NTNU – Trondheim
Norwegian University of
Science and Technology

Machine Learning: a Non-Technical Introduction

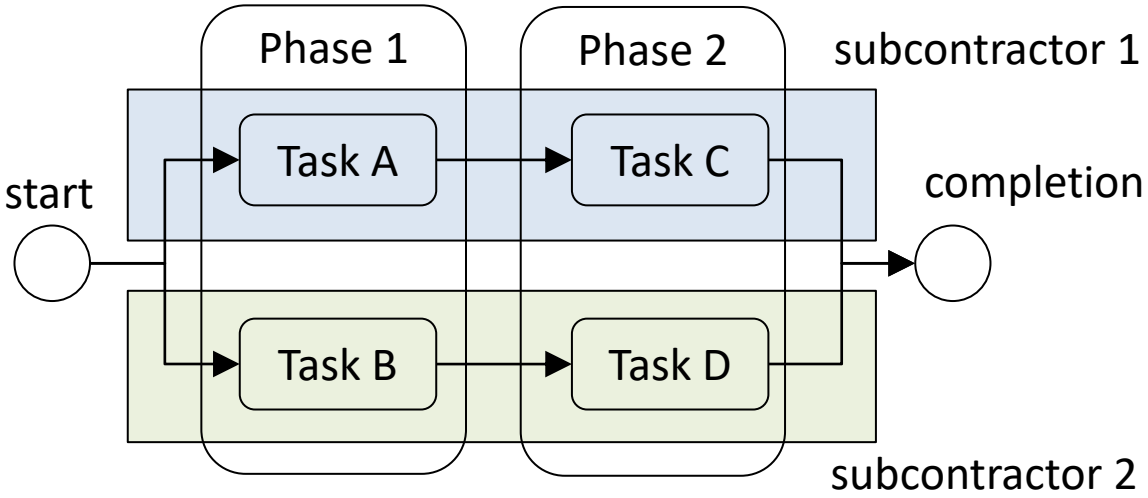
Antoine Rauzy
Antoine.Rauzy@ntnu.no

Agenda

- Motivating (Toy) Example
- Classification
- Regression
- Clustering
- Discussion

Presentation

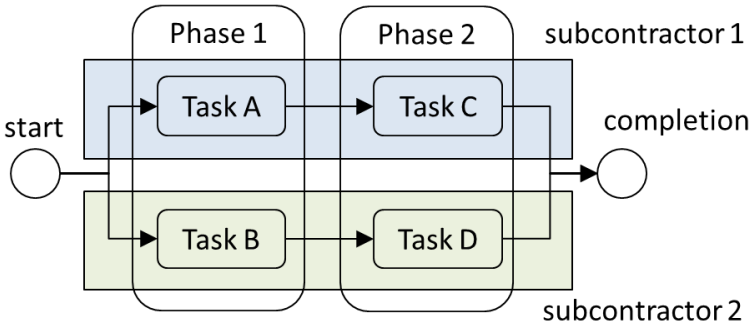
A contractor has 200 projects with the following pattern:



Performance indicators for tasks and projects:

- Competence in {Low, Medium, High}
- Durations of tasks A, B, C, and D

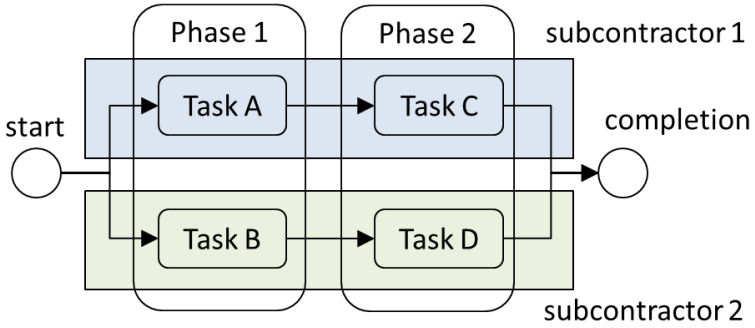
Objectives



What can we learn out of that?
Can Machine Learning help us?

What Can We Learn?

