Topics in Learning Theory

Lecture 1: Introduction

Outline of the Lectures

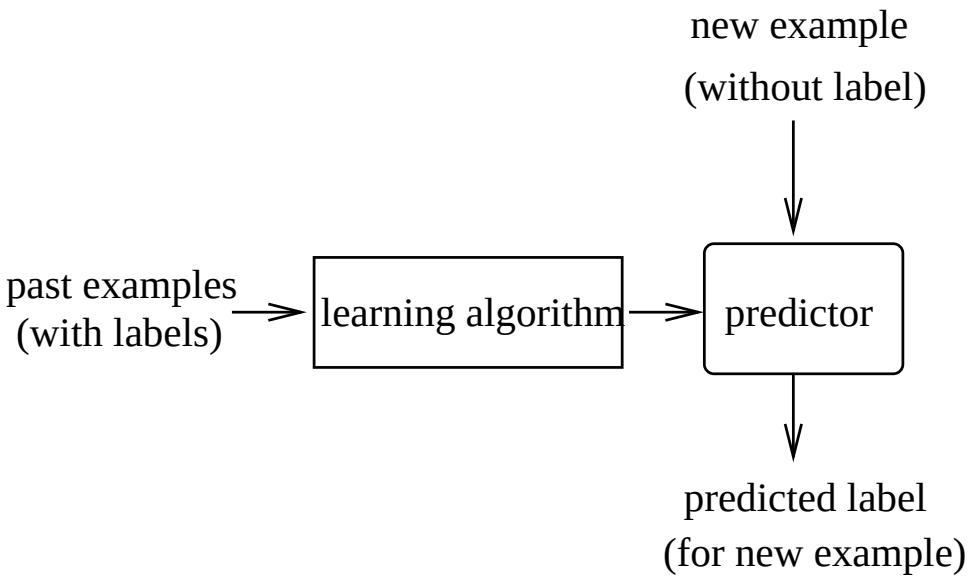
1. Introduction (what is machine learning and learning theory)
2. Generalization Error/Covering Numbers/Hoeffding Inequality
3. Concentration Inequalities
4. Binary Classification
5. Regularization
6. Loss Functions
7. Kernel Methods (I)
8. Kernel Methods (II)
9. Boosting
10. Brief Overview of Other Research Topics

What is machine learning

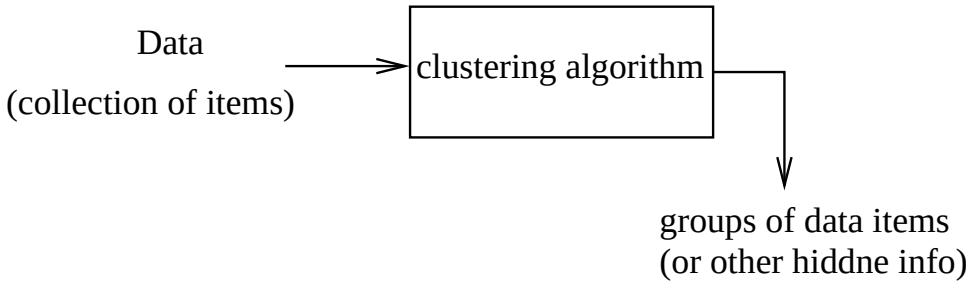
- Methods to automatically discover patterns or structures in data
  - e.g. more car accidents when it snows
- Methods to predict missing (unobserved) information using observation
  - will housing price go up or down?

