Challenges in Machine Learning and Data Mining

Tu Bao Ho, JAIST

Based on materials from

1. 9 challenges in ML (Caruana & Joachims)
2. 10 challenging problems in DM (Yang & Wu)

What is machine learning?

- The goal of machine learning is to build computer systems that can adapt and learn from their experience (Tom Dietterich).
- 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 measure by **P**, improves with experience (Tom Mitchell book, p. 2).
- ML problems can be formulated as
  - Given: (x_1, y_1), (x_2, y_2), …, (x_n, y_n)
    - x_i is description of an object, phenomenon, etc.
    - y_i is some property of x_i, if not available learning is unsupervised
  - Find: a function f(x) that $f(x_i) = y_i$

Finding hypothesis f in a huge hypothesis space F by narrowing the search with constraints (bias)

Overview of ML challenges

1. Generative vs. discriminative learning
2. Learning from non-vectorial data
3. Beyond classification and regression
4. Distributed data mining
5. Machine learning bottlenecks
6. Intelligible models
7. Combining learning methods
8. Unsupervised learning comes of age
9. More informed information access

1. Generative vs. discriminative methods

Training classifiers involves estimating **f: X → Y**, or **P(Y|X)**.
Examples: P(apple | red ∧ round), P(noun | "cá")

Generative classifiers	Discriminative classifiers
Assume some functional form for P(X|Y), P(Y)	Assume some functional form for P(Y|X)
Estimate parameters of **P(X|Y), P(Y)** directly from training data, and use Bayes rule to calculate **P(Y|X = x_i)**	Estimate parameters of **P(Y|X)** directly from training data
HMM, Markov random fields, Bayesian networks, Gaussians, Naïve Bayes, etc.	SVM, logistic regression, traditional neural networks, nearest neighbors, boosting, MEMM, conditional random fields, etc.

(cá: fish, to bet)

1. Generative vs. discriminative methods

Training classifiers involves estimating **f: X → Y**, or **P(Y|X)**.
Examples: P(apple | red ∧ round), P(noun | "cá")

Generative classifiers

- Assume some functional form for P(X|Y), P(Y)
- Estimate parameters of **P(X|Y), P(Y)** directly from training data, and use Bayes rule to calculate **P(Y|X = x_i)**
- HMM, Markov random fields, Bayesian networks, Gaussians, Naïve Bayes, etc.

Discriminative classifiers

- Assume some functional form for P(Y|X)
- Estimate parameters of **P(Y|X)**

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

$$P(apple \mid red \wedge round) = \frac{P(red \wedge round \mid apple)P(apple)}{P(red \wedge round)}$$

(cá: fish, to bet)

Generative vs. discriminative methods

Generative approach

- Try to build models for the underlying patterns
- Can be learned, adapted, and generalized with small data.

Discriminative approach

- Try to learn to minimize an utility function (e.g. classification error) but not to model, represent, or "understand" the pattern explicitly (detect 99.99% faces in real images and do not "know" that a face has two eyes).
- Often need large training data, say 100,000 labeled examples, and can hardly be generalized.

