Data Science Lecture Notes 07

Donghui Yan

University of Massachusetts Dartmouth

Outline

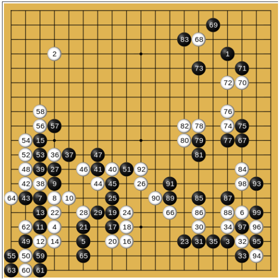
- Introduction
- Formalism and terminology
- Evaluation methodology

Machine learning in real life

- Search engine design
    - To max chance one gets what he searches in top K entries
- Computational advertising
    - Placement of ads to maximize profit
- Design of e-commerce web site
    - Selection of selling items to max click thru rate (or profit)
- Selection of headline news
    - e.g., which news as headline in news portal at Yahoo, CNN etc
- Object recognition
    - OCT hand digits recognition by USPS
- House (pool) cleaning robot.

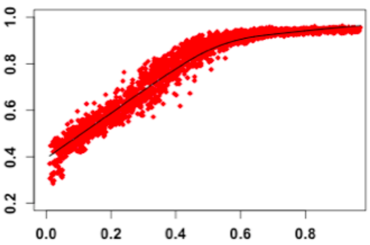
State of the art in AI

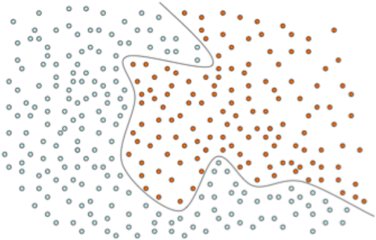
- AlphaGo
  - Beat Ke Jie (ranked #1 in world) 3:0 in 2017
  - Major milestone in AI research
- Self-driving
- Conversation robot.



What is machine learning (ML)?

Definition. Machine learning refers to application, methodology, and theory relevant to the automatic learning of *patterns or regularities* from data.





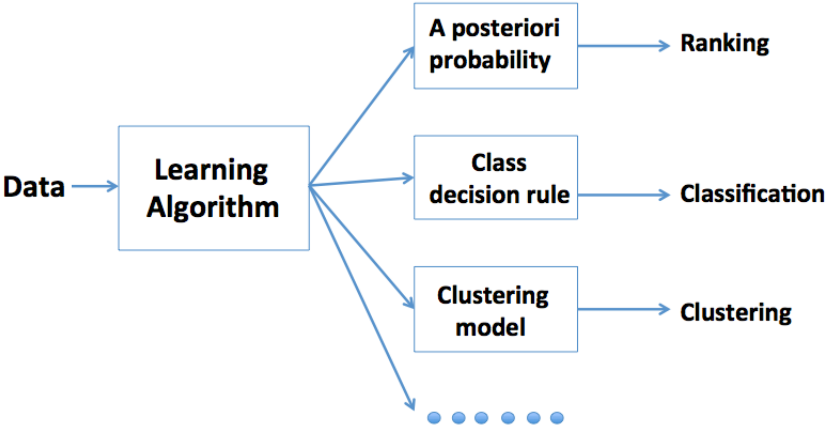
Two important *assumptions* in Machine learning

- The *future* is related to the *past*
  - The phenomenon is stationary, or the past and future drawn from the same probability distribution
- Knowledge about the problem under study
  - Generalization only possible when knowledge is encoded
  - Features being the most elementary form.

Problems in machine learning

- Classification
  - $Y \in \mathcal{C} = \{c_1, c_2, ..., c_k\}$, called labels
- Clustering
  - Y not given (often called unsupervised learning)
- Regression
  - $Y \in \mathcal{R}$, called response
- Ranking
- And a lot more new topics emerging in recent years
  - Topic model (e.g., what is the topic of a blogger article)
  - Manifold (topological) learning
  - Salient sentence extraction
  - Graph learning etc.

Problems in machine learning



A little history about the evolving of machine learning

- Early days
  - The AI approach
    - 1956 Dartmouth conference marks the start of AI
    - Perceptron (Rosenblatt, 1957)
    - Dying of the research on Neural network late 1960's
    - Various induction machine, expert system, fuzzy system etc
    - PAC learning (Valiant 1984)
  - The statistical approach
    - Statistical learning theory (Vapnik and Chervonenkis, 1964-1974)
    - Fisher's LDA, logistic regression, k-means, mixture analysis etc
    - Early nonparametric statistics (e.g., kNN)
- The revitalization of Neural network in mid 1980's
- SVM, boosting, Random Forests from early 1990's and on
  - The statistical approach is gaining popularity
- Neural network back again under guise of *deep learning* (2008-).