Topics in machine learning

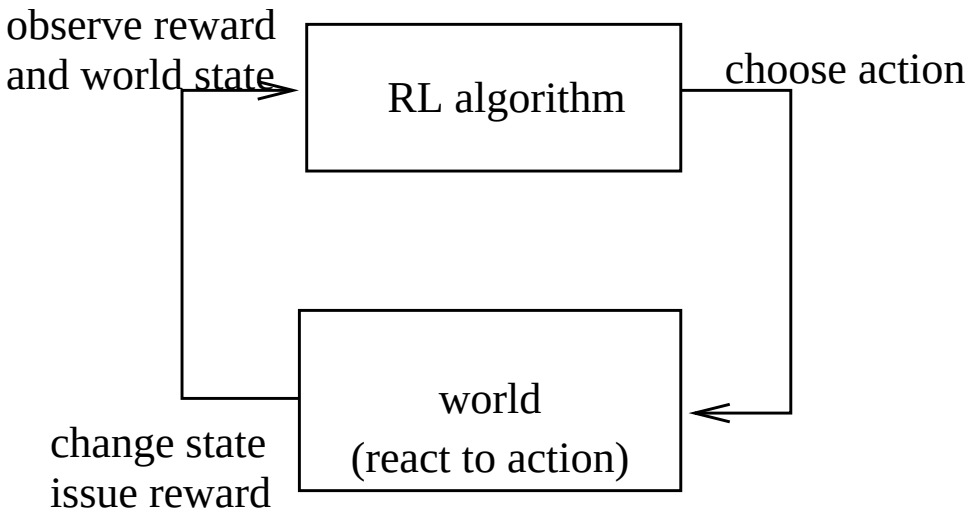
- Supervised learning
  - predict unknown information from known information.
  - learning prediction rule based on past examples



- Unsupervised learning
  - clustering and grouping data, anomaly detection, etc.



- Reinforcement learning (RL):
  - learning optimal action given state from feedbacks to maximize rewards.
  - delayed reward and exploration/exploitation.
  - robot exploration and mapping of unknown place; chess playing, etc.



Example supervised learning problem

Gender	Systolic BP	Weight	...	Disease Code
M	175	65	...	3
F	141	72	...	1
...	...	...	...	...
F	160	59	...	2

Figure 1: A Spreadsheet Example of Medical Data (patient records)

- Prediction problem: disease code (e.g. the odds that the person has heart disease) for a new patient given other columns

- Learning from known examples: want to find rules to predict disease code given existing patient records with known disease code

Example unsupervised learning problem

church, bishop, prayer
church, prayer, islam
stock, price, bear, market
stock, price, analyst, company
...

Figure 2: A hypothetical collection of news articles containing above key words

- Question: how to make sense out of them.
  - Do they have any structures, patterns that can help people understand these documents better?
- Clustering: partition documents into distinct groups so that documents within each groups are similar, and documents across groups are different

- ideally: each group has a certain semantic meaning that is consistent with human understanding
- religion, finance, sport ...

Example reinforcement learning problem

- Example: training a robot to roam around and avoid trouble based on sensor inputs:
  - inputs: camera, laser, sonar, touch, etc.
  - action: direction to go at the next step
  - feedback: whether the current position is a trouble spot (e.g. water, mud, etc)
- the goal is to learn to take action based on inputs that can avoid trouble

Other topics and issue

- inference: graphical models (build model and relationship of objects in the world)
- active learning: minimize the number of human labels for supervised learning
- semi-supervised learning: how to take advantage of unlabeled data.
- learning with structures: complicated models
  - complicated loss/prediction problem
  - complicated model structure, regularization
- scalability and optimization methods

- association rules: find pattern or correlation among data attributes
- data visualization: embedding of high dimension data into low-dimension
- summarization and interpretation
- evaluation

Supervised learning

- Prediction problem
  - Input X: known information.
  - Output (label) Y: unknown information.
  - Goal: to predict Y based on X
- Observe historical data $(X_1, Y_1), \ldots, (X_n, Y_n)$
- Find a prediction rule f that can map future observation X to the associated Y.

