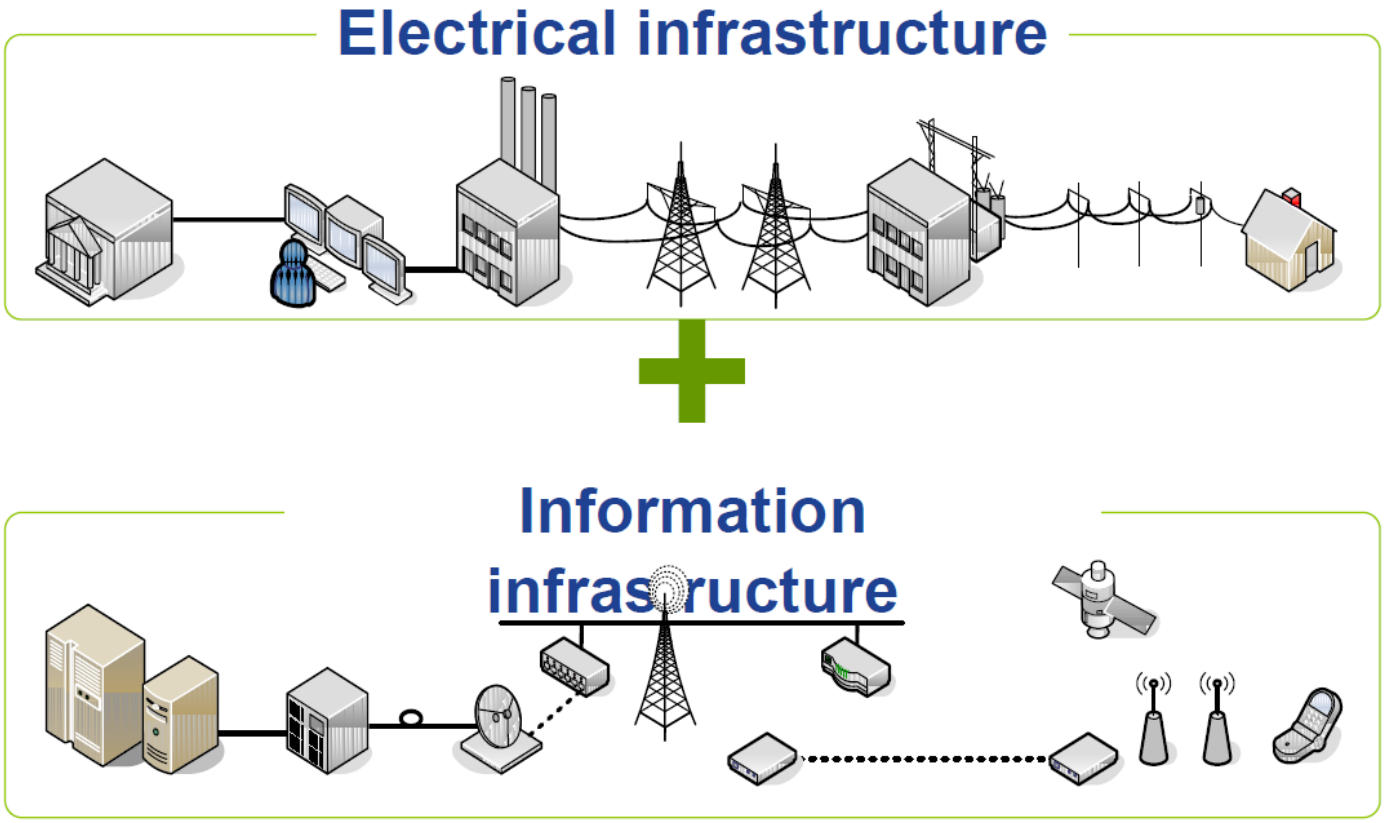
INSTITUTE OF
COMPUTER SCIENCE
Masaryk University

Machine learning on smart-grid data

Outline

- smart-grid infrastructure
- available data
- data quality
- example questions / problems
- machine learning applications
- machine learning on big data – boosted decision trees

Smart-grid infrastructure



Available data I

- static
  - topological data, consumer distribution tariff, FW version, ...
- dynamic
  - sensor measurements
    - consumption, production, voltage, current, ... (ca. 30 variables, 15 min period)
  - events reported by devices
    - tariff switching, power on/off, overvoltage, ...
    - % data transferred, transmission failure, ...
  - monitoring data
    - memory available, battery status, communication times, ...