Connections to other subjects

- Aspects of machine learning
  - Machine
    - Computer science (algorithms to realize machine learning)
  - Learning
    - Mathematics, probability and statistics (analysis and theory)
  - Applications
    - Provides motivation and ultimate testbed for learning algorithms.

- What's the connection of (nonparametric) statistics and ML?
  - Both learn from the data
  - Nonparametric statistics $\subseteq$ ML (by my definition)
  - But, as a matter of fact, ML focus more on discrete problems (e.g., classification) while (nonparametric) statistics more on the continuous world (e.g., regression).

Predictive learning or classification

- Given data $(X_1, Y_1), ..., (X_n, Y_n)$, we wish to learn the relationship $f : X \mapsto Y$ s.t.

The future prediction is the best

e.g., smallest error rate (or precision/recall, AUC etc)

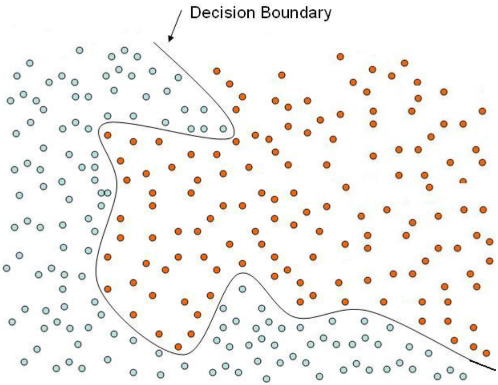
  - X_i called features, $Y_i \in \{1, 2, ..., J\}$ called *labels or classes*
  - $(X_1, Y_1), ..., (X_n, Y_n)$ is called a *training sample*
  - f is called the *trained or fitted model*

♠ *The best possible decision rule* (Bayes rule)

  - As if one knows the distribution (X, Y) (often unknown).

The classification problem

- What does the solution to a classification problem really do?
  - Identifying the decision (or class) boundary.



Loss function

Depends on the application, typical loss functions

- The 0-1 loss

$$cost = \begin{cases} 1, & \text{if } f(X) \neq Y \\ 0, & \text{otherwise.} \end{cases}$$

  - A loss function of special interest and most commonly used

- Cost-sensitive loss functions, i.e., a cost matrix, for $a \neq b$,

$$\begin{bmatrix} 0 & b \\ a & 0 \end{bmatrix}$$

  - Suitable when errors in diff classes have diff consequence
    - e.g., fraud detection, cost a when mistaking fraud as normal and b when mistaking normal as fraud.

Function class

Function class $\mathcal{F} = \{f\}$ determines the type of classifiers

- Linear classifiers
  - Logistic regression: $logit(P(Y|X)) = \boldsymbol{X\beta}$
  - SVM: $f(X) = \sum_{i=1}^{n} w_i K(X_i, X) + w_0$
- Boosting
  - $f(X) = \sum_{i=1}^{T(n)} a_i h(X_1, ..., X_n, X)$

  with h from some data dependent basis library
- Tree-based classifiers
  - C4.5, CART
  - Random Forests and its variants.

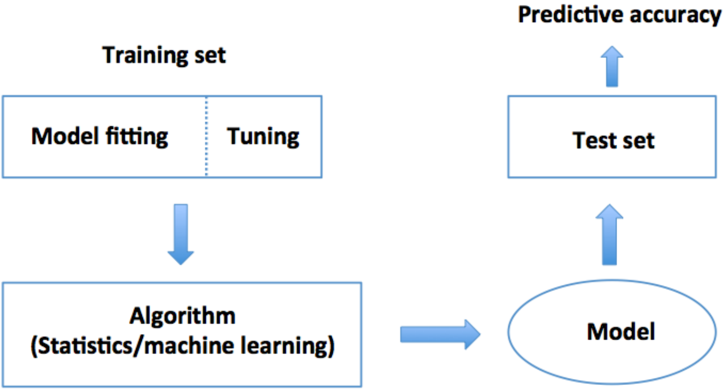
Evaluation of a machine learning algorithm

- Train on training data and evaluate on test data
  - Most common
- Cross-validation
  - Split data into J partitions
  - Use any one of the J partitions as test and rest for training
  - Average result on all J tests
- Bootstrap and use out of bag estimate
  - Train on a sample with replacement of all observations in the data
  - Test on the rest
  - Repeat many times and average results.

