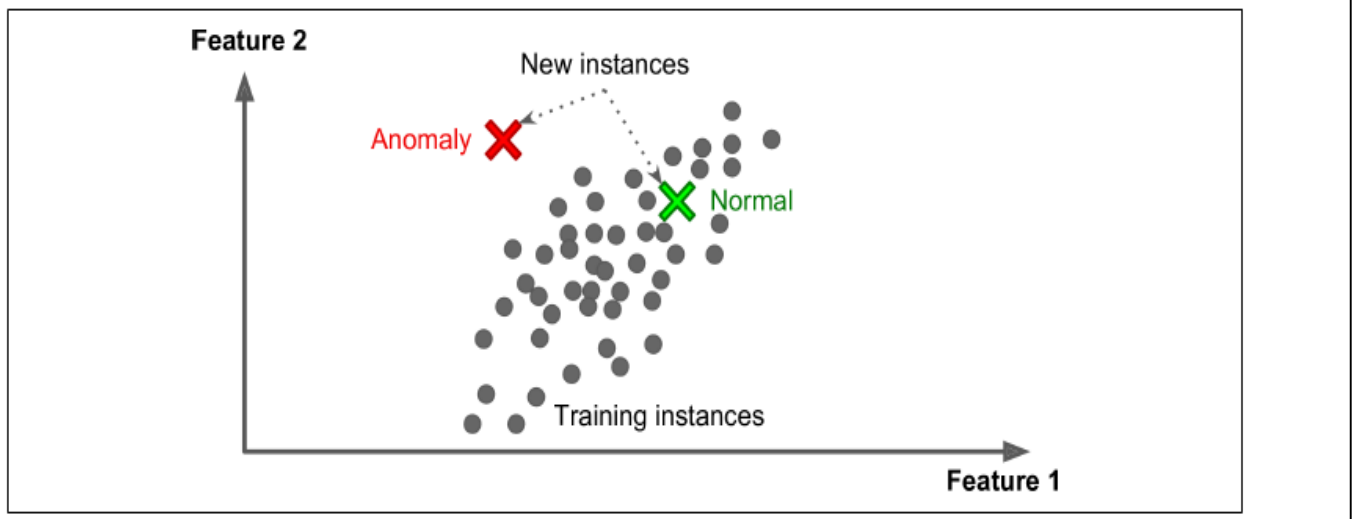


Figure 1-10. Anomaly detection

ML Semisupervised

- ***Some (few) features have labels some NOT***
- ***Mixing learning from data with labels***
- ***Google photos !***

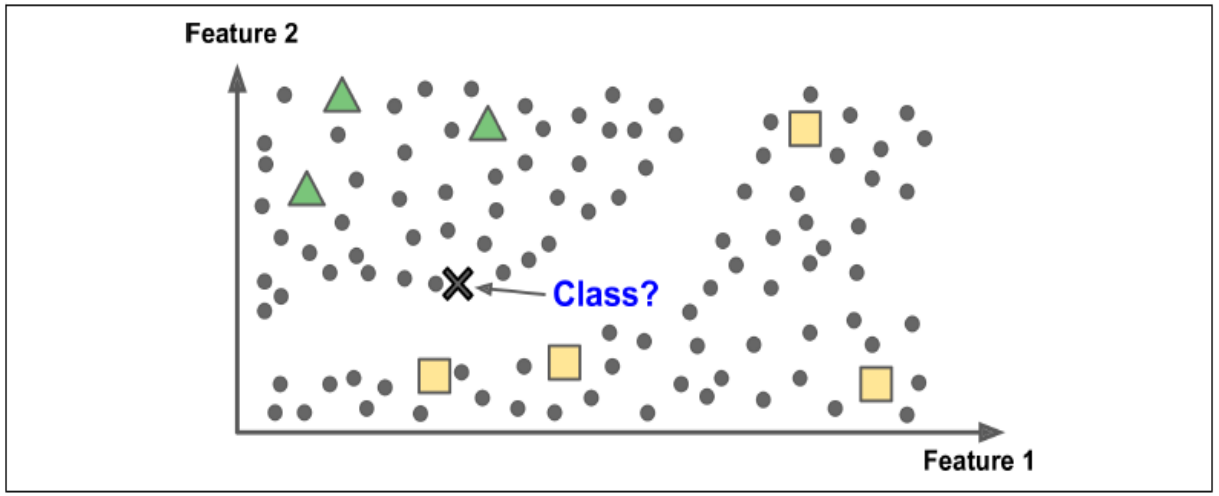


Figure 1-11. Semisupervised learning

Reinforcement learning

- ***Agent observes environment***
- ***Takes action***
- ***Reward or punishment***

