Machine Learning and Data Mining

29. April, 2021

Introduction and motivation: Machine Learning and Data Mining

- We live in the era of big data.
- Artificial Intelligence (AI)
- the vast range of applications
- theoretical frameworks in machine learning (supervised and unsupervised learning, statistical learning theory, etc.)

Buzzword Confusion

- Machine Learning is a subfield of AI, there is lots of non-learning AI
- Deep Learning is a subfield of Machine Learning, there is lots of non-deep (and/or non-neural) learning

Machine Learning definition

"How do we create computer programs that improve with experience?", Tom Mitchell
http://videolectures.net/mlas06_mitchell_itm/

Tom Mitchell. Machine Learning 1997.

"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 at tasks in T, as measured by P, improves with experience E.

Data types

Data types

Texts, Numbers, Tables, Images, Transactions, Videos

profusion of applications

- Detecting faces in images, Spam filtering, Digit recognition on checks, zip codes, Recommendation system, etc.
- All big IT companies I Google, Microsoft, Apple, Amazon, Netflix, etc. have large and prolific ML research groups
- for large and small consultancy firms
- recommender systems: Amazon wants to sell you more stuff; Netflix wants to keep you a happy viewer; Youtube, Facebook, Instagram want to keep you on their site
- Subgroup Discovery. What are my best customers? Whom should I target as my new customer?

Types of machine learning

The different types of machine learning problems

- Predictive or supervised learning approach
- Descriptive or unsupervised learning
- Reinforcement learning

supervised learning

- D is called the training set,
- N is the number of training examples
- the goal is to learn a mapping from inputs x to outputs y, given a labeled set of input-output pairs $D = \{(x_i, y_i)\}_{i=1}^{N}$
- $y_i \in \{1, ..., C\}$

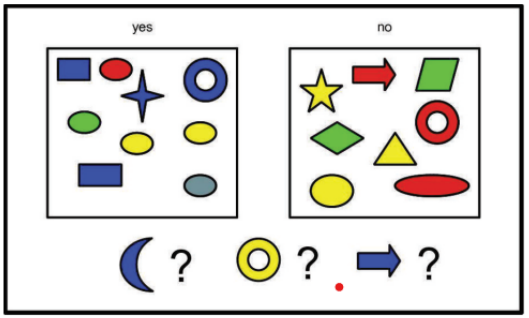
descriptive or unsupervised learning

- only given inputs $D = \{x_i\}_{i=1}^{N}$
- the goal is to find interesting patterns in the data.

Supervised learning

Supervised learning

- the goal is to learn a mapping from inputs x to outputs y
- $y \in \{1, ..., C\}$
- If $C = 2$ it is binary classification, if $C > 2$ it is multiclass classification
- we assume $y = f(x)$ for some unknown function f
- the goal of learning is to estimate the function f given a labeled training set
- to derive $\hat{y} = \hat{f}(x)$ (the hat symbol is usually used to denote an estimate.)



(a)

D features (attributes)

N cases

Color	Shape	Size (cm)	Label
Blue	Square	10	1
Red	Ellipse	2.4	1
Red	Ellipse	20.7	0

(b)

Figure 1.1 Left: Some labeled training examples of colored shapes, along with 3 unlabeled test cases. Right: Representing the training data as an $N \times D$ design matrix. Row i represents the feature vector $\mathbf{x}_i$. The last column is the label, $y_i \in \{0, 1\}$. Based on a figure by Leslie Kaelbling.

Figure 1: This Figure is Figure 1.1. in book: Machine Learning: A Probabilistic Perspective, K. Murphey

The need for probabilistic predictions

- $p(A)$ denotes the probability that the event A is true.
- $0 \leq p(A) \leq 1$
- $p(\bar{A}) = 1 - p(A)$

discrete random variable X

- the state space $\mathbb{X}$
- $p(X = x)$
- $0 \leq p(x) \leq 1$
- $\sum_{x \in X} p(x) = 1$

Fundamental rules

Given two events A, B

- $p(A \cup B) = p(A) + p(B) - p(A \cap B)$
- $p(A, B) = p(A \cap B) = p(A|B)p(B)$ the product rule

Given a joint distribution on two events $p(A, B)$, the marginal distribution is defined as

- $p(A) = \sum_b p(A, B = b) = \sum_b p(A|B = b)p(B = b)$
- the sum rule or the rule of total probability

the chain rule of probability

$p(X_1, ...X_M) =$
$p(X_1)p(X_2|X_1)p(X_3|X_1, X_2) \cdots p(X_M|X_1, X_2, ..., X_{M-1})$