Duration of task A: ~60 weeks
Duration of task B: ~50 weeks
Duration of task C: ~40 weeks
Duration of task D: ~50 weeks



Total Duration = max(duration(A) + duration(C), duration(B) + duration(D))

Low competence : time wasted due to a bad coordination
High competence : time saved due to a good coordination
Medium competence : time neither wasted nor saved due to coordination

Objective : Predict project duration after completion of tasks A and B

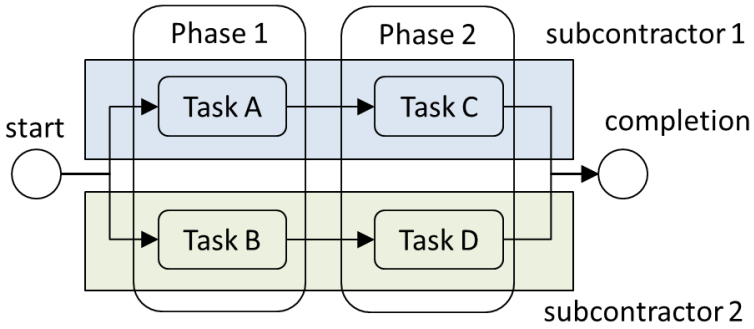
Agenda

- Motivating (Toy) Example
- Classification
- Regression
- Clustering
- Discussion

Classification

Expected duration: 100

labels
on-time: duration ≤ 115
delayed: $115 <$ duration ≤ 140
failed: duration > 140



Competence	Duration(A)	Duration(B)	Label
medium	75	69	delayed
low	58	53	on-time
…	…	…	…

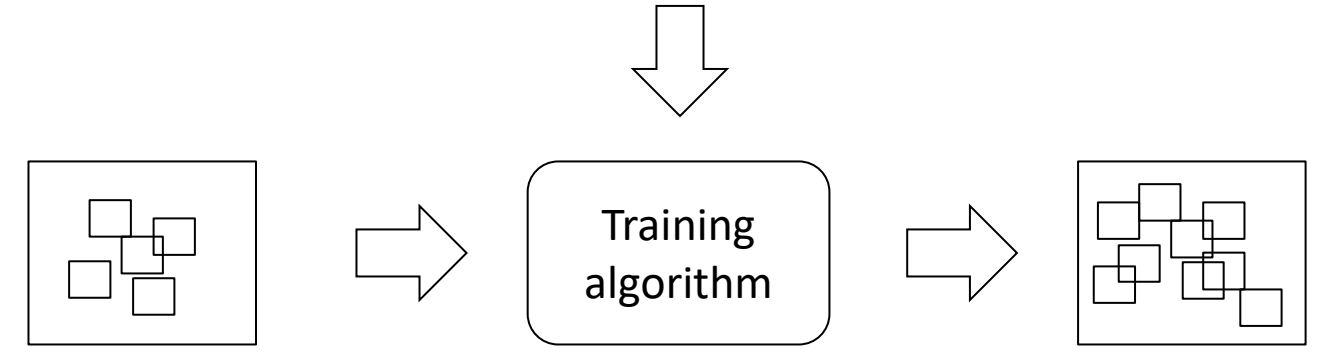
Classification problem:

- Given the performance indicators of phase 1, i.e. competence and durations of tasks A and B, predict the class of the project (on-time, delayed, or failed)

Phase 1: Training

Training set

Competence	Duration(A)	Duration(B)	Label
medium	75	69	delayed
low	58	53	on-time
...	...	...	...



Data structure (initial)