The Machine Learning Approach

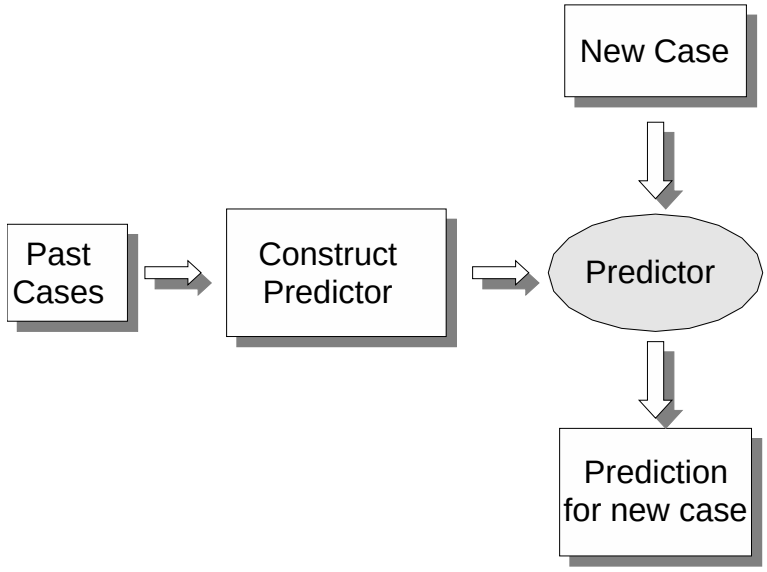


Figure 3: Predicting the Future Based on the Past

- past cases: training data (known input and known output)
- new cases: test data (known input and unknown output)

Components in supervised learning

- Output Y: encoding of desired information
  - regression: continuous real number (stock price tomorrow)
  - classification: discrete value (whether stock is up or down)
  - vector and complicated output: ranking, machine translation, speech recognition, etc.
- Input X: encoding of known information (features)
  - need to have predictive power (correlation to output Y).
  - often in vector representation: containing both categorical and numerical attributes.
- Prediction rule f: the learning algorithm assumes a certain form of f (linear, kernel, decision tree, neural network), and find a rule that fits the data.

Measuring output quality

- Truth output y
- Predicted output f
- Quality measured by loss function $\phi(f, y)$.
  - Regression: $\phi(f, y) = (f - y)^2$
  - Classification: $\phi(f, y) = I(f \neq y)$
  - Applications often require specific criteria.

Supervised Learning example: Email spam

- User send/receive emails from a computer server
- Spam email: unwanted email sent to multiple people
  - typically containing commercial advertising for dubious products, get-rich-quick schemes, etc.
- Computer server determines which email the user receives is a “spam”
  - Binary classification problem: whether an email is a spam email or not
- Actions: throw email away, or put into a “spam” folder.
  - require degrees of confidence: throw email away only when extremely confident about prediction

Example Feature Vector for Email Spam Detection

text:title			text:body			nontext	prediction target
...	cheap	...	enlargement	...	ink	from known spam host	spam
...	yes	...	yes	...	yes	yes	true
...	no	...	yes	...	no	yes	true
...	no	...	no	...	no	no	false
...	...	...	...	...	...	...	...

- Need to encode email into a feature vector: columns are called features
  - spreadsheet representation (encode unstructured text into structured data)
  - bag-of-word binary feature representation of email text
  - using known spam host as nontext features

Digit recognition: feature vector

- Input: images
- Output: digits 0-9 which the image represents
- Feature vector:
  - each pixel is a feature component (column).
  - value is whether the pixel is active or not.