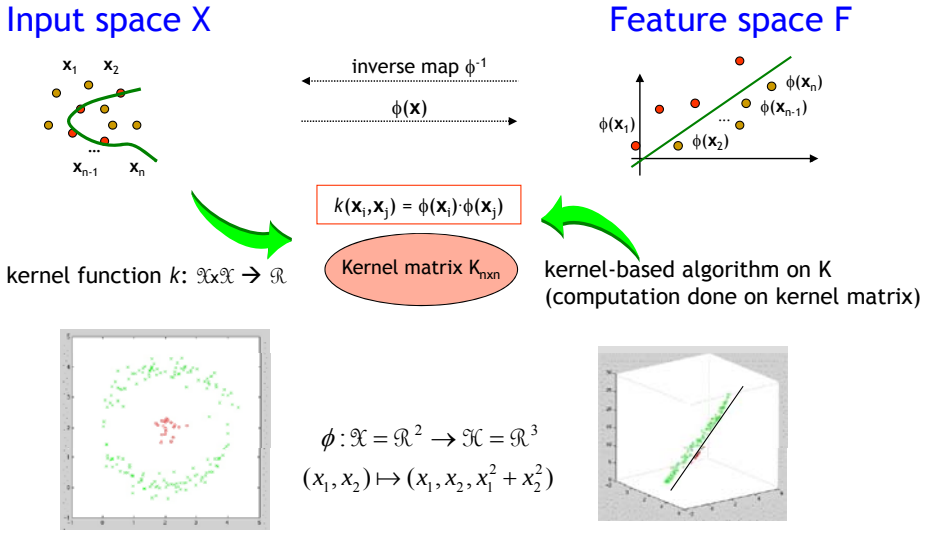
Generative vs. discriminative learning

- **Objective**: determine which is better for what, and why
- **Current**:
  - Discriminative learning (ANN, DT, KNN, SVM) typically more accurate
    - Better with larger data
    - Faster to train
  - Generative learning (graphical models, HMM) typically more flexible
    - More complex problems
    - More flexible predictions
- **Key Challenges**:
  - Vapnik: "When solving a problem, don't first solve a harder problem"
  - Making generative methods computationally more feasible
  - When to prefer discriminative vs. generative

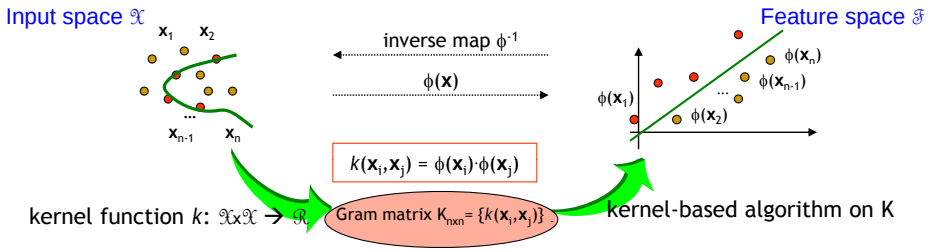
2. Kernel methods

- Objective: learning from non-vectorial input data
- Current:
  - Most learning algorithms work on flat, fixed length feature vectors
  - Each new data type requires a new learning algorithm
  - Difficult to handle strings, gene/protein sequences, natural language parse trees, graph structures, pictures, plots, …
- Key Challenges:
  - One data-interface for multiple learning methods
  - One learning method for multiple data types
- Research already in progress

Kernel methods: the basic ideas



Kernel methods: math background



Linear algebra, probability/statistics, functional analysis, optimization

- **Mercer theorem**: Any positive definite function can be written as an inner product in some feature space.
- **Kernel trick**: Using kernel matrix instead of inner product in the feature space.
- **Representer theorem**:

Every minimizer of $\min_{f\in\mathfrak{X}}\{C(f,\{x_i,y_i\}) + \Omega(\|f\|_H)$ admits a representation of the form $f(.) = \sum_{i=1}^{m}\alpha_i K(.,x_i)$

3. Beyond classification and regression

- **Objective**: learning to predict complex objects
- **Current**:
  - Most machine learning focuses on classification and regression
  - Discriminative methods often outperform generative methods
  - Generative methods used for learning complex objects (e.g. language parsing, protein sequence alignment, information extraction)
- **Key Challenges**:
  - Extend discriminative methods (ANN, DT, KNN, SVM, …) to more general settings
  - Examples: ranking functions (e.g. Google top-ten, ROC), natural language parsing, finite-state models
  - Find ways to directly optimize desired performance criteria (e.g. ranking performance vs. error rate)

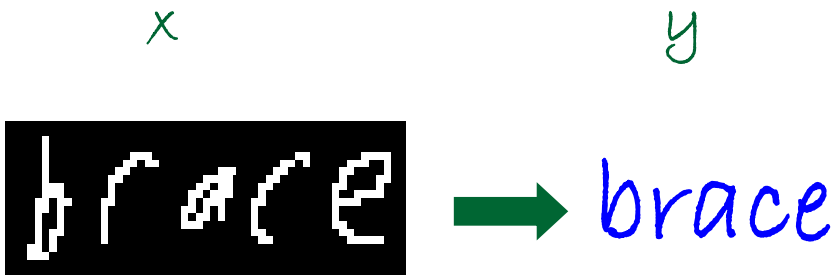
What is structured prediction? (Daume)

- Structured prediction is a fundamental machine learning task involving classification or regression in which the output variables are mutually dependent or constrained.
- Such dependencies and constraints reflect sequential, spatial or combinatorial structure in the problem domain, and capturing these interactions is often as important for the purposes of prediction as capturing input-output dependencies.
- Structured prediction (SP) – the machine learning task of generating outputs with complex internal structure .