Available data II

- additional computed / derived data
- data from external sources
  - weather forecast, cellular infrastructure data, ...
- millions of customers ⇒ millions of devices ⇒ billions of measurements per day
  - 3.5 millions of smart-meters (ČEZ)
  - 30 measured variables
  - 96 measurements a day
  - $3.5 \times 10^6 \times 96 \times 30 \times 4B \sim 40$ GB / day
  - soon becomes „BIG DATA“

Data quality

- high reliability (but not always!)
- communication issues
  - ⇒ missing data
  - ⇒ inhomogeneity
- inconsistency issues

- need for complex validation
- data quality and completeness determination
- missing values imputation / estimation

Example questions / problems

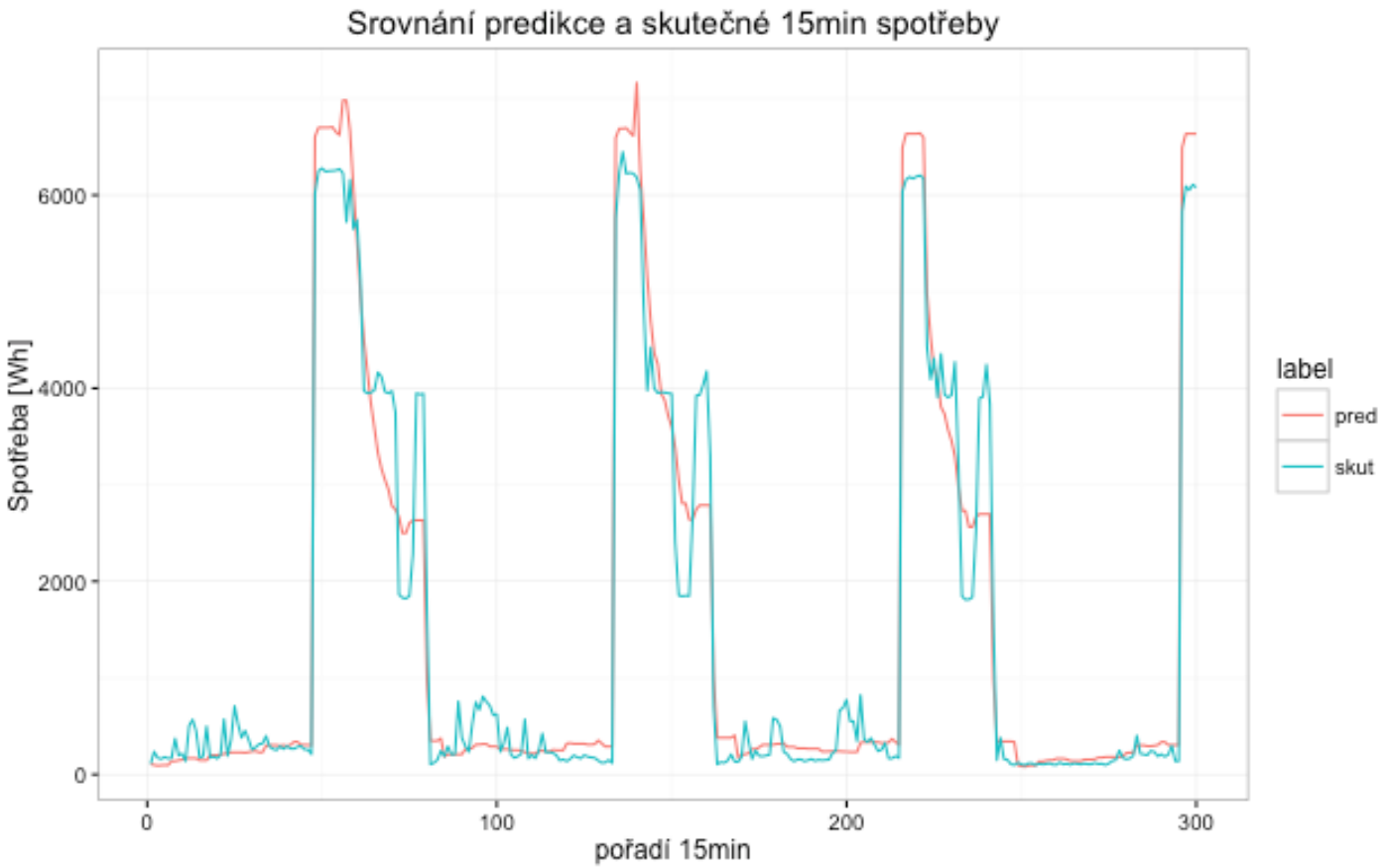
- operational problems detection and identification
- local load control
  - can we balance local consumption and production?
  - solar plants x water heaters, el. heating, batteries, ...
- technical / non-technical losses
- problematic localities identification
- customer clustering
- ...

