Introduction to intelligent systems

Machine learning

Mikkel N. Schmidt

Technical University of Denmark,
DTU Compute, Department of Applied Mathematics and Computer Science.

Overview

1. Machine learning problems
2. Machine learning algorithms
3. Linear regression
4. Generalization
5. Tasks

Feedback group

- David Svane-Petersen
- Yuxuan Zhang
- sebastian vargr
- Peter Vestereng Larsen

Learning objectives

I Types of machine learning problems.

I Generalization: Training and test error.

II Linear regression. Model, parameters, and cost function.

I Understand the concepts and definitions, and know their application. Reason about the concepts in the context of an example. Use correct technical terminology.

II As above plus: Read, manipulate, and work with technical definitions and expressions (mathematical and Python code). Carry out practical computations. Interpret and evaluate results.

Machine learning problems

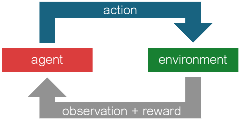
Machine learning problems

Categorization of learning problems

Unsupervised Learn function that describes the structure in data

Supervised Learn function that maps input to output to optimize cost

Reinforcement Learn a function (policy) that maps inputs to actions to optimize cumulative reward

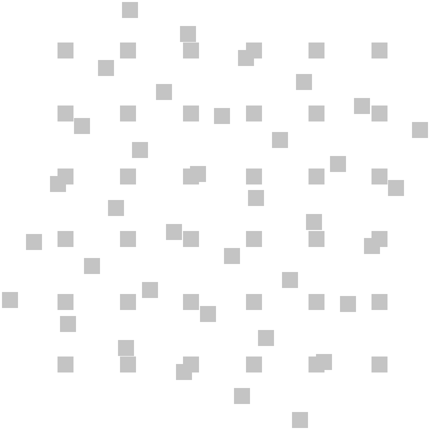


Unsupervised learning

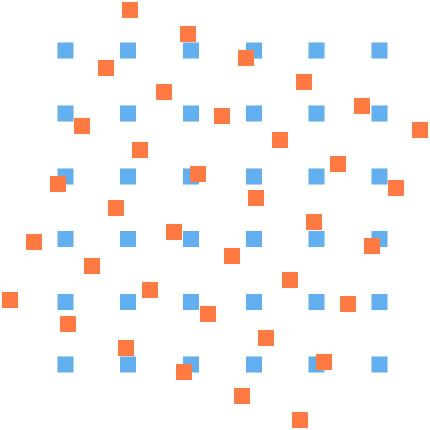
Learn a function that describes the structure in a data set

Clustering	Find a way to group data points into meaningful components
Dimensionality reduction	Find a lower-dimensional representation of the data
Anomaly detection	Find data points that deviate from “normal” behaviour

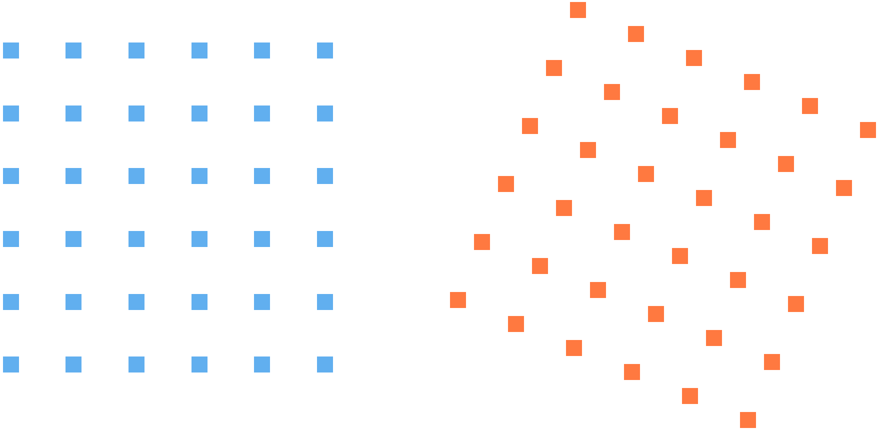
Unsupervised learning: Discover patterns in data



