

Introduction to machine learning

Victor Kitov



November-December 2015 г.

Course information

- Instructor - Victor Vladimirovich Kitov
 - MSU, NES
 - practical experience
 - academic experience
 - ensemble learning
- 6 weeks+1 application project+exam
- seminars:
 - theoretical tasks
 - programming using python
 - ipython notebook, numpy, scipy, pandas, scikit-learn.
- course information
stellar.mit.edu/S/project/baymethodsadv/.

Course information

- Students
- Grading
 - attendance
 - homework
 - application project
- Bonus points
 - kaggle.com
 - commits to github

Recommended materials

- **Курс лекций по машинному обучению.** Константин Воронцов. (In Russian). machinelearning.ru.
- **Statistical Pattern Recognition.** 3rd Edition, Andrew R. Webb, Keith D. Copsey, John Wiley & Sons Ltd., 2011.
- **The Elements of Statistical Learning: Data Mining, Inference, and Prediction.** Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2nd Edition, Springer, 2009. <http://statweb.stanford.edu/~tibs/ElemStatLearn/>.
- **Machine Learning: A Probabilistic Perspective.** Kevin P. Murphy. Massachusetts Institute of Technology. 2012.
- **Pattern Recognition and Machine Learning.** Christopher M. Bishop. Springer. 2006.
- **Any additional public sources** - wikipedia, articles, tutorials, video-lectures.

Table of Contents

- 1 Tasks solved by machine learning
- 2 Main concepts of machine learning.
- 3 Practical applications of machine learning

Formal definitions of machine learning

- Machine learning is a field of study that gives computers the ability to learn without being explicitly programmed.
- 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 P at tasks in T improves with experience E .
- Examples: spam filtering, speech recognition, image recognition (face detection, eyes detection, pose detection, person identification).

Major niches of ML

- hard to formulate explicit rules (image recognition)
- further adaptation to usage conditions is required (voice detection)
- fast adaptation to changing conditions (stock prices prediction)

Connections with other fields

- Computer science
- Pattern recognition
 - recognize patterns and regularities in the data
- Artificial intelligence
 - create devices capable of intelligent behavior
- Time-series analysis
- Theory of probability, statistics
 - rely on probabilistic model
- Optimization methods
- Theory of algorithms

General problem statement

- Set of objects O
- Each object is described by a vector of known characteristics $\mathbf{x} \in \mathcal{X}$ and predicted characteristics $y \in \mathcal{Y}$.

$$o \in O \longrightarrow (\mathbf{x}, y)$$

- Usually $\mathcal{X} = \mathbb{R}^D$, \mathcal{Y} - a scalar, but they may be any structural descriptors of objects in general.

General problem statement

- Task: find a mapping f , which could accurately approximate $\mathcal{X} \rightarrow \mathcal{Y}$.
 - using a finite «training» set of objects with known (x, y) .
 - to apply on a set of objects of interest
- Questions solved in ML:
 - how to select object descriptors - features
 - in what sense a mapping f should approximate true relationship
 - how to construct f

Variants of problem statement

- For each new object x need to associate y .
- What is known:
 - $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$ - supervised learning:
 - x_1, x_2, \dots, x_N - unsupervised learning
 - dimensionality reduction
 - clustering
 - $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N), x_{N+1}, x_{N+2}, \dots, x_{N+M}$ - semi-supervised learning.
- If predicted objects x'_1, x'_2, \dots, x'_K for which y is forecasted, are known in advance, then this is «transductive» learning.