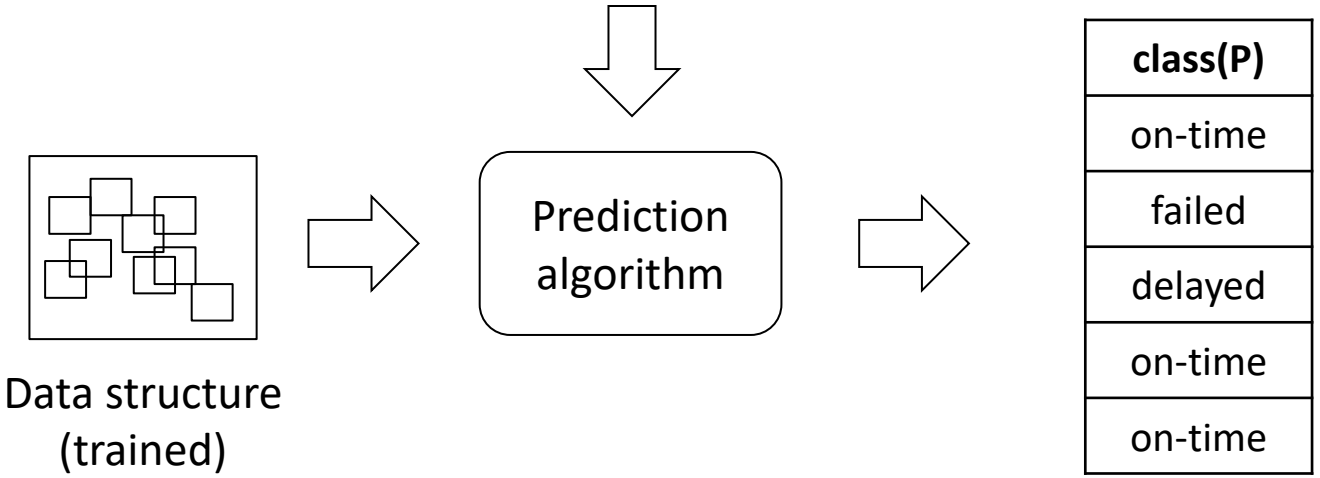
Data structure (trained)

Classification is a **supervised learning** problem: labels are known

Phase 2: Prediction

Test set

Competence	Duration(A)	Duration(B)	Label
medium	75	69	?
low	58	53	?
...	...	...	...



Performance

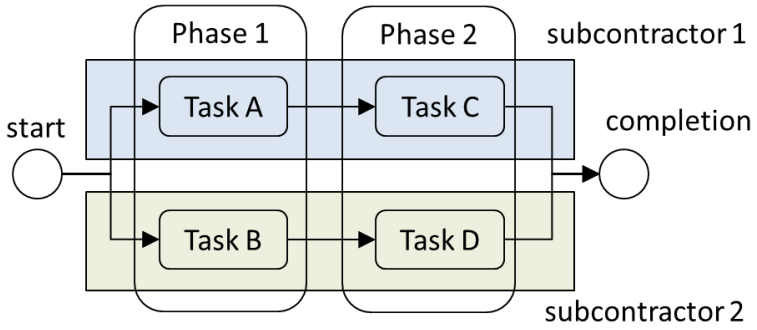
predicted classes

actual classes	on-time	delayed	failed
on-time	8	3	3
delayed	6	5	0
failed	2	7	6

Agenda

- Motivating (Toy) Example
- Classification
- Regression
- Clustering
- Discussion

Definition



Competence	Duration(A)	Duration(B)	Reward
medium	75	69	138
low	58	53	112
…	…	…	…

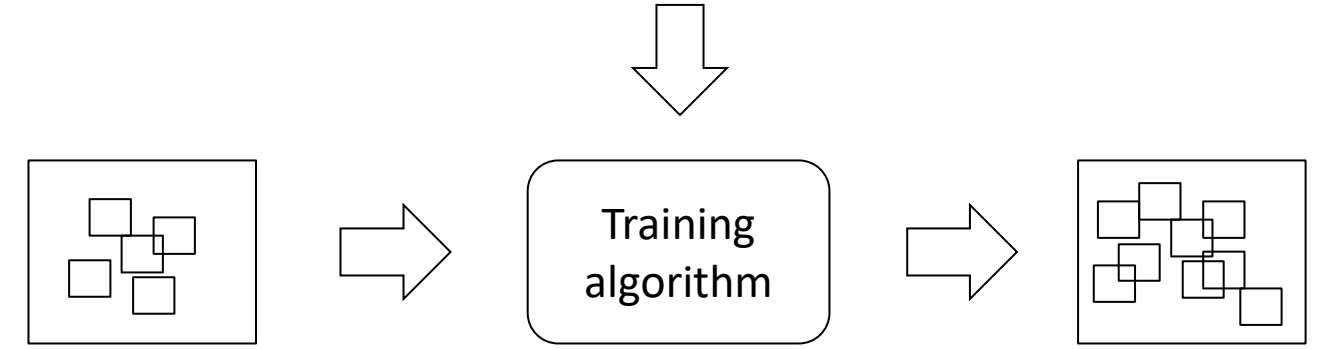
Regression problem:

- Given the performance indicators of phase 1, i.e. competence and durations of tasks A and B, predict the total duration

Phase 1: Training

Training set

Competence	Duration(A)	Duration(B)	Reward
medium	75	69	138
low	58	53	112
…	…	…	…



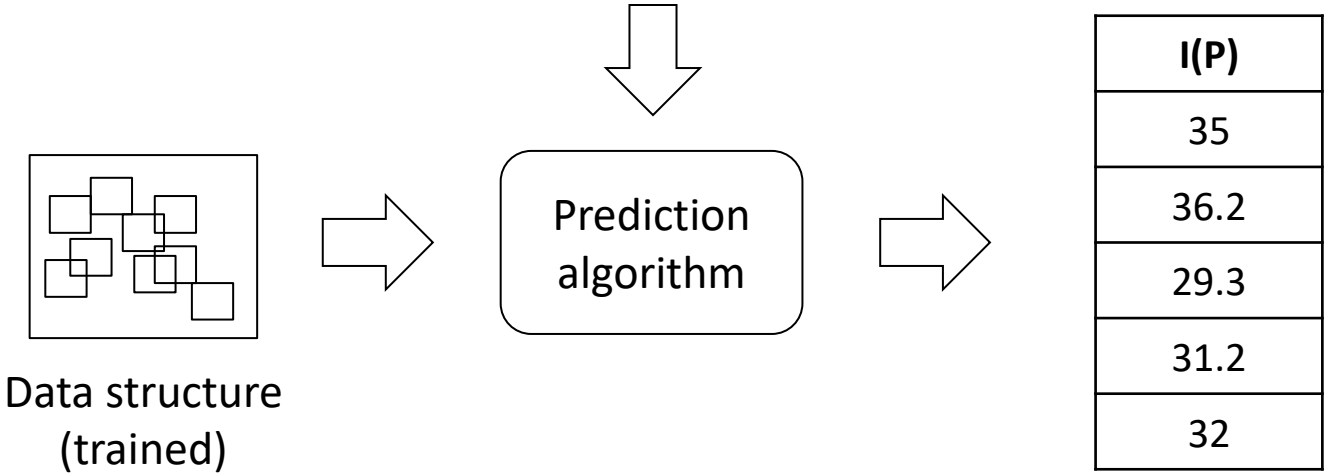
Data structure (initial)

Data structure (trained)

Regression is a **supervised learning** problem: rewards are known

Phase 2: Prediction

I(A)	I(B)
18.2	12.9
18.7	11.9
14.3	10.4
16	10.1
16.4	11.4



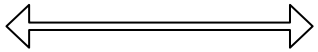
I(P)
35
36.2
29.3
31.2
32

Performance

Actual values

Duration
137
141
84
121
92

distance



Predicted values

Duration
130
130
120
126
123

Mean absolute error: $\frac{\sum |p - a|}{n}$

Mean squared error: $\frac{\sum (p - a)^2}{n}$

R^2

Agenda

- Motivating (Toy) Example
- Classification
- Regression
- Clustering
- Discussion

Definition

Competence	Duration task A	Duration task B	Duration task C	Duration task D	Total
medium	75	69	69	69	138
low	58	53	59	59	112
medium	58	60	63	63	123
low	61	80	93	93	173
high	62	55	43	43	98
high	50	70	55	55	125
medium	60	81	76	76	157
low	88	72	79	79	161
medium	51	53	57	57	110

Clustering problem:

- Can we group projects into clusters?

Unsupervised learning

Agenda

- Motivating (Toy) Example
- Classification
- Regression
- Clustering
- Discussion

Conclusion

	Strengths	Weaknesses
Internal	• Easiness of implementation (no programming competences required)	• There must be something to learn? • Data preprocessing
External	• Results • New way of thinking	• Availability of a lot of good data • Performance

Specific issues:

- Time dependencies
- Accidents are non-deterministic