Unsupervised learning: Discover patterns in data



Unsupervised learning: Discover patterns in data

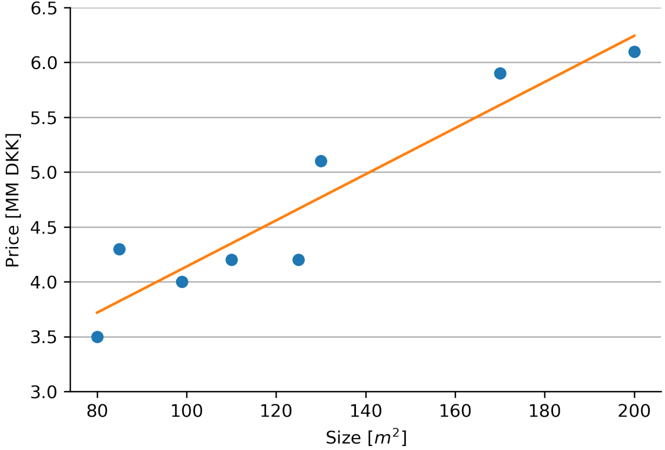


Supervised learning

Learn a function that maps an input to an output to optimize a cost

Regression	Outputs are continuous variables
Classification	Outputs are discrete classes
Ranking	Output is a ranking of the data objects

Supervised learning: House price prediction



Reinforcement learning

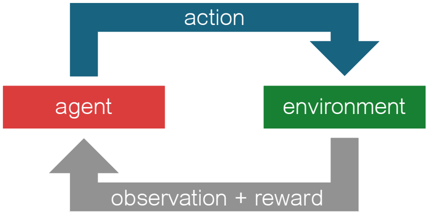
Learn a function (policy) that maps inputs to actions to optimize cumulative reward

Evaluation vs. control

- **Passive** Evaluate the future reward for a given policy
- **Active** Estimate the optimal policy by exploration

Obervability of the environment

- **Full** Agent knows the state of the environment
- **Partial** Agent must learn a representation of the environment



Exercise: What is human learning?

Is human learning best characterized as

- Unsupervised learning
- Supervised learning
- Reinforcement learning

(If you think the answer is somehow obvious, see if you can come up with an argument against)

Machine learning algorithms

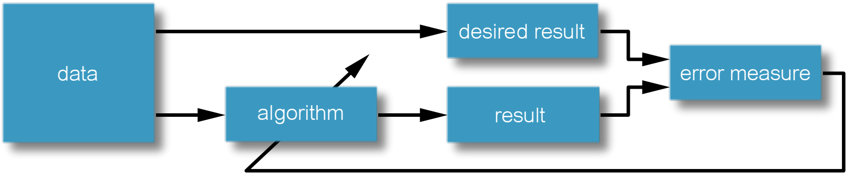
Machine learning algorithms

Machine learning

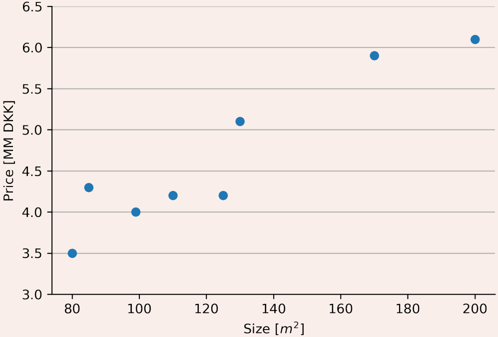
- Algorithm with tunable parameters
- Takes in some data and produces some output
- Measure error between algorithm's output and the desired output
- Tune parameters to minimize error

Goal: Generalization = good performance on future/unseen data

Machine learning algorithms

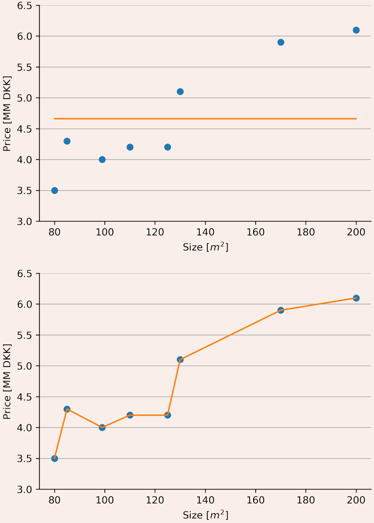


Exercise: Price of a 150 m^2 house



- What would you expect the price of a 150 m^2 house to be?
- Discuss which “algorithm” you used to come up with your answer

Exercise: House price regression



- Which of the above regression curves is best?
- Discuss how you could define a criteria for which is “best”

House price regression

Possible criteria for a good regression line

Fit	the observed data well
Robust	to small changes in the data
Generalize	to (unseen) future data

