Introduction into machine learning

November 5, 2016

Aleksey Kurov
kurov@pixsee.com

Course plan

1. Introduction into machine learning
2. Linear regression. Basics methods of ML.
3. SVM. Decision trees. Clustering. Dimension reduction.
4. Basics of neural networks.
5. Convolutional neural networks.
6. Recurrent neural networks. Natural language processing.
7. Reinforcement Learning.
8. Project presentations.

Useful resources

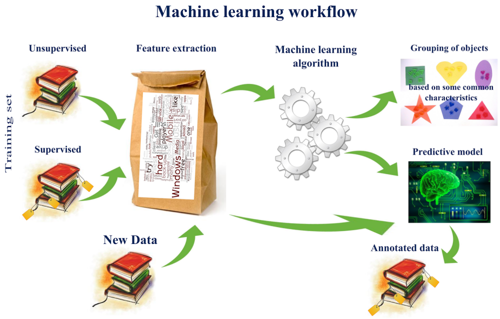
1. Coursera. Machine learning (Andrew Ng)
2. Coursera. Введение в машинное обучение (Воронцов).
3. Russsian ml resource - http://www.machinelearning.ru/
4. Slack - opendatascience; **samara-ml** - slack of current course
5. ML books - https://github.com/josephmisiti/awesome-machine-learning/blob/master/books.md
6. https://arxiv.org
7. https://vk.com/deeplearning
   https://vk.com/modeloverfit
   https://vk.com/datascience

Agenda

1. What is the Machine learning?
2. Types of machine learning.
    1. Supervised learning.
    2. Unsupervised learning.
    3. Semi-supervised learning.
    4. Reinforcement learning.
3. Machine learning techniques
    1. Classification.
    2. Regression.
    3. Clustering.
4. Basics of Linear algebra and Probability theory and Optimization

What is machine learning?

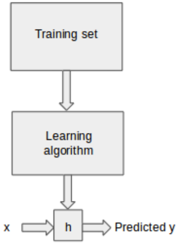
Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.



Supervised learning

Supervised learning (predictive model, "labeled" data) - learning some mapping from input data to output.

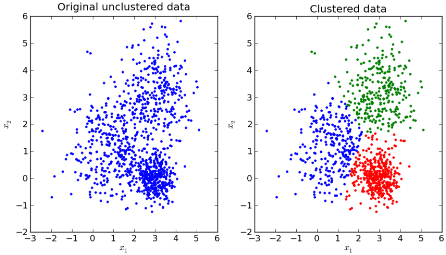
- Classification (Logistic Regression, Decision Tree, KNN, Random Forest, SVM, Naive Bayes, etc)
- Numeric prediction (Linear Regression, KNN, Gradient Boosting & AdaBoost, etc)



Unsupervised learning

Unsupervised learning (descriptive model, "unlabeled" data) - finds hidden data structure from unlabeled data.

- clustering (K-Means, Gaussian mixtures, DBSCAN and etc.)
- pattern discovery
- dimension reduction (PCA, linear discriminant analysis)



Semi-supervised learning

Semi-supervised learning is a class of supervised learning tasks and techniques that also make use of unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data.

- unlabeled data is cheap
- labeled data can be hard to get
- human annotation is boring
- labels may require experts

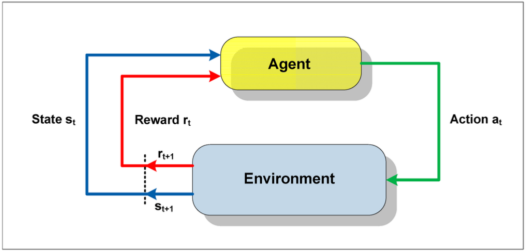
Semi-supervised algorithms

- Self Training
- Generative Models
- S3VMs
- Graph-Based Algorithms
- Multiview Algorithms

Reinforcement learning

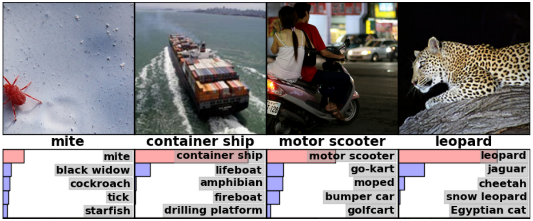
Reinforcement learning is learning what to do–how to map situations to actions–so as to maximize a numerical reward signal

- Rules are unknown
- No supervisor, only reward signal
- Agent's actions affect the subsequent data it receives



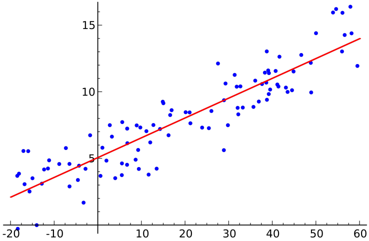
Classification

- Data is labeled. It means it is assigned a class, for example spam/non-spam or fraud/non-fraud.
- The decision being modelled is to assign labels to new unlabelled pieces of data.
- This can be thought of as a discrimination problem, modelling the differences or similarities between groups.
- Usually supervised



Regression

- Data is labelled with a real value (think floating point) rather than a label.
- Examples that are easy to understand are time series data like the price of a stock over time.
- The decision being modelled is what value to predict for new unpredicted data.