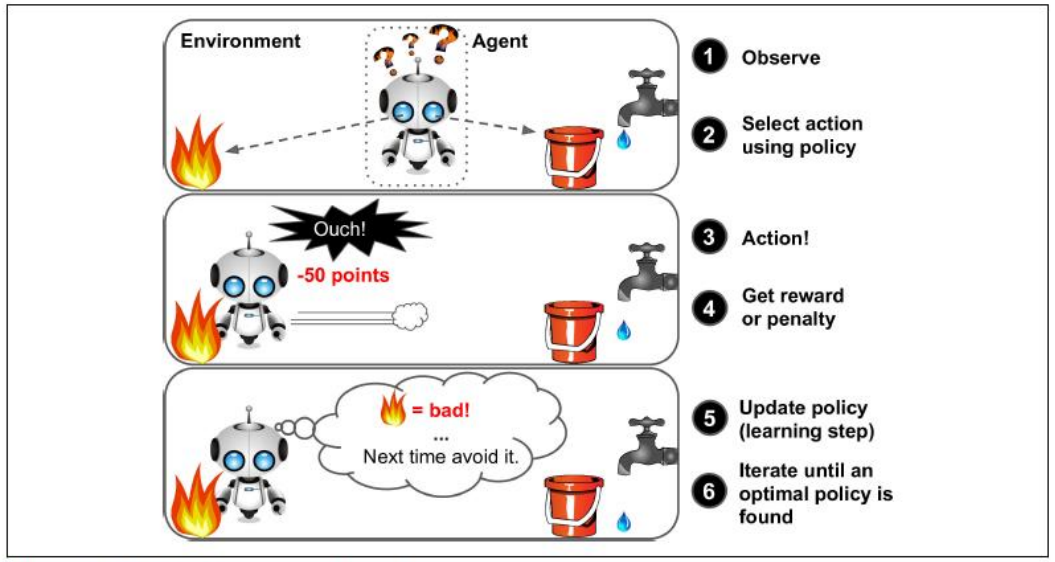


Figure 1-12. Reinforcement Learning

Batch learning

- ***Train offline on available data set then launch***
- ***Advantage: Simple***
- ***Disadvantage: CPU-time, Cannot handle new data types***
- ***Cannot train a new system every day***

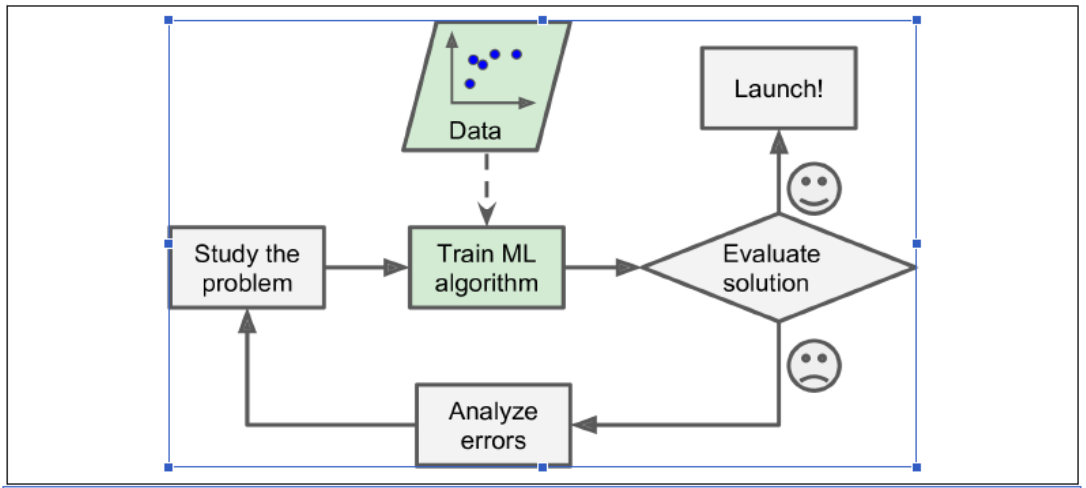


Figure 1-2. Machine Learning approach

Online learning

- ***Feed data in mini-batches -> Train -> Adapt to new data***
- ***Learning rate how fast should it learn ?***
- ***Advantage: Handle Continuous data flow. Handle Huge data***
- ***Disadvantage: Bad data can decline the performance. E.g. Bot attack or a malfunctioning sensor.***