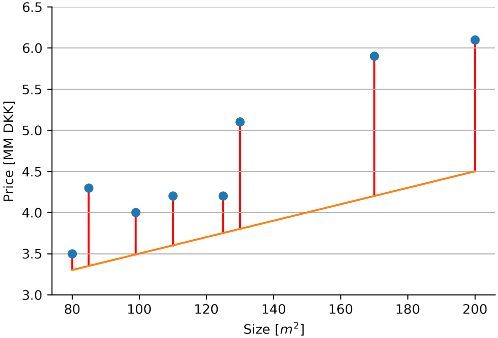
Linear regression

Linear regression

- Regression line: $f(x) = ax + b$
- Error: Squared distance between data and regression line

$$E = \sum_{n=1}^{N} (y_n - f(x_n))^2$$

- Find values of a and b to minimize E

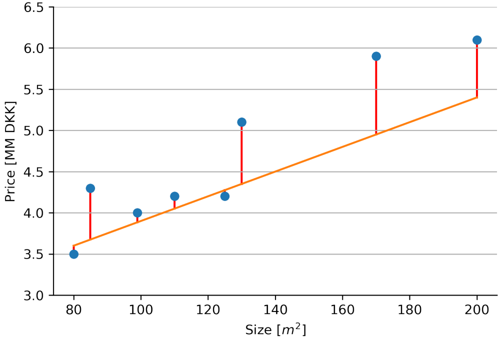


Linear regression

- Regression line: $f(x) = ax + b$
- Error: Squared distance between data and regression line

$$E = \sum_{n=1}^{N} (y_n - f(x_n))^2$$

- Find values of a and b to minimize E

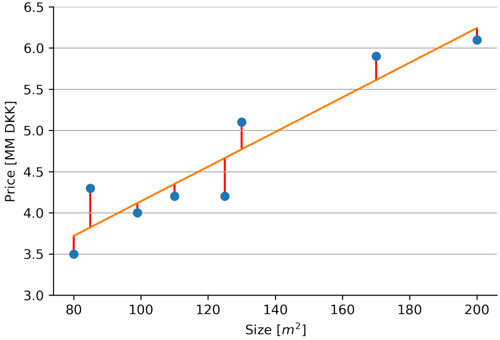


Linear regression

- Regression line: $f(x) = ax + b$
- Error: Squared distance between data and regression line

$$E = \sum_{n=1}^{N} (y_n - f(x_n))^2$$

- Find values of a and b to minimize E



Exercise: Least squares regression

Solve the least square regression problem by minimizing the error

- Differentiate the error measure wrt. the parameters a and b
- This gives you two equations in two unknowns to solve

Problem specification

- Data

$$x = \{80, 85, 99, 110, 125, 130, 170, 200\}$$
$$y = \{3.5, 4.3, 4, 4.2, 4.2, 5.1, 5.9, 6.1\}$$

- Regression function

$$f(x) = ax + b$$

- Error measure

$$E = \sum_{n=1}^{N} (y_n - f(x_n))^2$$

Some useful definitions

$$\bar{x} = \sum_{n=1}^{N} x_n = 999$$

$$\bar{y} = \sum_{n=1}^{N} y_n = 37.3$$

$$\overline{xy} = \sum_{n=1}^{N} x_n y_n = 4914.5$$

$$\overline{xx} = \sum_{n=1}^{N} x_n^2 = 136951$$

Solution: Equation for a

Differentiate wrt. a and equate to zero

$$E = \sum_{n=1}^{N} (y_n - f(x_n))^2 = \sum_{n=1}^{N} (y_n - a x_n - b)^2$$

$$\frac{dE}{da} = \sum_{n=1}^{N} -2(y_n - a x_n - b) x_n$$

$$= -2 \sum_{n=1}^{N} y_n x_n + 2a \sum_{n=1}^{N} x_n^2 + 2b \sum_{n=1}^{N} x_n$$

$$= -2\overline{xy} + 2a\overline{xx} + 2b\bar{x} = 0$$

$$\Rightarrow \underline{\overline{xx} \cdot a + \bar{x} \cdot b = \overline{xy}}$$

Some useful definitions

$$\bar{x} = \sum_{n=1}^{N} x_n = 999$$

$$\bar{y} = \sum_{n=1}^{N} y_n = 37.3$$

$$\overline{xy} = \sum_{n=1}^{N} x_n y_n = 4914.5$$

$$\overline{xx} = \sum_{n=1}^{N} x_n^2 = 136951$$

Solution: Equation for b

Differentiate wrt. b and equate to zero

$$E = \sum_{n=1}^{N} (y_n - f(x_n))^2 = \sum_{n=1}^{N} (y_n - ax_n - b)^2$$

$$\begin{aligned}
\frac{dE}{db} &= \sum_{n=1}^{N} -2(y_n - ax_n - b) \\
&= -2\sum_{n=1}^{N} y_n + 2a\sum_{n=1}^{N} x_n + 2Nb \\
&= -2\bar{y} + 2a\bar{x} + 2Nb = 0 \\
&\Rightarrow \underline{\bar{x} \cdot a + N \cdot b = \bar{y}}
\end{aligned}$$