Machine learning application – regression I

- quantitative output
- looking for function $\mathbb{R}^n \rightarrow \mathbb{R}$
- many methods
- metrics
  - mean square error (MSE)
  - mean absolute error (MAE)
- missing data imputation
- prediction of future values

c15	*t*	*15m*	*day*	*cavg*	*tavg*	…
76	12.36	77	119	494	7.21	…
4158	3.64	2	89	842	5.71	…
1041	8.76	89	89	494	9.20	…
267	-3.46	94	5	47	-2.94	…
1131	-10.9	13	21	494	-7.88	…
…						
?	4.56	23	103	97	9.54	…
?	20.74	72	208	125	19.17	…
?	10.37	24	102	842	11.34	…

Machine learning application – regression II



Machine learning applications – classification

- qualitative output (f: $\mathbb{R}^n \rightarrow C$)
- finite set of classes
- metrics
  - accuracy
  - precision / recall
  - AUC
- localities
  - controllable / uncontrollable
  - problematic / stable
- operational problems

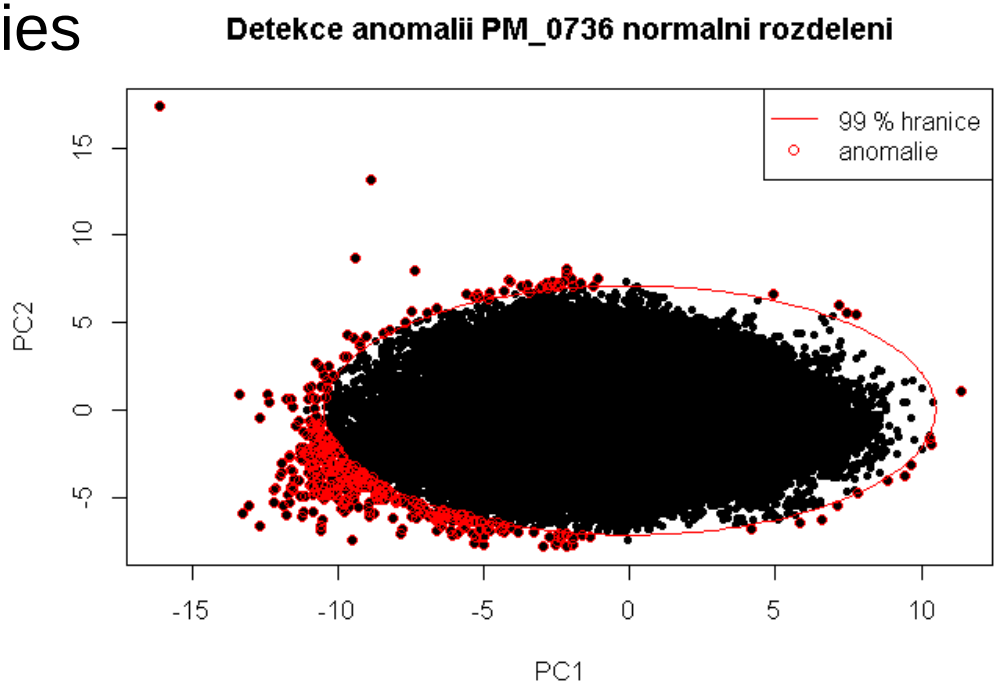
y	x_1	x_2	x_3	…
A	2.5	0.1	-3.1	…
C	-9.3	-3.7	8.0	…
B	-2.1	1.9	-9.2	…
A	6.3	3.3	-3.2	…
D	7.0	-7.7	3.8	…
…				
?	1.8	5.4	3.8	…
?	-3.5	-0.8	2.2	…
?	7.7	9.9	1.9	…

Machine learning applications – clustering

- grouping similar objects together
- unsupervised learning
- many different metrics / algorithms
- hard to evaluate