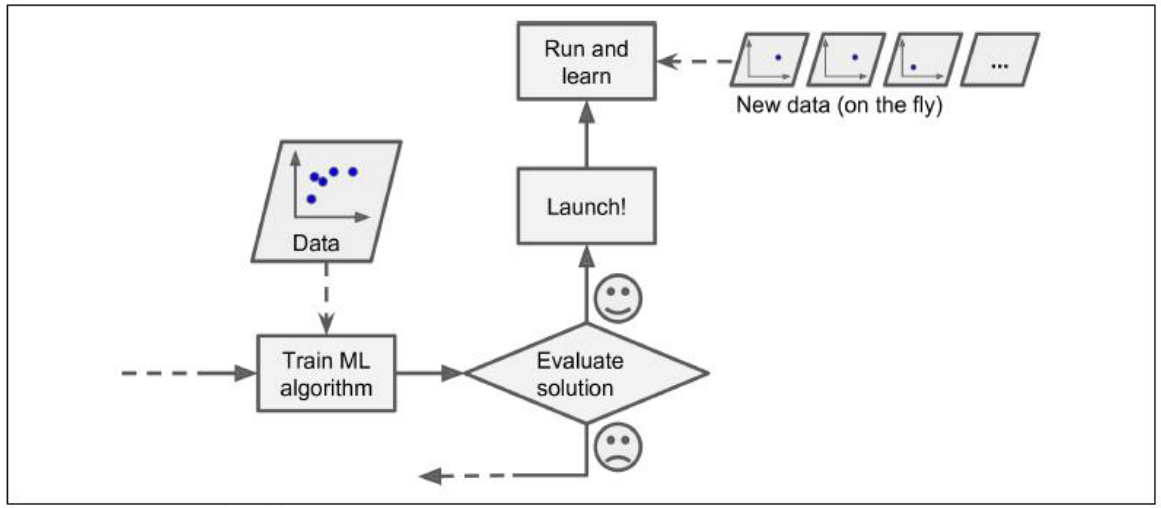


Figure 1-13. Online learning

Instance based learning

- ***Learn -> Use similarities to Generalize -> Make Predictions***

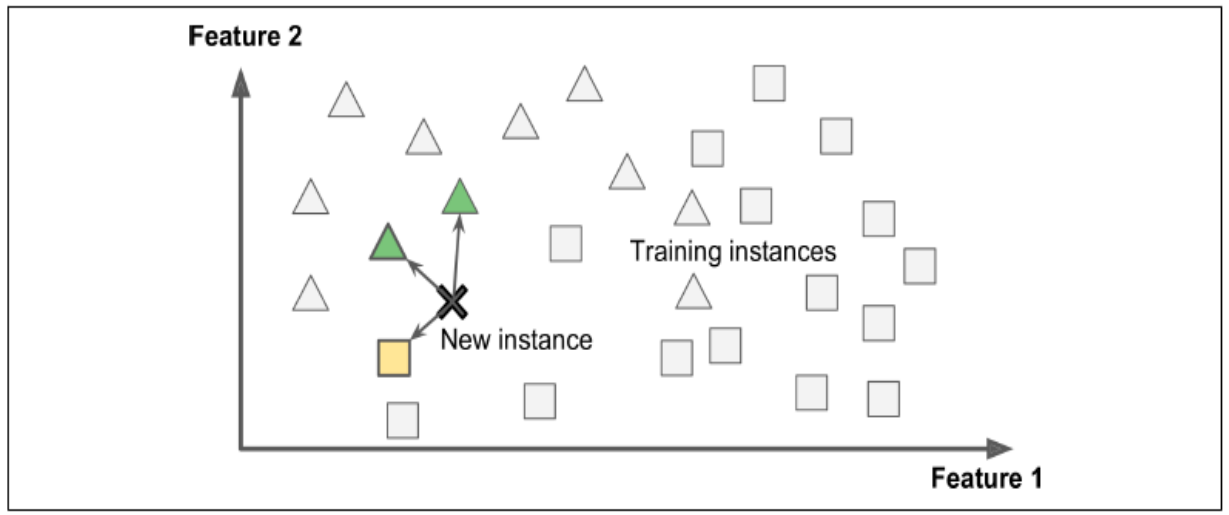


Figure 1-15. Instance-based learning

Model based learning: Classification

- ***Learn by examples -> Build a model -> Make Predictions***