Evaluation methodology

- Have separate training/test set
- Fit a model (e.g., logistic model) on the training set
- Evaluate the trained model on the test set
  - Correct classification when the label matches (0/1 loss)
  - More advanced metric like AUC.

Evaluation illustrated



Selection of tuning parameters

- Treat the training set as the entire data
- Split *training data* = *training set* + *tuning set*
- Treat tuning set → test set and proceed as usual evaluation
  - Calculate a performance metric, e.g., accuracy
  - Select parameters that lead to the best performance
  - Use selected parameters for final performance evaluation.

Performance metrics

Some popular performance metrics

- Error rate
  - Most commonly used in statistics and machine learning
- Kappa statistics
  - Commonly used in remote sensing, medical assessment
- Area under curve (AUC)
  - When detection and false alarm rate matter, e.g., biomarker discovery, anomaly/fraud detection.

Kappa statistic

- The idea is to measure the amount of departure from that arises purely by chance
  - κ is calculated from the confusion matrix
  - κ takes into account sample sizes for different classes
  - Controversial (many modifications going on).

The confusion matrix

Label	1	...	j	...	C	Total
1	n_{11}	...	n_{1j}	...	n_{1C}	$n_{1.}$
...	...	...	...	...	...	...
i	n_{i1}	...	n_{ij}	...	n_{iC}	$n_{i.}$
...	...	...	...	...	...	...
C	n_{C1}	...	n_{Cj}	...	n_{CC}	$n_{C.}$
Total	$n_{.1}$	...	$n_{.j}$	...	$n_{.C}$	n

- C = # classes
- n_{ij} = # points from class i but classified as j
- n = size of the sample.

Calculation of κ

Definition. The κ coefficient is calculated as

$$\kappa = \frac{P_{observed} - P_{expected}}{1 - P_{expected}}$$

where

$$P_{expected} = \sum_{i=1}^{C} \frac{n_{i.}}{n} \cdot \frac{n_{.i}}{n}, \qquad P_{observed} = \frac{1}{n} \sum_{i=1}^{C} n_{ii}.$$

- $P_{expected}$ measures chance that observed and true labels agree
- $P_{observed}$ measures proportion of observations labeled correctly.

Example κ statistics

Confusion matrix of Logit on the South Africa Heart data

True/Predicted	1	2	Total
1	130	41	171
2	22	38	60
Total	152	79	231

- $n_{1.} = 171,\ n_{2.} = 60,\ n_{.1} = 151,\ n_{.2} = 79,\ n = 231$
- $P_{expected} = n_{1.}n_{.1}/n + n_{2.}n_{.2}/n = 171 \cdot 152/231^2 + 60 \cdot 79/231^2 = 0.5759$
- $P_{observed} = (n_{11} + n_{22})/n = (130 + 38)/231 = 0.7273$
- $\kappa = (P_{observed} - P_{expected})/(1 - P_{expected}) = (0.7273 - 0.5759)/(1 - 0.5759) = 0.36.$

Area under curve

- ROC curve is a graphical plot of true positive rate (TPR) vs. false positive rate (FPR) as discrimination threshold varies
  - TPR = % true positives out of the positives
  - FPR = % false positives out of the negatives
- Example (assume class 1 = positive, 2 = negative)

True/Predicted	1	2	Total
1	130 (true pos)	41 (false neg)	171 (pos)
2	22 (false pos)	38 (true neg)	60 (neg)
Total	152	79	231

Area under curve (AUC)

- Another measure to assess a machine learning algorithm
- Often used when cost for mis-classification is asymmetric
  - e.g., intrusion as normal Vs normal as intrusion in cyber security
- R package "AUC".