Example: web-search

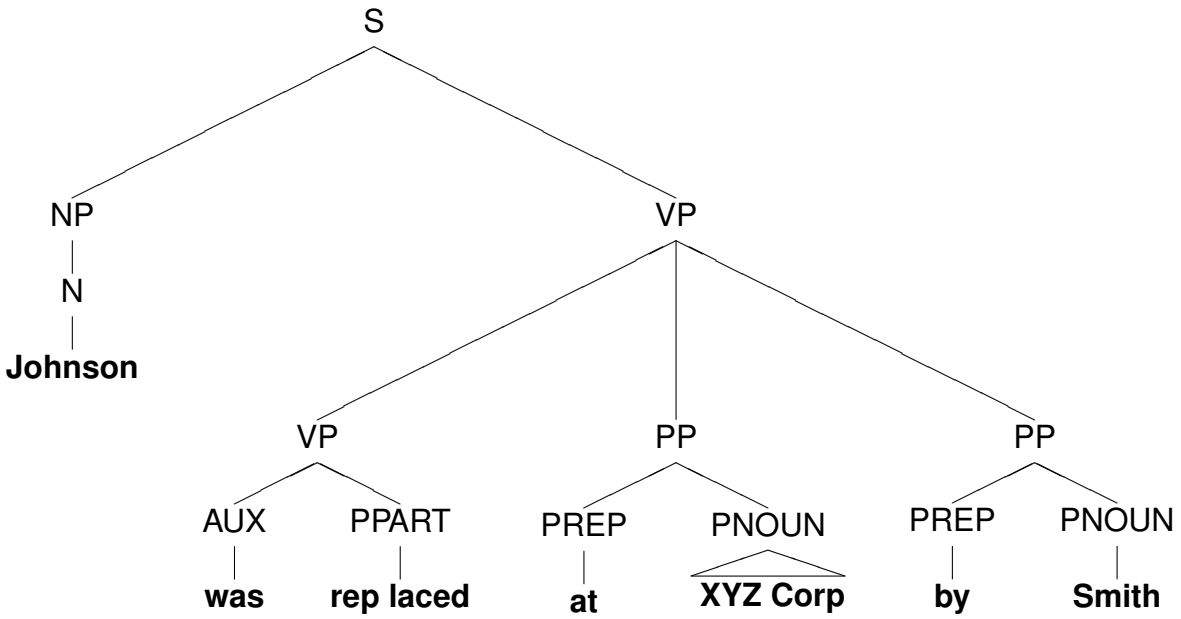
- User types a query, search engine returns a result page:
  - selects from billions of pages.
  - assign a quality score for each page, and return pages ranked by the scores.

- Prediction problem: predict relevant score for page p given query q.

- Training:
  - randomly select queries q, and web-pages p for each query.
  - use editorial judgment to assign relevance grade $y(p, q)$.
  - construct a feature $x(p, q)$ for each query/page pair.
  - learn scoring function $f(x(p, q))$ to preserve the order of $y(p, q)$ for each q.

Complicated output prediction: statistical parsing

- **Input**: English sentence; **output**: parse tree



Machine learning approach

- Decompose parse tree generation into simpler steps
  - bottom up: group consecutive chunks
- Create linguistic features that are most predictive of the best grouping.
- Output quality: matched chunks.

Summary of Prediction Approach

- Can solve many real-world problems
- Steps:
  - Formulate the problem as prediction problem
    - fill unknown information from known information
  - Design quality evaluation criterion
  - Incorporate prior knowledge in feature generation
  - Using standard learning algorithms as black boxes with given input features to construct prediction rule
    - or use prior information to construct problem specific learning algorithm

Standard model for supervised learning

- Data (X, Y) are randomly drawn from an underlying distribution D.
  - Y may not be deterministic given X.
- Assume training data are iid samples from D: $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- Want to construct prediction function f from training data to minimize future loss over D
  - $\mathbf{E}_{(X,Y)\sim D}\phi(f(X), Y)$.

Some Issues in applying prediction methods

- Testing data differs from training data
  - data change through time
  - training data available in one domain (news article) but test data in another domain (e.g. web-pages)
  - sampling of training data is biased
- How to learn with small number of data and adapt under domain change
  - construct prior from unlabeled data
  - learn or construct task-invariant representation from related tasks
- Discovering spurious patterns

Overfitting

- The goal: predict well on unseen data
- Two aspects:
  - rule should fit the data well
    - requires a more expressive model.
  - behavior of rule on test data should match that on training data
    - requires a less expressive (more stable) model.
- Related concepts:
  - training versus test error, bias variance trade-off, overfitting, model complexity, generalization performance, regularization

Example of Overfitting

- Given training data (X_i, Y_i) $(i = 1, \ldots, n)$
  - assume X_i are all different
- The following prediction rule fits data perfectly:

$$f(X) = \begin{cases} Y_i & \text{if } X = X_i \text{ for some } i \\ 0 & \text{otherwise} \end{cases}$$

- Have no prediction capability when X not in the training data.

Learning Theory

- Estimate generalization performance (prediction accuracy on unseen data) of learning algorithms.

- Quantify the degree of overfitting and derive rigorous estimates.