Machine learning applications – anomaly detection

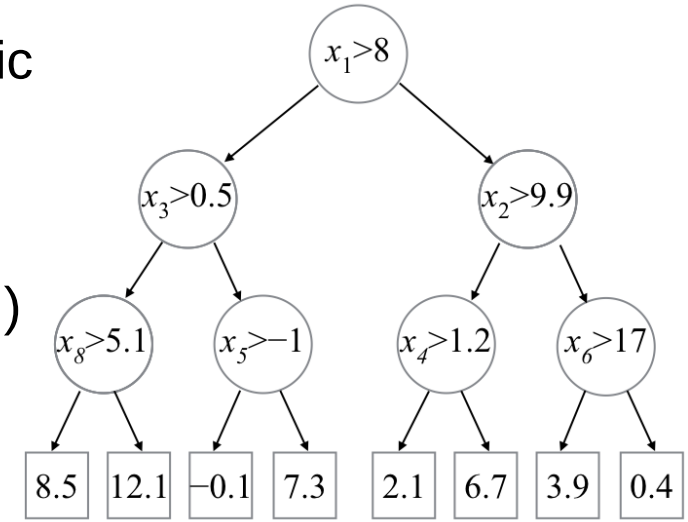
- detection of suspicious measurements
- detection of operational problems
- multivariate time-series



Machine learning on big data – boosted decision trees I

Decision trees

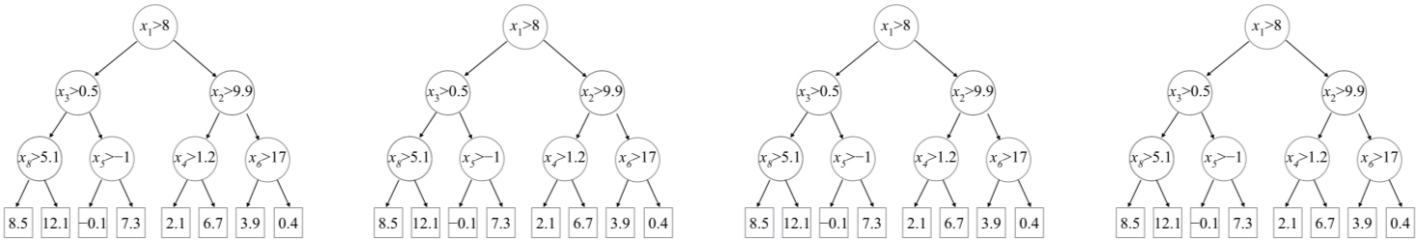
- split nodes
  - greedy algorithm
  - minimize cost function acc. to metric
- leaves
  - mean of outputs (reg.)
  - majority / distribution of classes (cl.)



Machine learning on big data – boosted decision trees II

Boosting

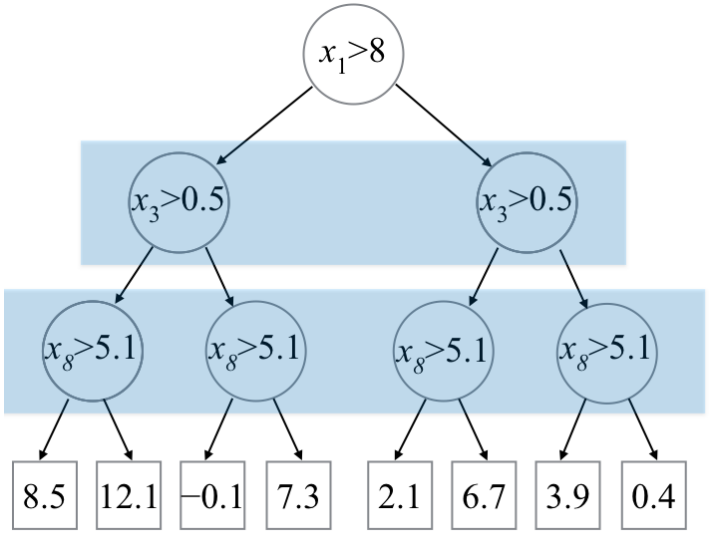
- ensemble model – forest rather than just one tree
- each tree built to minimize the error of the previous ones
- robustness



Machine learning on big data – boosted decision trees III

Obliviousness

- splits on each level are the same
- computational speed boost
  - learning phase
  - evaluation phase (no if‘s)
- robustness
  - resistance to overfitting
  - resistance to outliers



Machine learning on big data – boosted decision trees IV

Training on big data – distributed version

- Ph.D. thesis
- cost function computable „part by part“
  - not all of them satisfy the condition
- minimize the number of passes through data
  - efficiency
- fit into map-reduce or similar paradim