What is structured prediction? (Lafferty)

- Text, sound, event logs, biological, handwriting, gene networks, linked data structures like the Web can be viewed as graphs connecting basic data elements-.
- Important tasks involving structured data require the computation of a labeling for the nodes or the edges of the underlying graph. E.g., POS tagging of natural language text can be seen as the labeling of nodes representing the successive words with linguistic labels.
- A good labeling will depend not just on individual nodes but also the contents and labels of nearby nodes, that is, the preceding and following words--thus, the labels are not independent.

Handwriting recognition

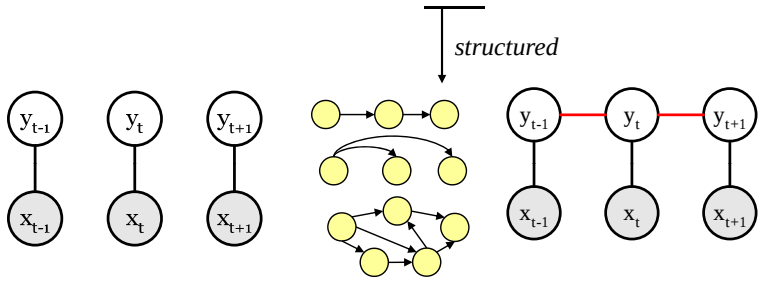


Sequential structure

Structured prediction

- Structured Learning / Structured Prediction

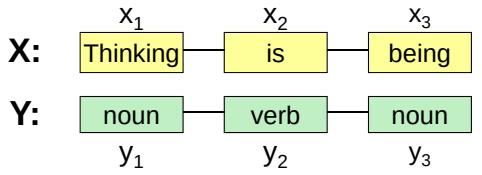
$$h : \mathcal{X} \rightarrow \mathcal{Y}$$



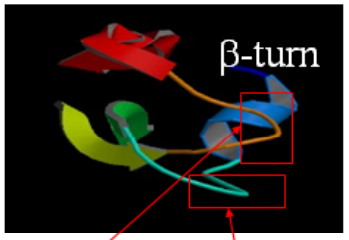
(a) Unstructured Model (b) Linear-Structured Model

Labeling sequence data problem

- X is a random variable over data sequences
- Y is a random variable over label sequences whose labels are assumed to range over a finite label alphabet A
- Problem: Learn how to give labels from a closed set Y to a data sequence X



- POS tagging, phrase types, etc. (NLP),
- Named entity recognition (IE)
- Modeling protein sequences (CB)
- Image segmentation, object recognition (PR)
- etc.



X KARIIRYFYNAKAGLCQTFCRAKRNNFKSAED
Y nnnnnnnnnTTtttnnnnnnnnTtttnnnnnn

4. Distributed learning

- Objective: DM/ML with distributed data
- Current:
  - Most ML algorithms assume random access to all data
  - Often data comes from decentralized sources (e. g. sensor networks, multiple organizations, learning across firewalls, different security systems)
  - Many projects infeasible (e.g. organization not allowed to share data)
- Key Challenges:
  - Develop methods for distributing data while preserving privacy
  - Develop methods for distributed learning without distributing the data

5. Full auto: ML for the masses

- **Objective**: make ML easier to apply to real problems
- **Current:**
  - ML applications require detailed knowledge about the algs
  - Preparing/Preprocessing takes at least 75% of the effort
- **Key Challenges:**
  - Automatic selection of machine learning method
  - Tools for preprocessing the data
    - Reformatting, Sampling, Filling in missing values, Outlier detection
  - Robust performance estimation and model selection
  - "Data Scoup"

6. Interpretable models

- **Objective**: make learning results more understandable
- **Current:**
  - Methods often achieve good prediction accuracy
  - The prediction rule appears complex & is difficult to verify
  - Lack of trust in the rule
  - Lack of insight
- **Key Challenges:**
  - Machine learning methods that are understandable & generate accurate rules
  - Methods for generating explanations
  - Model verification for user acceptance