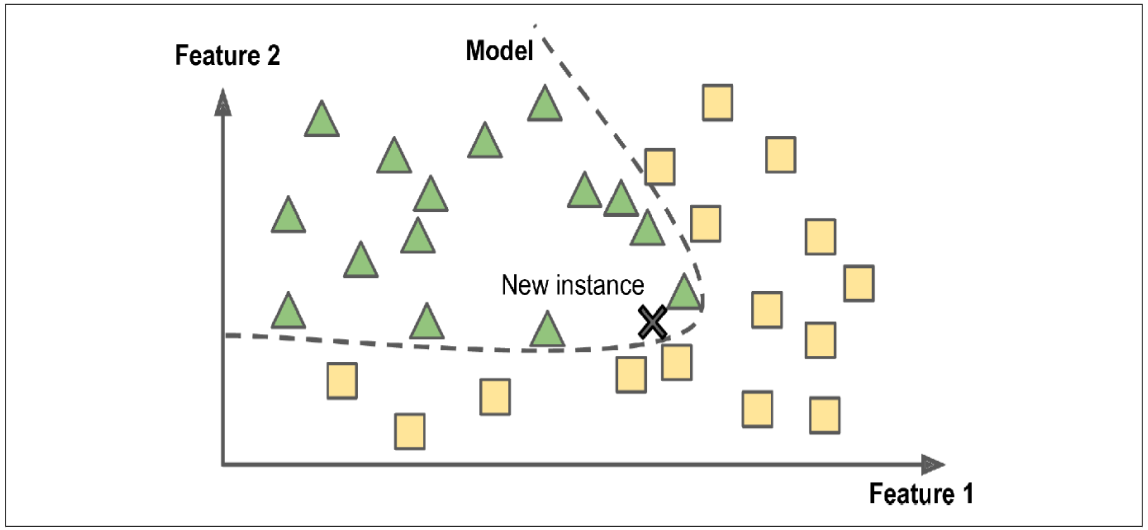


Figure 1-16. Model-based learning

Iris: Probability plot with decision boundaries

- **Probability function using petal length and petal width and 3 classes**

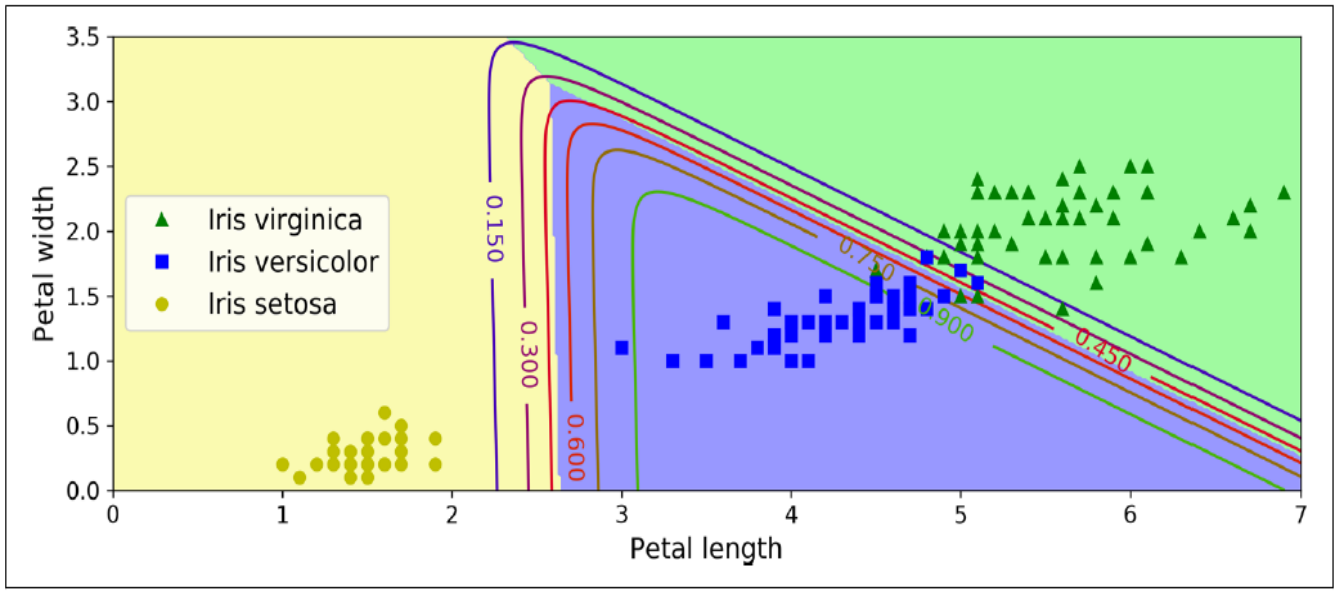


Figure 4-25. Softmax Regression decision boundaries

- **Notice the linear decision boundaries (e.g. green 90%probability) for Iris Versicolor**