Some useful definitions

$$\bar{x} = \sum_{n=1}^{N} x_n = 999$$

$$\bar{y} = \sum_{n=1}^{N} y_n = 37.3$$

$$\overline{xy} = \sum_{n=1}^{N} x_n y_n = 4914.5$$

$$\overline{xx} = \sum_{n=1}^{N} x_n^2 = 136951$$

Solution: Two equations in two unknowns

Two equations

$$\overline{xx} \cdot a + \bar{x} \cdot b = \overline{xy}$$
$$\bar{x} \cdot a + N \cdot b = \bar{y}$$

In matrix notation

$$\begin{bmatrix} \overline{xx} & \bar{x} \\ \bar{x} & N \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \overline{xy} \\ \bar{y} \end{bmatrix}$$

Some useful definitions

$$\bar{x} = \sum_{n=1}^{N} x_n = 999$$

$$\bar{y} = \sum_{n=1}^{N} y_n = 37.3$$

$$\overline{xy} = \sum_{n=1}^{N} x_n y_n = 4914.5$$

$$\overline{xx} = \sum_{n=1}^{N} x_n^2 = 136951$$

Solution by substitution

Solve for b in eq. (B)

$$b = \frac{\bar{y} - \bar{x} \cdot a}{N}$$

Insert in eq. (A)

$$\overline{xx} \cdot a + \bar{x} \cdot \underbrace{\frac{\bar{y} - \bar{x} \cdot a}{N}}_{b} = \overline{xy}$$

and solve for a

$$\begin{aligned} a &= \frac{N \cdot \overline{xy} - \bar{x} \cdot \bar{y}}{N \cdot \overline{xx} - \bar{x}^2} \\ &= \frac{8 \cdot 4914.5 - 999 \cdot 37.3}{8 \cdot 136951 - 999^2} \approx \underline{0.0210} \end{aligned}$$

Insert, and solve for b

$$b = \frac{37.3 - 999 \cdot 0.0210}{8} \approx \underline{2.04}$$

Equations

$$\begin{aligned} (A) \quad & \overline{xx} \cdot a + \bar{x} \cdot b = \overline{xy} \\ (B) \quad & \bar{x} \cdot a + N \cdot b = \bar{y} \end{aligned}$$

Constants

$$\begin{aligned} \bar{x} &= 999 \\ \bar{y} &= 37.3 \\ \overline{xy} &= 4914.5 \\ \overline{xx} &= 136951 \end{aligned}$$

Solution by solving matrix equation in Python

Two equations in matrix notation

$$\begin{bmatrix} \overline{xx} & \bar{x} \\ \bar{x} & N \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \overline{xy} \\ \bar{y} \end{bmatrix}$$

```
>>> N, x, y, xy, xx = 8, 999, 37.3, 4914.5, 136951

>>> X = np.array([[xx, x],[x, N]])
>>> print(X)
[[136951    999]
 [   999      8]]

>>> y = np.array([xy, y])
>>> print(y)
[4914.5   37.3]

>>> a, b = np.linalg.solve(X, y)
>>> print(f'a = {a:.3}, b = {b:.3}')
a = 0.021, b = 2.04
```

Some useful definitions

$$\bar{x} = \sum_{n=1}^{N} x_n = 999$$

$$\bar{y} = \sum_{n=1}^{N} y_n = 37.3$$

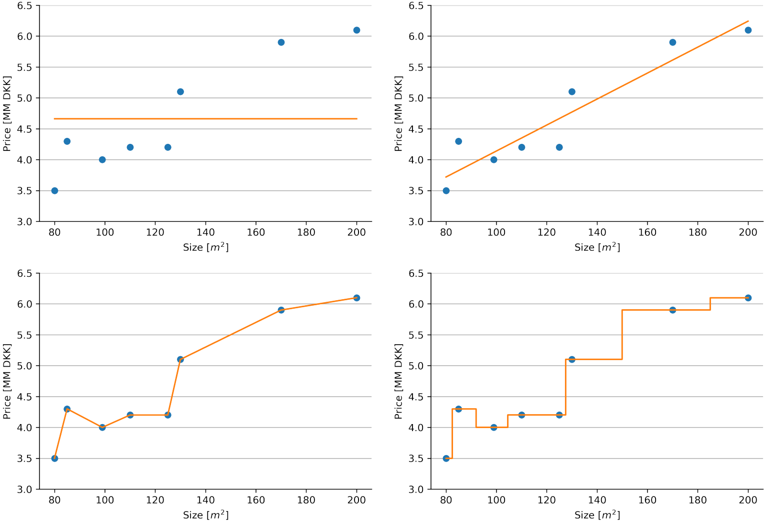
$$\overline{xy} = \sum_{n=1}^{N} x_n y_n = 4914.5$$