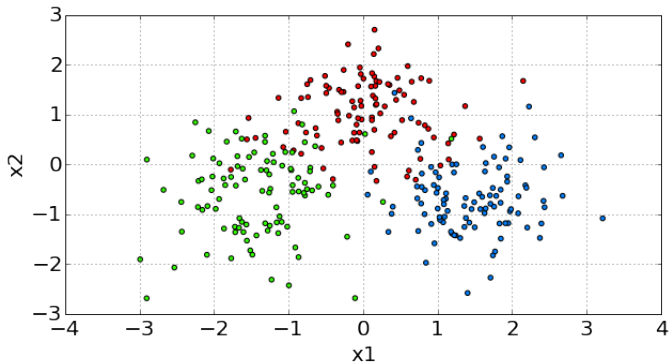
Types of target variable

- Types of target variable:
 - $\mathcal{Y} = \mathbb{R}$ - regression (in supervised learning)
 - $\mathcal{Y} = \mathbb{R}^M$ - vector regression (in supervised learning) or feature extraction (in unsupervised learning)
 - $\mathcal{Y} = \{\omega_1, \omega_2, \dots, \omega_C\}$ - classification (in supervised learning) or clustering (in unsupervised learning).
 - $C=2$: binary classification, encoding - $\mathcal{Y} = \{+1, -1\}$ or $\mathcal{Y} = \{0, 1\}$.
 - $C>2$: multiclass classification
 - \mathcal{Y} -set of all sets of $\{\omega_1, \omega_2, \dots, \omega_C\}$ - labeling
 - $\mathcal{Y} = \{y \in \mathbb{R}^C : y_i \in \{0, 1\}\}$, $y_i = 1 \Leftrightarrow$ object is associated with ω_i .

Types of features

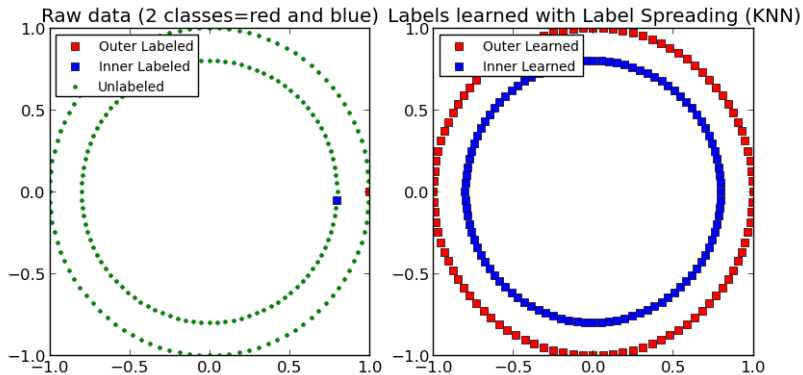
- Full object description $\mathbf{x} \in \mathcal{X}$ consists of individual features $x_i \in \mathcal{X}_i$
- Types of feature:
 - $\mathcal{X}_i = \{0, 1\}$ - binary feature
 - $|\mathcal{X}_i| < \infty$ - discrete (nominal) feature
 - $|\mathcal{X}_i| < \infty$ and \mathcal{X}_i is ordered - ordinal feature
 - $\mathcal{X}_i = \mathbb{R}$ - real feature

Example of classification



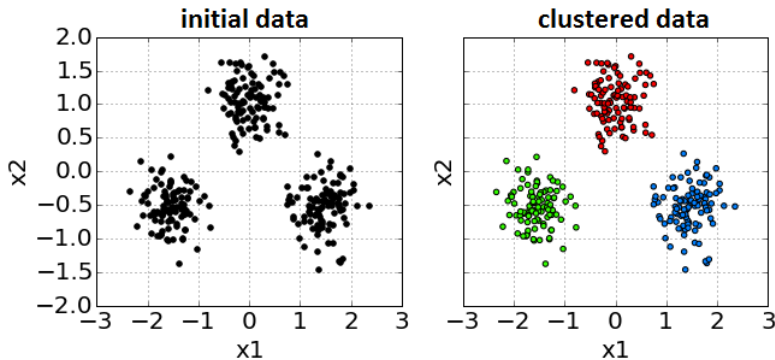
Supervised learning: $x = (x_1, x_2)$, y is shown with color

Example of semi-supervised learning



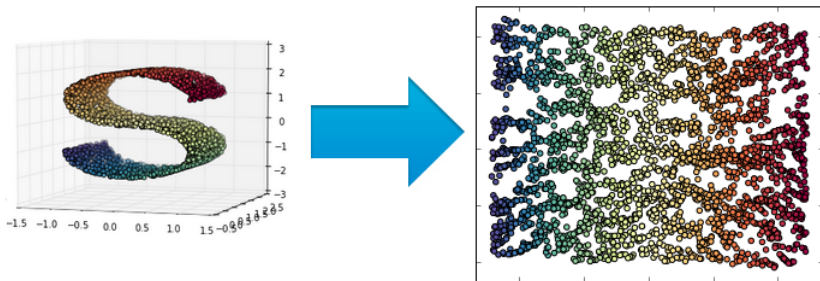
Semi-supervised learning.