Model based learning. Regression

- ***Learn by examples -> Build a model -> Make Predictions***

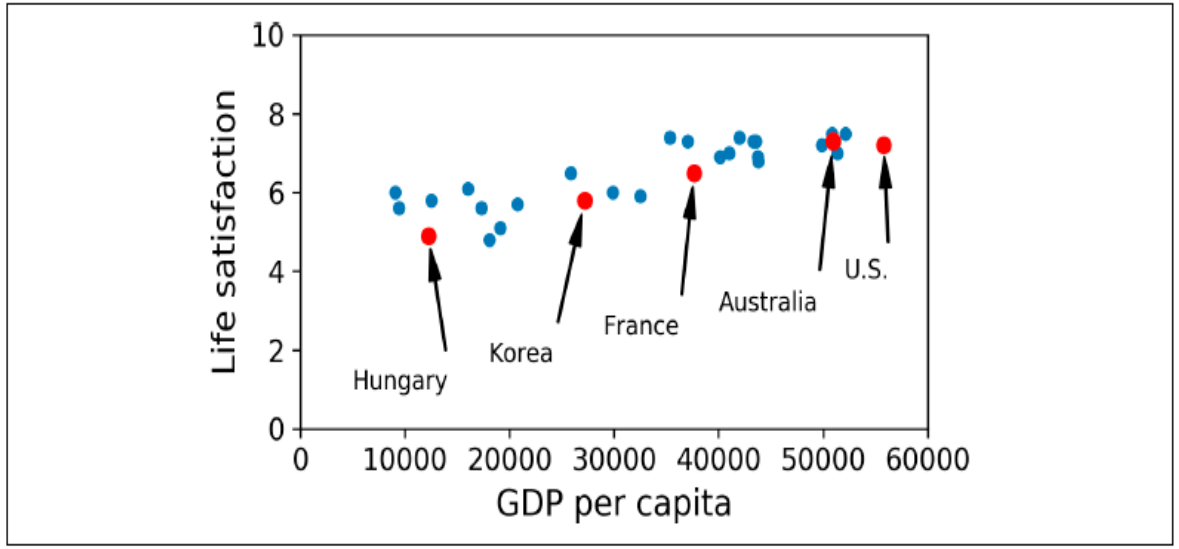


Figure 1-17. Do you see a trend here?