7. Ensemble methods

- **Objective**: combining learning methods automatically
- **Current:**
  - We do not have a single DM/ML method that "does it all"
  - Results indicate that combining models results in large improvements in performance
  - Focus on boosting and bagging
- **Key Challenges:**
  - Develop methods that combine the best of different learning algs
  - Searching for good combinations might be more efficient than designing one "global" learning algorithm
  - Theoretical explanation for why and when ensemble methods help

Selective ensemble

LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn

Many Could be Better Than All: When a number of learners are available, … …, ensembling **many** of the available learners may be better than ensembling **all** of them

[Z.-H. Zhou et al., IJCAI'01 & AIJ02]

The learner k to be excluded from the ensemble satisfies:

$$(2N-1)\sum_{i=1}^{N}\sum_{j=1}^{N}C_{ij} \le 2N^2\sum_{\substack{i=1\\ i\neq k}}^{N}C_{ik} + N^2E_k \quad \text{in classification}$$

or

$$\sum_{\substack{j=1\\ j\in\{j\|Sum_j|\le 1\}}}^{m} Sgn\left(\left(Sum_j + f_{kj}\right)d_j\right) \le 0 \quad \text{in regression}$$

8. Unsupervised learning

- **Objective:** improve state-of-the-art in unsupervised learning
- **Current:**
  - Research focus in 90's was supervised learning
  - Much progress on supervised learning methods like neural networks, support vector machines, boosting, etc.
  - Unsupervised learning needs to "catch up"
- **Key Challenges:**
  - More robust and stable methods for clustering
  - Include user feedback into unsupervised learning (e.g. clustering with constraints, semi-supervised learning, transduction)
  - Automatic distance metric learning
  - Clustering as an interactive data analysis process

9. Information access

- **Objective:** more informed information access
- **Current:**
  - Bag-of-words
  - Retrieval functions exploit document structure and link structure
  - Information retrieval is a process without memory
- **Key Challenges:**
  - Develop methods for exploiting usage data
  - Learn from the query history of a user / user group
  - Preserving privacy while mining access patterns
  - Exploiting common access patterns and finding "groups" of users
  - Web Expert: agent that learns the web (beyond Google)
  - Topic modelling

What is data mining?

"Data-driven discovery of models and patterns from massive observational data sets"

Statistics, Inference | **Languages, Representations** | **Data Management**

Applications

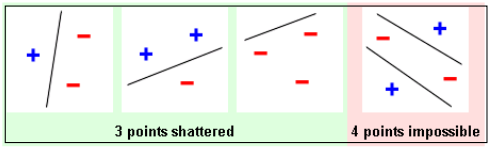
Overview of DM challenges (ICDM'05)

1. Developing a Unifying Theory of Data Mining
2. Scaling Up for High Dimensional Data/High Speed Streams
3. Mining Sequence Data and Time Series Data
4. Mining Complex Knowledge from Complex Data
5. Data Mining in a Network Setting
6. Distributed Data Mining and Mining Multi-agent Data
7. Data Mining for Biological and Environmental Problems
8. Data-Mining-Process Related Problems
9. Security, Privacy and Data Integrity
10. Dealing with Non-static, Unbalanced and Cost-sensitive Data

1. Developing a unifying theory of DM

- The current state of the art of data-mining research is too "ad-hoc"
  - techniques are designed for individual problems
  - no unifying theory
- Needs unifying research
  - Exploration vs explanation
- Long standing theoretical issues
  - How to avoid spurious correlations?
- Deep research
  - Knowledge discovery on hidden causes?
  - Similar to discovery of Newton's Law?

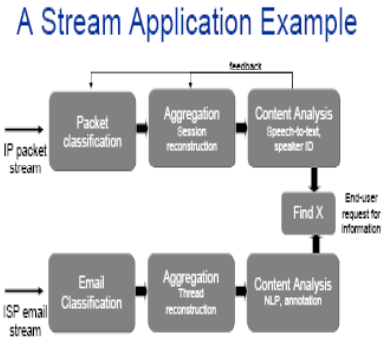
- **Example: VC dimension.** In statistical learning theory, or sometimes computational learning theory, the VC dimension (for Vapnik-Chervonenkis dimension) is a measure of the capacity of a statistical classification algorithm, defined as the cardinality of the largest set of points that the algorithm can shatter.