$$\overline{xx} = \sum_{n=1}^{N} x_n^2 = 136951$$

Generalization

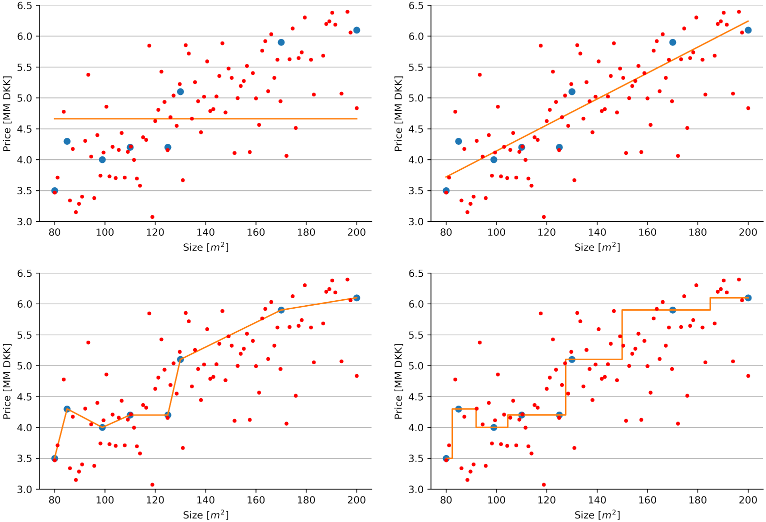
House price regression: Generalization

- If we knew future house prices, we could measure generalization error



House price regression: Generalization

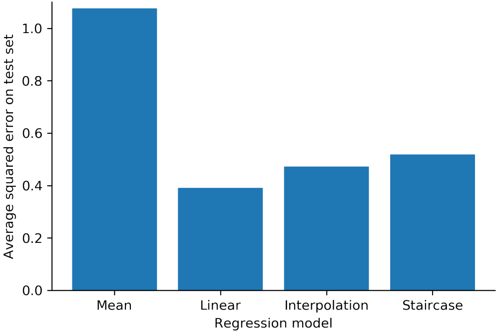
- If we knew future house prices, we could measure generalization error



House price regression: Generalization

Generalization error / out-of-sample error

- Average error on future data

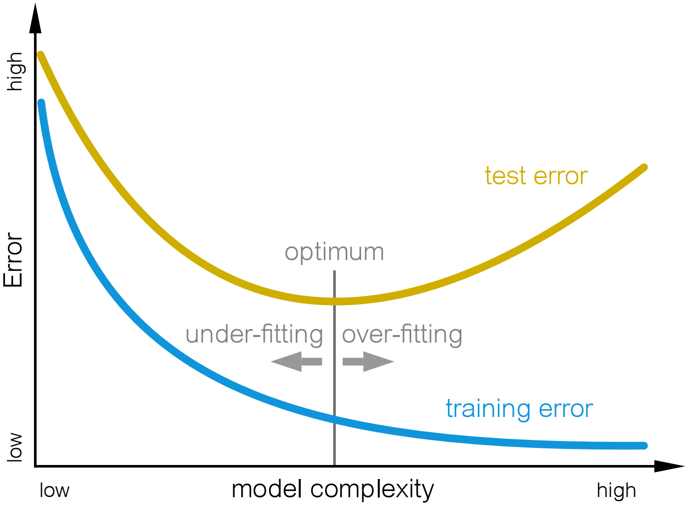


But, of course, we don't actually have access to future data

Cross-validation

- We only have access to a finite data sample
- Split the data sample into two parts called the *training set* and the *test set*
- Fit the models using the training set
- Evaluate and compare model performance on the test set

Model complexity



Tasks

Tasks for today

Tasks today

1. Work through the *regression complexity* notebook
   `06-RegressionComplexity.ipynb`

Today's feedback group

- David Svane-Petersen
- Yuxuan Zhang
- sebastian vargr
- Peter Vestereng Larsen