Basics of machine learning

Tamás Grósz

Machine learning

Formal definition

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E. (Tom M. Mitchell)

Machine learning

Formal definition

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E. (Tom M. Mitchell)

- Machine learning algorithms build a mathematical model of sample data (training data)

Machine learning

Formal definition

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E. (Tom M. Mitchell)

- Machine learning algorithms build a mathematical model of sample data (training data)
- The goal is to create an agent that can make predictions or decisions without being explicitly programmed to perform the task.

Types of machine learning

Reinforcement learning

Reinforcement Learning(RL) is a type of machine learning technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences. Commonly used for game AI.

Types of machine learning

Reinforcement learning

Reinforcement Learning(RL) is a type of machine learning technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences.
Commonly used for game AI.

Unsupervised learning

It tries to identify commonalities in the data and reacts based on the presence or absence of such commonalities in each new piece of data.
Most common algorithms: Clustering, AutoEncoders, PCA, etc.

Types of machine learning

Supervised learning learning

In supervised learning, each example is a pair consisting of an input vector (x) and an expected output value (y).

- We will focus on this area.
- Two version: classification and regression

Supervised learning

Classification

- The task is identifying to which category a new observation belongs, on the basis of the training data
- y in this case is nominal
- Examples: image classification, NLP, speech recognition

Supervised learning

Classification

- The task is identifying to which category a new observation belongs, on the basis of the training data
- y in this case is nominal
- Examples: image classification, NLP, speech recognition

Regression

- The problem of predicting a continuous quantity output for an example
- y is numerical
- Examples: predicting stock prices, predicting price of a house

Classification

Statistical pattern recognition

- For each class we learn a function that predicts $p(y_i|x)$
- The learning process maximizes the $p(y_j|x)$ for the correct label (y_j)
- The decision boundary is where these function intersect

Classification

Statistical pattern recognition

- For each class we learn a function that predicts $p(y_i|x)$
- The learning process maximizes the $p(y_j|x)$ for the correct label (y_j)
- The decision boundary is where these function intersect

Geometric view

- We represent the decision hyperplane directly
- In 2D it could be lines

Evaluation

- After training we need to measure how well the model generalizes

Evaluation

- After training we need to measure how well the model generalizes
- Test set: previously unseen examples with labels, it is used to calculate the P metric

Evaluation

- After training we need to measure how well the model generalizes
- Test set: previously unseen examples with labels, it is used to calculate the P metric
- Many different metrics exists: Accuracy, EER, F_1-score, AUC etc.