The conditional probability

$p(A|B) = p(A, B)/p(B)$

- $$p(Activity =' lakeside'|Weather =' Hot') = \frac{p(Activity =' lakeside' \wedge Weather =' Hot')}{p(Weather =' Hot')}$$

Figure 2: Example of the conditional probability

Bayes rule

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)}$$

Bayes rule

$$p(A = a \wedge B = b) = p(B = b \wedge A = a)$$

$$p(A = a|B = b) \times p(B = b) = p(B = b|A = a) \times p(A = a)$$

$$p(A = a|B = b) = \frac{p(B = b|A = a) \times p(A = a)}{p(B = b)}$$

$$p(A = a|B = b) = \frac{p(B = b|A = a) \times p(A = a)}{\sum_{a' \in A} p(B = b|A = a') \times p(A = a')}$$

Figure 3: Bayes rule

The need for probabilistic predictions

- the probability distribution over possible labels, given the input vector x and training set $\mathcal{D}$ by $p(y|x, \mathcal{D})$

$$\hat{y} = \hat{f}(\mathbf{x}) = \underset{c=1}{\overset{C}{\operatorname{argmax}}}\, p(y = c|\mathbf{x}, \mathcal{D})$$

Figure 4: Our "best guess"

Real world application of classification

Classification

- Probably the most widely used form of machine learning
- Various real world applications

Example: Classifying flower 1/2

Figure 1.3 Three types of iris flowers: setosa, versicolor and virginica. Source: `http://www.statlab.uni-heidelberg.de/data/iris/` . Used with kind permission of Dennis Kramb and SIGNA.

Figure 5: This Figure is Figure 1.3. in book: Machine Learning: A Probabilistic Perspective, K. Murphey

Example: Classifying flower 2/2

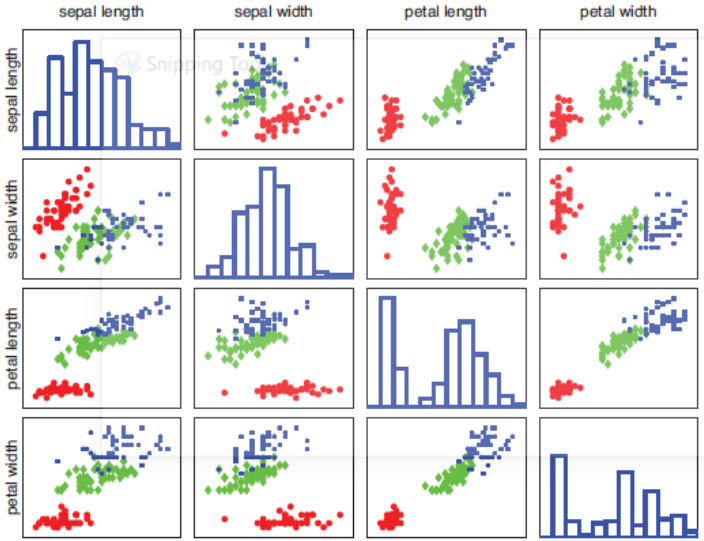


Figure 1.4 Visualization of the Iris data as a pairwise scatter plot. The diagonal plots the marginal histograms of the 4 features. The off diagonals contain scatterplots of all possible pairs of features. Red circle = setosa, green diamond = versicolor, blue star = virginica. Figure generated by `fisheririsDemo`.

Figure 6: This Figure is Figure 1.4. in book: Machine Learning: A Probabilistic Perspective, K. Murphey

Example:

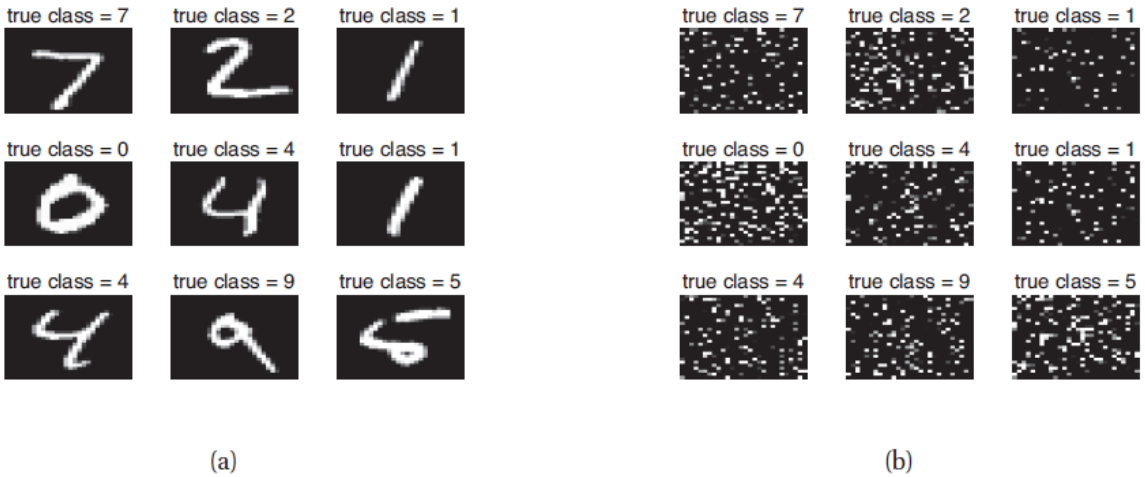


Figure 1.5 (a) First 9 test MNIST gray-scale images. (b) Same as (a), but with the features permuted randomly. Classification performance is identical on both versions of the data (assuming the training data is permuted in an identical way). Figure generated by `shuffledDigitsDemo`.

Figure 7: This Figure is Figure 1.5. in book: Machine Learning: A Probabilistic Perspective, K. Murphey

Example: Face detection and recognition

(a) (b)

Figure 1.6 Example of face detection. (a) Input image (Murphy family, photo taken 5 August 2010). Used with kind permission of Bernard Diedrich of Sherwood Studios. (b) Output of classifier, which detected 5 faces at different poses. This was produced using the online demo at `http://demo.pittpatt.com/`. The classifier was trained on 1000s of manually labeled images of faces and non-faces, and then was applied to a dense set of overlapping patches in the test image. Only the patches whose probability of containing a face was sufficiently high were returned. Used with kind permission of Pittpatt.com

Figure 8: This Figure is Figure 1.6. in book: Machine Learning: A Probabilistic Perspective, K. Murphey

Regression

- the response variable is continuous
- $x \in R$, $y \in R$

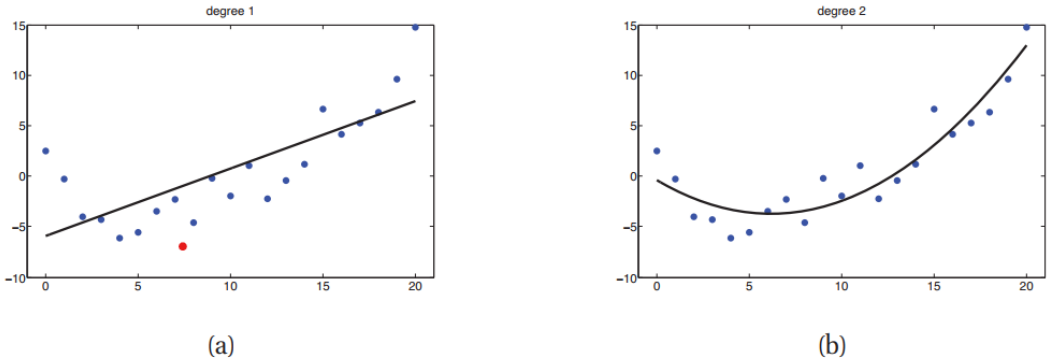


Figure 1.7 (a) Linear regression on some 1d data. (b) Same data with polynomial regression (degree 2). Figure generated by `linregPolyVsDegree`.

Figure 9: This Figure is Figure 1.7. in book: Machine Learning: A Probabilistic Perspective, K. Murphey

real-world regression problems

- Predict tomorrow's stock market price
- Predict the age of a viewer watching a given video on YouTube
- Predict the temperature at any location inside a building using weather data, time, door sensors

number of parameters

Some models come with a fixed number of parameters and the others parameter sets grow with the amount of training data. Therefore we distinguish two elementary types of model:

- parametric model - a model which comes with the fixed amount of parameters.
- non-parametric model - a model which parameters grow with the amount of data in the training set D.

Basic Data Mining Tasks

- Classification: Assign each data element into the one of the predefined classes (e.g. Spam e-mail detection, etc.)
- Regression: map item into the real valued prediction variable (e.g. predict a value of savings in future)
- Clustering: similar to classification but the groups are not predefined (e.g. group people into communities from a given social network)

Task 1 Seminar presentation (11. May 2021.)

Seminar presentation

20min

interactive discussion

real-world Application

article/paper from scientific journal