Model based learning. Regression

- ***Find the best linear model: bx + a***
- **$\theta_0 + \theta_1$xGDP**

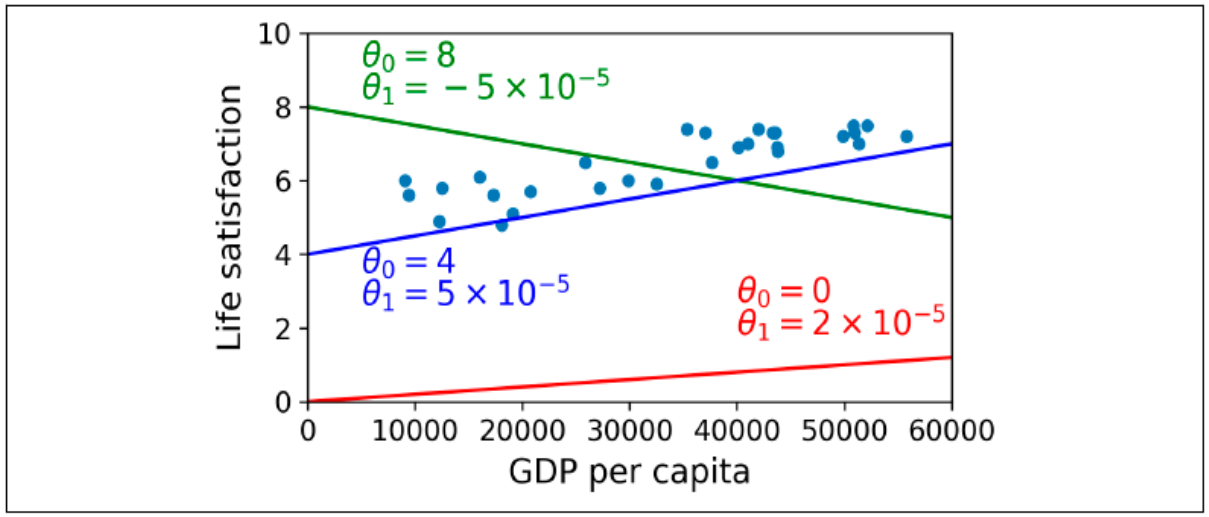


Figure 1-18. A few possible linear models

Model based learning. Regression

- ***Find the best linear model: bx + a***
- **$\theta_0 + \theta_1$xGDP**

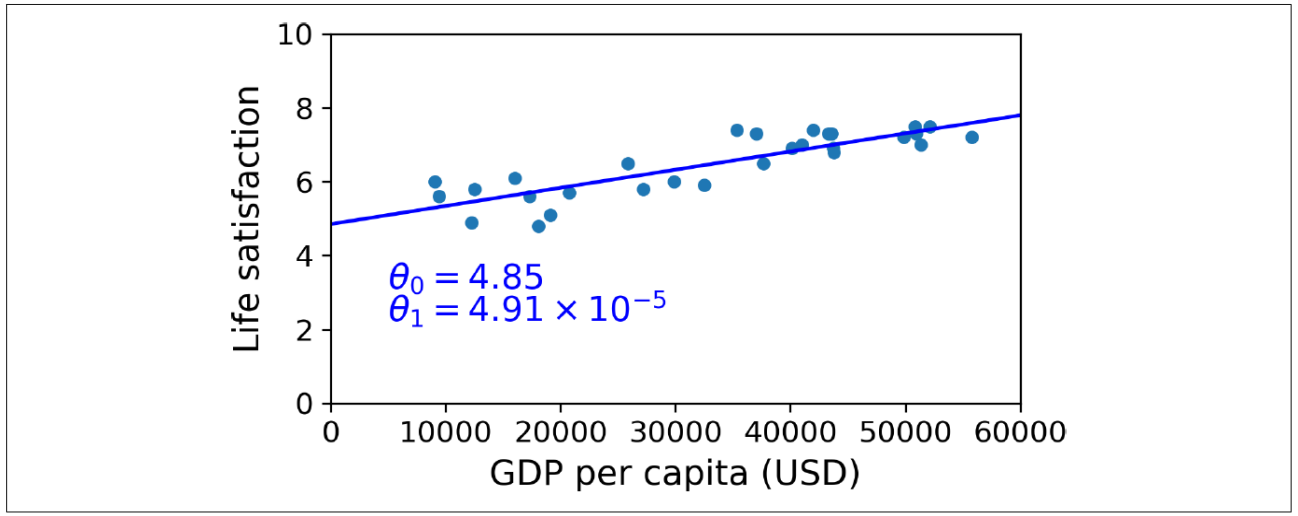


Figure 1-19. The linear model that fits the training data best

Regresison Code Example

- **Using ML to find the best linear model $\theta_0 + \theta_1 x$GDP**
- **Training and running the program !**
- **Show and look at: Life satisfaction code !**

Challenges of Machine Learning

- **Insufficient quantity of training data**
- **Nonrepresentative training data**
- **Poor quality data: outliners, noise, missing features for some data**
- **Irrelevant features**
- **Overfitting the training data**
- **Underfitting the training data**

Nonrepresentative training data

- ***Nonrepresentative data leads to a model with low accuracy for some data***

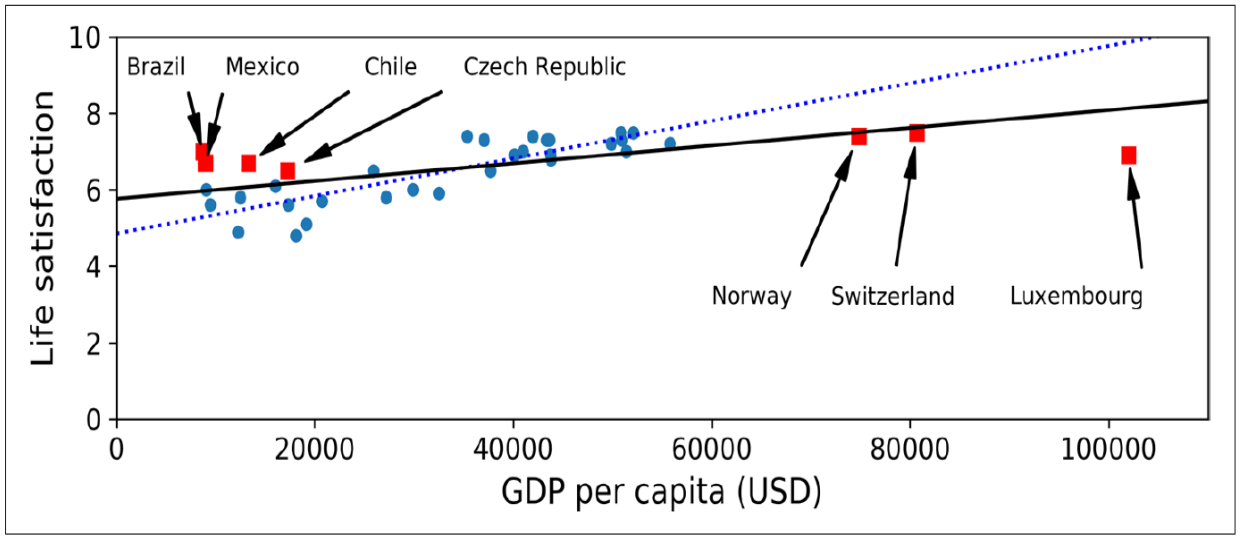


Figure 1-21. A more representative training sample

Overfitting

- ***Making the model too good on training set.***
- ***But low accuracy for other test data***
- ***Solution: use a more simple model and regularization***