Example of clustering



Unsupervised learning: clustering

Example of dimensionality reduction



Unsupervised learning: dimensionality reduction

Table of Contents

- 1 Tasks solved by machine learning
- 2 Main concepts of machine learning.
- 3 Practical applications of machine learning

Training set

- Training set: $X \in \mathbb{R}^{N \times D}$ - design matrix, $Y \in \mathbb{R}^N$ - predicted outputs (target values)
- Using X, Y the task is to estimate unknown parameters $\hat{\theta}$ of mapping $\hat{y} = f_{\theta}(x)$ so that it will approximate true relationship $y = y(x)$
- It is assumed that $z_n = (x_n, y_n)$ for $n = 1, 2, \dots, N$ - are independent and identically distributed random variables (i.i.d).
- Two steps of ML:
 - training
 - application

Loss function

- Loss function $\mathcal{L}(\hat{y}, y, x)$ is usually assumed to be $\mathcal{L}(\hat{y}, y)$
- Examples:
 - classification:

- misclassification rate

$$\mathcal{L}(\hat{y}, y) = \mathbb{I}[\hat{y} \neq y]$$

- regression:
 - MAE (mean absolute error):

$$\mathcal{L}(\hat{y}, y) = |\hat{y} - y|$$

- MSE (mean squared error):

$$\mathcal{L}(\hat{y}, y) = (\hat{y} - y)^2$$

- absolute relative error: $\frac{|\hat{y}-y|}{|y|}$, squared relative error: $\left(\frac{\hat{y}-y}{y}\right)^2$

Function class

- **Function class** - parametrized set of functions

$F = \{f_\theta, \theta \in \Theta\}$, from which the true relationship $\mathcal{X} \rightarrow \mathcal{Y}$ is approximated.

- Examples of linear class functions:
 - regression:

$$f(x) = \theta_0 + \theta_1 x^1 + \theta_2 x^2 + \dots + \theta_D x^D,$$

where x^i - is the i -th feature of x .

- two-class classification $y \in \{+1, -1\}$:

$$f(x) = \text{sign}\{\theta_0 + \theta_1 x^1 + \theta_2 x^2 + \dots + \theta_D x^D\},$$

where

$$\text{sign}(a) = \begin{cases} 1, & a \geq 0 \\ 0, & a < 0 \end{cases}$$

Empirical risk

- **Machine learning algorithm** associates $f_{\hat{\theta}}(\cdot)$ to (X, Y)
 - in the function class $F = \{f_{\theta}, \theta \in \Theta\}$
 - for given loss function $\mathcal{L}(\hat{y}, y)$
- **Empirical risk:**

$$L(\theta|X, Y) = \frac{1}{N} \sum_{n=1}^N \mathcal{L}(f_{\theta}(x_n), y_n)$$

- Method of empirical risk minimization:

$$\hat{\theta} = \arg \min_{\theta} L(\theta|X, Y)$$

Estimation of empirical risk

- Generally it holds that:

$$L(\hat{\theta}|X, Y) < L(\hat{\theta}|X', Y')$$

where X, Y is the training sample and X', Y' is the new data.

- $L(\hat{\theta}|X', Y')$ can be estimated using :
 - validation set (not used in optimization of $L(\theta|X, Y)$)
 - cross-validation
 - leave-one-out method

Levels of fitting

Underfitted model

Model that oversimplifies true relationship $\mathcal{X} \rightarrow \mathcal{Y}$.

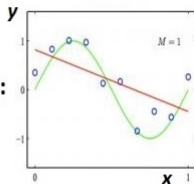
Overfitted model

Model that is too tuned on particular peculiarities (noise) of the training set instead of the true relationship $\mathcal{X} \rightarrow \mathcal{Y}$.

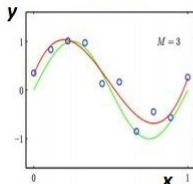
Examples of overfitted/underfitted models

- true relationship
- estimated relationship with polynimes of order M
- objects of the training sample

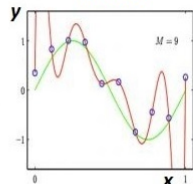
underfitted
model



relevant
model



overfitted
model



классификация:

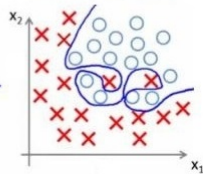
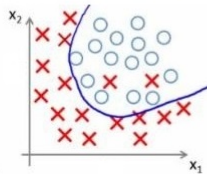
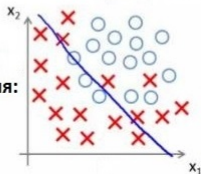


Table of Contents

- 1 Tasks solved by machine learning
- 2 Main concepts of machine learning.
- 3 Practical applications of machine learning

Examples of ML applications

Classification:

- spam filtering
- search engine: do query and document match each other?
- is series of network transactions regular or a hacking attempt?
- will the client with given characteristics switch his mobile operator?
- will given client of a bank return his debt?
- does the signal correspond to the target or noise in radar detection?