VC dimension of perceptron is 3.

2. Scaling up for high dimensional data and high speed streams

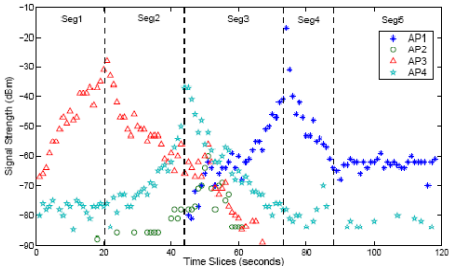
- Scaling up is needed
  - ultra-high dimensional classification problems (millions or billions of features, e.g., bio data)
  - Ultra-high speed data streams
- Streams
  - continuous, online process
  - e.g. how to monitor network packets for intruders?
  - concept drift and environment drift?
  - RFID network and sensor network data



Excerpt from Jian Pei's Tutorial
http://www.cs.sfu.ca/~jpei/

3. Sequential and time series data

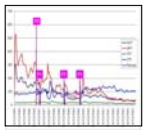
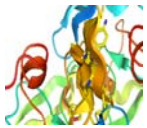
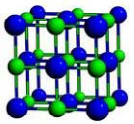
- How to efficiently and accurately cluster, classify and predict the trends?
- Time series data used for predictions are contaminated by noise
  - How to do accurate short-term and long-term predictions?
  - Signal processing techniques introduce lags in the filtered data, which reduces accuracy
  - Key in source selection, domain knowledge in rules, and optimization methods

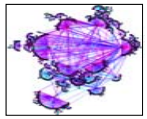


Real time series data obtained from Wireless sensors in Hong Kong UST CS department hallway

4. Mining complex knowledge from complex data (complexly structured data)

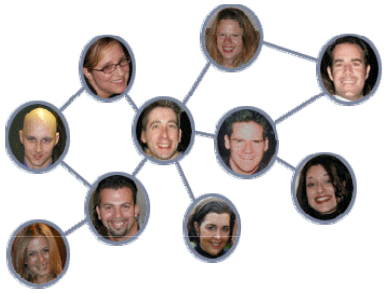
- Mining graphs
- Data that are not i.i.d. (independent and identically distributed)
  - many objects are not independent of each other, and are not of a single type.
  - mine the rich structure of relations among objects,
  - E.g.: interlinked Web pages, social networks, metabolic networks in the cell
- Integration of data mining and knowledge inference
  - The biggest gap: unable to relate the results of mining to the real-world decisions they affect - all they can do is hand the results back to the user.
- More research on interestingness of knowledge



5. Data mining in a network setting

- Community and Social Networks
  - Linked data between emails, Web pages, blogs, citations, sequences and people
  - Static and dynamic structural behavior
- Mining in and for Computer Networks
  - detect anomalies (e.g., sudden traffic spikes due to a DoS (Denial of Service) attacks
  - Need to handle 10Gig Ethernet links (a) detect (b) trace back (c) drop packet



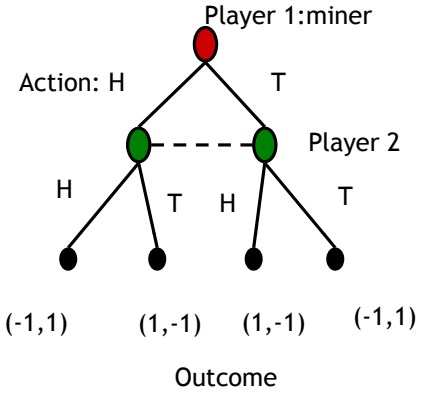
Picture from Matthew Pirretti's slides,penn state

An Example of packet streams (data courtesy of NCSA, UIUC)

6. Distributed data mining and mining multi-agent data

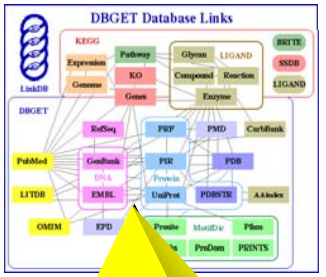
- Need to correlate the data seen at the various probes (such as in a sensor network)
- Adversary data mining: deliberately manipulate the data to sabotage them (e.g., make them produce false negatives)
- Game theory may be needed for help