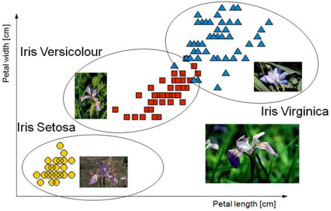


Clustering

Data is not labeled, but can be divided into groups based on similarity and other measures of natural structure in the data.

Algorithms

- Connectivity-based clustering (hierarchical clustering)
- Centroid-based clustering (k-means clustering)
- Density-based clustering (DBSCAN)



Basics of linear algebra

Matrix notation

$$A_{m,n} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix}$$

Matrix multiplication

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj},$$

where $A \in R^{m \times n}$ and $B \in R^{n \times p}$

Basics of linear algebra

Identity matrix

The identity matrix, denoted $I \in R^{n \times n}$ is a square matrix with ones on the diagonal and zeros everywhere else. That is,

$$I_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

Diagonal matrix

A diagonal matrix is a matrix where all non-diagonal elements are 0. This is typically denoted $\mathrm{D} = \mathrm{diag}(d_1, d_2, \ldots, d_n)$, with

$$D_{ij} = \begin{cases} d_i & i = j \\ 0 & i \neq j \end{cases}$$

Basics of linear algebra

Transpose matrix

The transpose of a matrix results from “flipping” the rows and columns. Given a matrix $A \in R^{m \times n}$, its transpose, written $A^T \in R^{n \times m}$, is the $n \times m$ matrix whose entries are given by

$$(A^T)_{ij} = A_{ji}$$

The following properties of transposes :

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $(A + B)^T = A^T + B^T$

Basics of linear algebra

Norm

A norm of a vector $\|x\|$ is informally a measure of the "length" of the vector. For example, we have the commonly-used Euclidean or ℓ_2 norm,

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

More formally, a norm is any function $f : R^n \rightarrow R :$ that satisfies 4 properties:

1. For all $x \in R^n, f(x) \geq 0$ (non-negativity).
2. $f(x) = 0$ if and only if $x = 0$ (definiteness).
3. For all $x \in R^n, t \in R, f(tx) = |t| f(x)$ (homogeneity).
4. For all $x, y \in R^n, f(x + y) \leq f(x) + f(y)$ (triangle inequality).

Basics of probability theory

Axioms of Probability

- **Sample space** Ω : The set of all the outcomes of a random experiment. Here, each outcome $\omega \in \Omega$ can be thought of as a complete description of the state of the real world at the end of the experiment.
- **Set of events (or event space)** F: A set whose elements $A \in F$ (called events) are subsets of Ω (i.e., $A \subseteq \Omega$ is a collection of possible outcomes of an experiment)
- **Probability measure**: A function $P : F \rightarrow R$ that satisfies the following properties,
  - $P(A) \geq 0$, for all $A \in F$
  - $P(\Omega) = 1$
  - If $A_1, A_2, \ldots$ are disjoint events (i.e., $A_i \cap A_j = \emptyset$ whenever $i \neq j$)

Basics of probability theory

Properties

- If $A \subseteq B \rightarrow P(A) \leq P(B)$
- $P(A \cap B) \leq min(P(A), P(B))$
- (Union Bound) $P(A \cup B) \leq P(A) + P(B)$
- $P(\Omega \setminus A) = 1 - P(A)$
- (Law of Total Probability) If $A_1, \ldots, A_k$ are a set of disjoint events such that $\cup_{i=1}^{k} A_i = \Omega$, then $\sum_{i=1}^{k} P(A_k) = 1$

Conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Basics of probability theory

Random variables

A **random variable** $X : \Omega \to E$ is a measurable function from the set of possible outcomes Ω to some set E. If X can take only a finite number of values, so it is known as a **discrete random variable**. Here, the probability of the set associated with a random variable X taking on some specific value k is

$$P(X = k) := P(\{\omega : X(\omega) = k\})$$

If $X(\omega)$ takes on a infinite number of possible values, so it is called a **continuous random variable**. We denote the probability that X takes on a value between two real constants a and b (where $a < b$) as

$$P(a \leq X \leq k) := P(\{\omega : a \leq X(\omega) \leq b\})$$

Basics of probability theory

Cumulative distribution functions

A **cumulative distribution function (CDF)** is a function $F_X : R \rightarrow [0, 1]$ which specifies a probability measure as,

$$F_X(x) = P(X \leq x)$$

Properties:

- $0 \leq F_X(x) \leq 1$
- $lim_{x \rightarrow -\infty} F_X(x) = 0$
- $lim_{x \rightarrow \infty} F_X(x) = 1$
- $x \leq y \rightarrow F_X(x) \leq F_X(y)$

Basics of probability theory

Probability mass functions

Probability mass function (PMF) is a function $p_X : \Omega \rightarrow R$ such that:

$$p_X(x) = P(X = x)$$

In the case of discrete random variable, we use the notation $Val(X)$ for the set of possible values that the random variable X may assume. For example, if $X(\omega)$ is a random variable indicating the number of heads out of ten tosses of coin, then $Val(X) = 0, 1, 2 \ldots, 10$
Properties:

- $0 \leq p_X(x) \leq 1$
- $\sum_{x \in Val(x)} p_X(x) = 1$
- $\sum_{x \in A)} p_X(x) = P(X \in A)$

Basics of probability theory

Probability density functions

We define the **Probability Density Function** or PDF as the derivative of the CDF

$$f_X(x) = \frac{dF_X(x)}{dx}$$

Properties:

- $f_X(x) \geq 0$
- $\int_{-\infty}^{\infty} f_X(x) = 1$
- $\int_{x \in A} f_X(x) dx = P(X \in A)$

Basics of probability theory

Expectation

IF X is a discrete random variable with PMF $p_X(x)$ and $g : R \to R$ is an arbitrary function **Expectation (mean)** or **expected value** of $g(X)$ is

$$E[g(X)] = \sum_{x \in X} g(x) p_X(x)$$

If X is a continuous random variable with PDF $f_X(x)$, then the expected value of $g(x)$ is defined as,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Basics of probability theory

Expectation

Properties:

- $E[a] = a$ for any constant $a \in R$
- $E[af(X)] = aE[f(X)]$ for any constant $a \in R$
- Linearity of Expectation
  $E[f(X) + g(X)] = E[f(X)] + E[g(X)]$

Basics of probability theory

Variance

The **variance** of a random variable X is a measure of how concentrated the distribution of a random variable X is around its mean.

$$Var[X] = E[(X - E(X))^2]$$

Properties:

- $E[(X - E(X))^2] = E[X^2] - E[X]^2$
- $Var[a] = 0$ for any constant $a \in R$
- $Var[af(X)] = a^2 Var[f(X)]$ for any constant $a \in R$

Basics of probability theory

Discrete random variables

- $X \sim \mathit{Bernoulli}(p)$ (where $0 \leq p \leq 1$): one if a coin with heads probability p comes up heads, zero otherwise.

$$p(x) = \begin{cases} p & \text{if } p = 1 \\ 1 - p & \text{if } p = 0 \end{cases}$$

- $X \sim \mathit{Binomial}(n, p)$ (where $0 \leq p \leq 1$): the number of heads in n independent flips of a coin with heads probability p.

$$p(x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

Basics of probability theory

Discrete random variables

- $X \sim \mathit{Geometric}(p)$ (where $p > 0$): the number of flips of a coin with heads probability p until the first heads.

$$p(x) = p(1-p)^{x-1}$$

- $X \sim \mathit{Poisson}(\lambda)$ (where $\lambda > 0$):a probability distribution over the nonnegative integers used for modeling the frequency of rare events.

$$p(x) = e^{\lambda} \frac{\lambda^x}{x!}$$

Basics of probability theory

Continuous random variables

- $X \sim \mathit{Uniform}(a, b)$ (where $a < b$): equal probability density to every value between a and b on the real line

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

- $X \sim \mathit{Exponential}(\lambda)$ (where $\lambda > 0$):decaying probability density over the nonnegative reals.

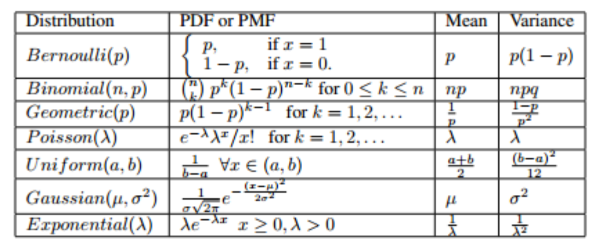
$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Basics of probability theory

Continuous random variables

- $X \sim \mathit{Normal}(\mu, \sigma^2)$ also known as the Gaussian distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n,p)$	$\binom{n}{k} p^k(1-p)^{n-k}$ for $0 \leq k \leq n$	np	npq
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\lambda^x/x!$ for $k = 1, 2, \ldots$	λ	λ
$Uniform(a,b)$	$\frac{1}{b-a}$ $\forall x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ $x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$