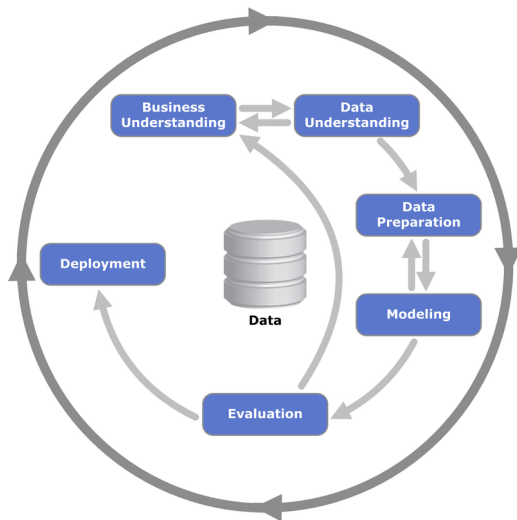
Labelling:

- assignment of topics to text documents

Регрессия:

- determine the flat price by its characteristics
- predict demand for certain product

CrispDM methodology



CrispDM general comments

- Log each step
 - quantitative: procedures and results in report.
 - qualitative: explain why certain option was taken and alternative options ignored.

CrispDM - Business understanding

- Understand business goals and constraints
- State business objective in business terms
- State relevant data mining objective in technical terms
- State success criteria
- Produce plan of project

CrispDM - Data understanding

- Collect data
- Understand data
 - qualitative meaning (what and how was measured)
 - quantitative distribution (data type, range, variance, skewness)
- Explore data
 - basic dependencies
 - interesting subsets
 - statistical analysis
- Quality check
 - outliers
 - missing data
 - errors in measurements

CrispDM - Data preparation

usually takes most of the time

- Select data (select datasets, records, attributes)
- Clean data
 - missing values
 - outliers
 - erroneous values
 - inconsistent groups of attributes
- Construct data
 - derive attributes (normalization, aggregation, composition)
 - use background knowledge
 - fill missing values
- Integrate data together into connected structures (e.g. joined tables)
- Format data (uppercase/lowercase, encoding, etc.)

CrispDM - Modeling

- Select relevant models
 - depending on data mining objective
 - depending on data properties (possibly need to return to data preparation)
- Divide dataset into training/validation/test sets
- Build models
 - choose initial values for model parameters
 - choose parameter estimation techniques
 - estimate parameters
 - post-process results using domain knowledge

CrispDM - Evaluation

- evaluate model output quality using technical data mining criteria
 - compare to baseline
 - reliability of results (statistical significance, dependence on specific data assumptions)
 - check for systematic errors and interpret them (may be caused by missed factors/constraints)
- evaluate resulting models (interpretability, efficiency, scalability)
- analyze final business effect

CrispDM - Deployment

- plan deployment
- plan monitoring and maintenance
- produce final report
- review project experience
 - from project team
 - from customers

Notation used in the course

- If this corresponds the context and there are no redefinitions, then:
 - x - vector of known input characteristics of an object
 - y - predicted target characteristics of an object specified by x
 - x_i - i -th object of a set, y_i - corresponding target characteristic
 - x^k - k -th feature of object specified by x
 - x_i^k - k -th feature of object specified by x_i
 - D - dimensionality of the feature space: $x \in \mathbb{R}^D$
 - N - the number of objects in the training set
 - X - design matrix, $X \in \mathbb{R}^{N \times D}$
 - $Y \in \mathbb{R}^N$ - target characteristics of a training set
 - $\mathcal{L}(\hat{y}, y)$ - loss function, where y is the true value and \hat{y} is the predicted value.
 - $\{\omega_1, \omega_2, \dots, \omega_C\}$ - possible classes, C - total number of classes.
 - \hat{z} defines an estimate of z , based on the training set: for example, $\hat{\theta}$ is the estimate of θ , \hat{y} is the estimate of y , etc.