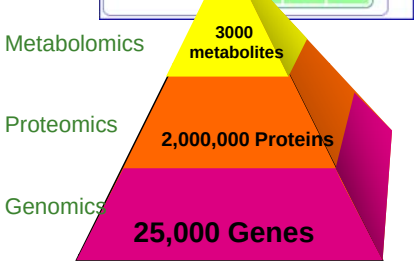
- Games



7. Data mining for biological and environmental problems

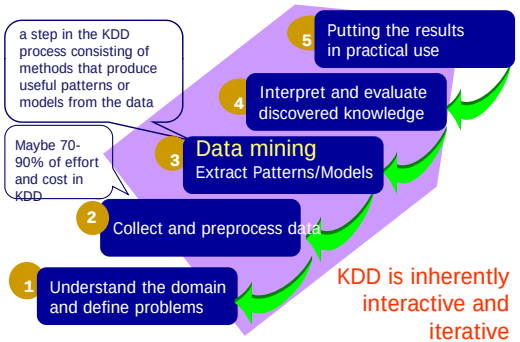
- New problems raise new questions
- Large scale problems especially so
  - Biological data mining, such as HIV vaccine design
  - DNA, chemical properties, 3D structures, and functional properties → need to be fused
  - Environmental data mining
  - Mining for solving the energy crisis





8. Data-mining-process related problems

- How to automate mining process?
  - the composition of data mining operations
  - Data cleaning, with logging capabilities
  - Visualization and mining automation
- Need a methodology: help users avoid many data mining mistakes
  - What is a canonical set of data mining operations

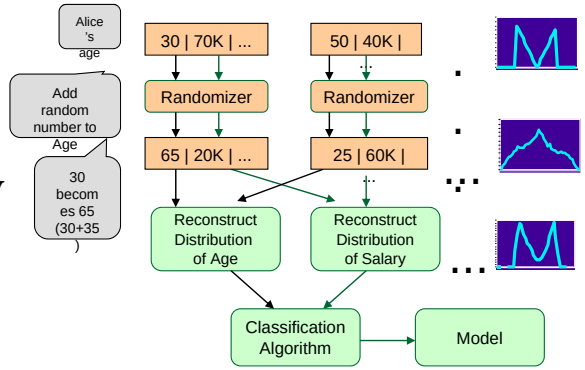


(In many cases, viewed KDD as data mining)

9. Security, privacy and data integrity

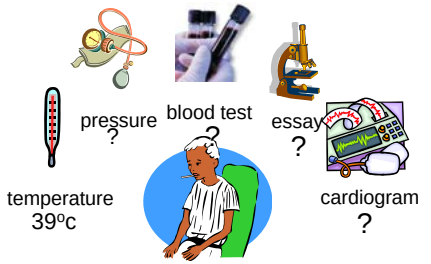
- How to ensure the users privacy while their data are being mined?
- How to do data mining for protection of security and privacy?
- Knowledge integrity assessment

- Perturbation Methods
- Secure Multi-Party Computation (SMC) Methods



10. Dealing with non-static, unbalanced and cost-sensitive data

- The UCI datasets are small and not highly unbalanced
- Real world data are large (10^5 features) but only < 1% of the useful classes (+'ve)
- There is much information on costs and benefits, but no overall model of profit and loss
- Data may evolve with a bias introduced by sampling



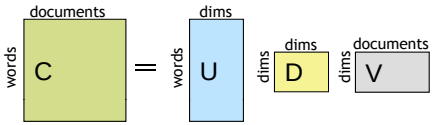
- Each test incurs a cost
- Data extremely unbalanced
- Data change with time

- Some papers can be found here
  http://www.jaist.ac.jp/~bao/K417-2008

Để máy hiểu được nghĩa các văn bản?

- Tìm tài liệu trên Google liên quan 3 chủ đề "thực phẩm", "mắm tôm", "dịch bệnh".
- Google cho ra rất nhiều tài liệu, với precision và recall thấp.
- Làm sao máy tính hiểu được nội dung văn bản để tìm kiếm cho hiệu quả?
- Thông qua chủ đề của văn bản
- Latent semantic analysis (Deerwester et al., 1990; Hofmann, 1999): Biểu diễn văn bản trong một không gian Euclid, mỗi chiều là một tổ hợp tuyến tính các từ (giống PCA).