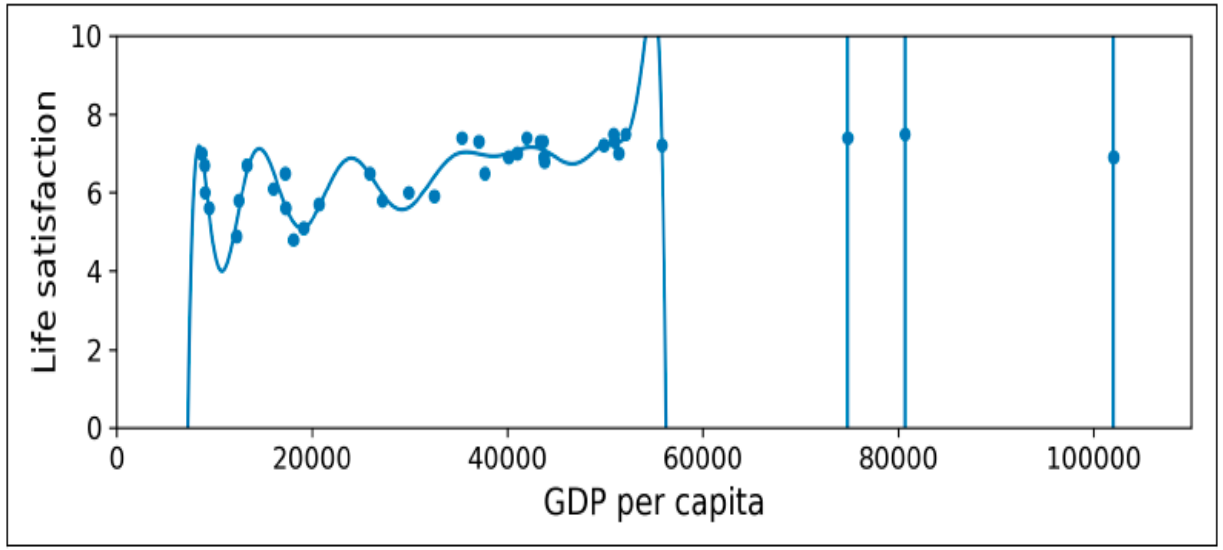
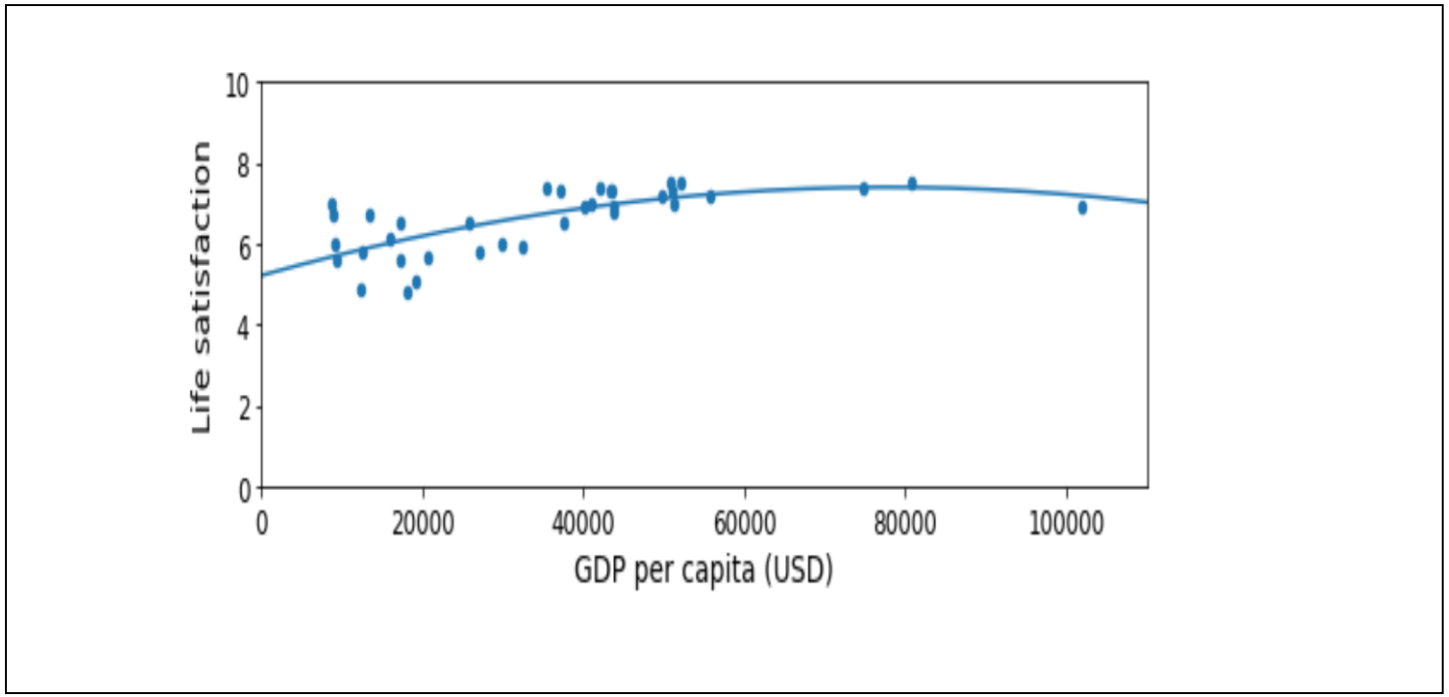