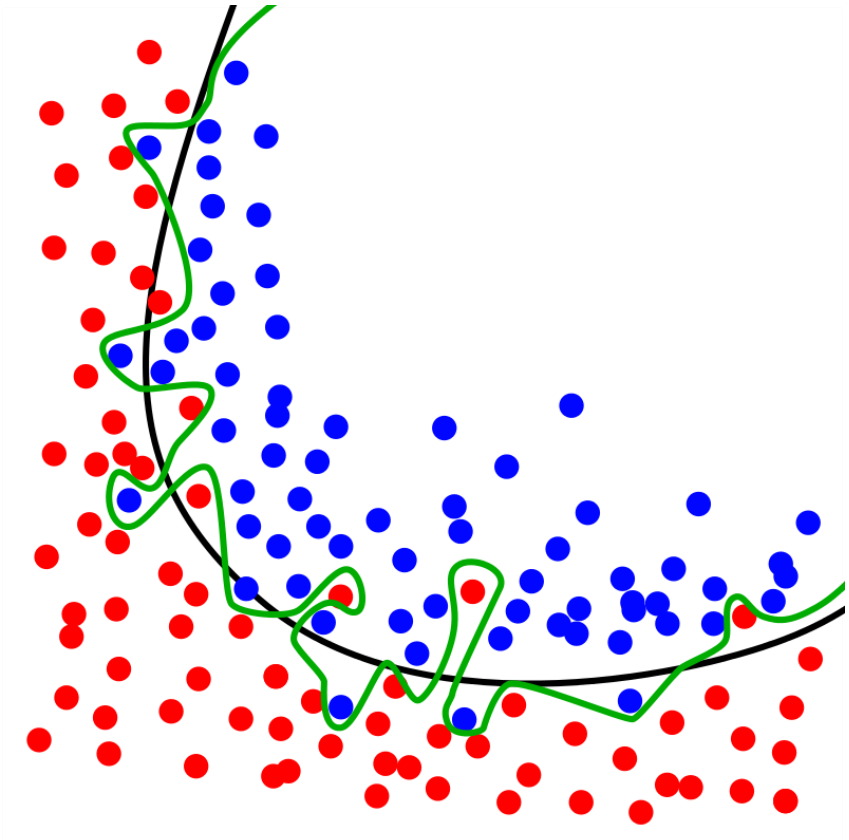
Parameters and Hyperparameters

- All machine learning algorithm has many parameters
- These parameters are learned by the algorithm
- For example: the steepness of the line

Hyperparameters

- These values are set before the learning process begins
- They control the optimizer, the model capacity and other things
- Examples: Learning rate, number of optimization steps, etc.

Overfitting



Definition: "the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably"

Best practices

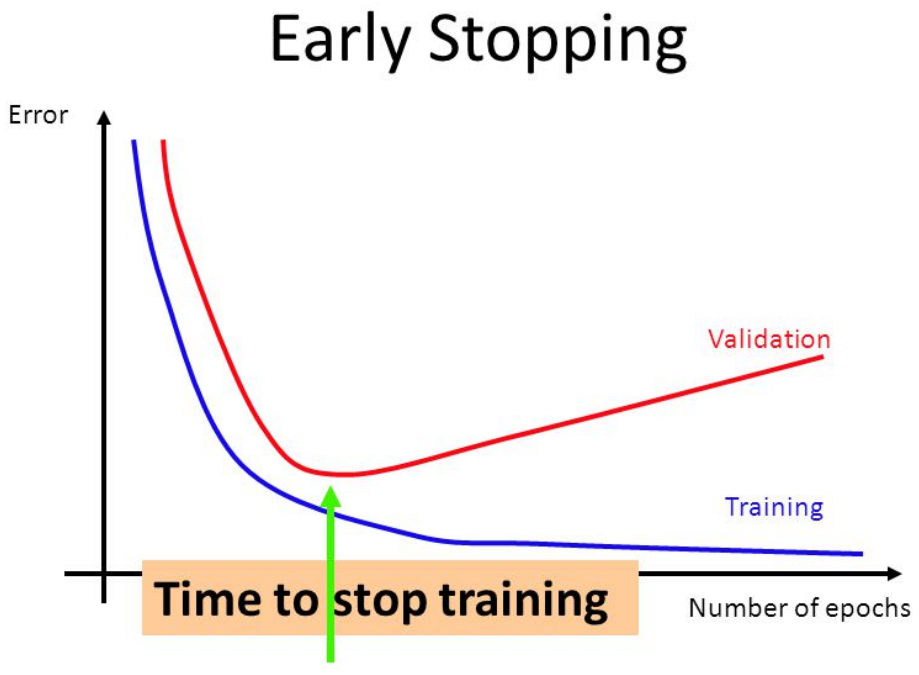
Development set

- It could help, if we simulate the test data during training
- Dev set: part of the training data, which are not used to optimize the parameters, but rather to set the hyperparameters

- Besides validation, we can use other methods to avoid overfitting
- There is no data like more data
- Use validation to set the hyperparameters and only evaluate the best model on the test

Best practices

- Always try to have a small model (faster evaluation and usualy better generalization)
- Avoid peeking (using the test data too much)



Practice

`www.inf.u-szeged.hu/~groszt/index.php?id=teaching`
Google cloud platform: https://cloud.google.com/
Alternative: google colab (in google drive)
Python basics: practice_02.ipynb