Figure 1-22. Overfitting the training data

Overfitting solution

- ***Making the model better on training set.***
- ***And hiogh accuracy for other test data***
- ***Solution: use a more simple model and regularization like degree = 2***



Regularization

- ***Setting constraints on the model parameters***

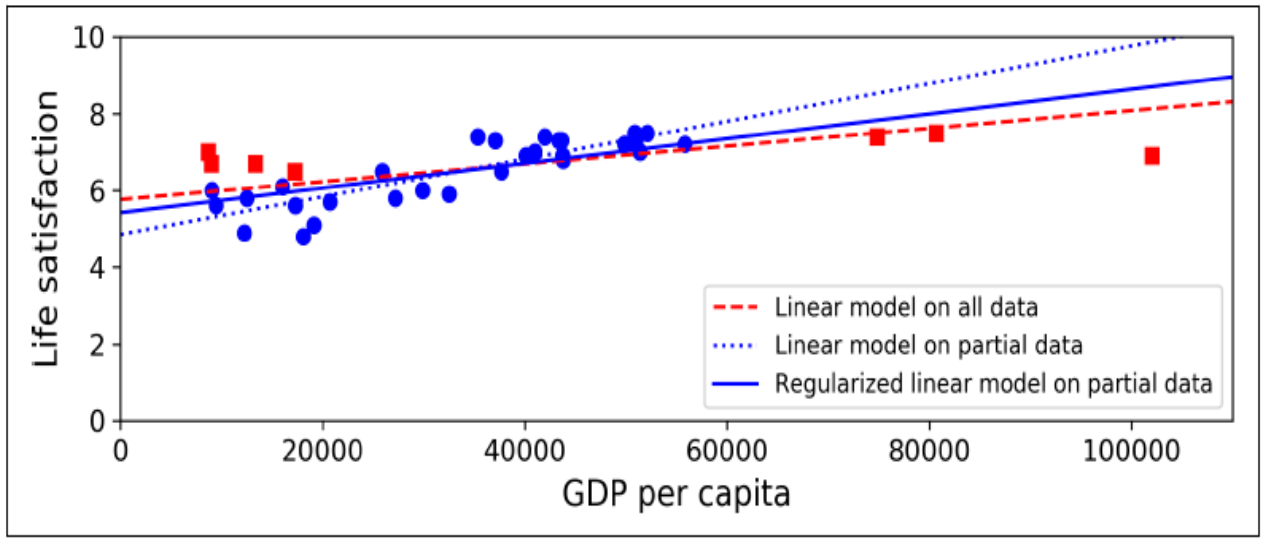


Figure 1-23. Regularization reduces the risk of overfitting

Underfitting

- ***Too simple model to find a structure in the data***
- ***Solution: Model with more parameters***
- ***Better features***
- ***Remove any constraints and regularization***

Testing and validation

- ***Split data set into training set and a test set***
- ***Train model on training set (80%-90% data)***
- ***Finally test finally model on test set (10%-20% data)***

Exercise

- **It is time for discussion, setting up the environment Anaconda and coding Python in Spyder !**

- Chapter 1 Assignments: No. 1 - 14
- Anaconda Installation Guide
- Jupiter Test
- Python Basic No. 1