Latent semantic analysis



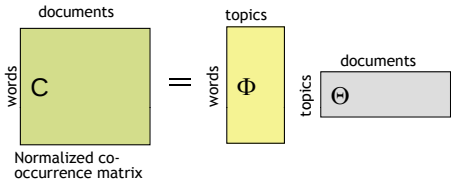
	D1	D2	D3	D4	D5	D6	Q1
rock	2	1	0	2	0	1	1
granite	1	0	1	0	0	0	0
marble	1	2	0	0	0	0	1
music	0	0	0	1	2	0	0
song	0	0	0	1	0	2	0
band	0	0	0	0	1	0	0

	D1	D2	D3	D4	D5	D6	Q1
dim1	-0.888	-0.759	-0.615	-0.961	-0.388	-0.851	-0.845
dim2	0.460	0.652	0.789	-0.276	-0.922	-0.525	0.534

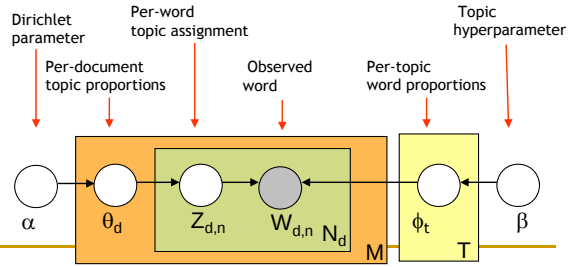
Topic modeling: key ideas

- Topic modeling key idea (LDA, Blei, JMLR 2004)
  - mỗi văn bản là một mixture của các chủ đề
  - mỗi chủ đề là một phân bố xác suất trên các từ.
- Thí dụ
  - "thực phẩm" = {an toàn, rau, thịt, cá, không ngộ độc, không đau bụng …}
  - "mắm tôm" = {tôm, mặn, đậu phụ, thịt chó, lòng lợn, …}
  - "dịch bệnh" = {nhiều người, cấp cứu, bệnh viện, thuốc, vacine, mùa hè, … }
  - D1= {thực phẩm 0.6, mắm tôm 0.35, dịch bệnh 0.8}

Topic modeling



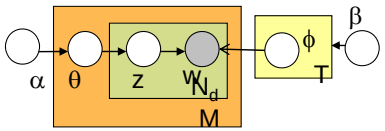
Latent Dirichlet Allocation (LDA)



Latent Dirichlet allocation (LDA) model

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}$$

Dirichlet prior on the per-document topic distributions



$$p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta) = p(\theta|\alpha)\prod_{n=1}^{N}p(z_n|\theta)p(w_n|z_n,\beta)$$

Joint distribution of topic mixture θ, a set of N topic z, a set of N words w

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}p(z_n|\theta)p(w_n|z_n,\beta)\right)d^k\theta$$

Marginal distribution of a document by integrating over θ and summing over z

$$p(D|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}}p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d^k\theta_d$$

Probability of collection by product of marginal probabilities of single documents

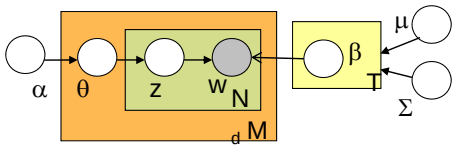
Example of topics learned

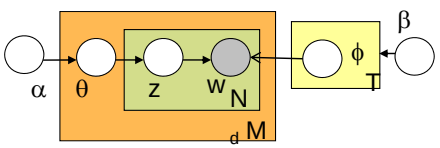
- From 16000 documents of AP corpus → 100-topic LDA model.
- Each color codes a different factor from which the word is putatively generated

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Dirichlet-Lognormal (DLN) topic model





Model descrption:

$$\theta \mid \alpha \sim Dirichlet(\alpha)$$
$$\beta \mid \mu, \Sigma \sim Lognormal(\mu, \Sigma)$$
$$z \mid \theta \sim Multinomial(\theta)$$
$$w \mid \beta \sim Multinormial(f(\beta))$$

$$\Pr(x_1,...,x_n) = \frac{1}{(2\pi)^{n/2}.\sqrt{\Sigma}.x_1...x_n} \exp\left\{-\frac{1}{2}(\log x-\mu)^T \Sigma^{-1}(\log x-\mu)\right\}$$
$$where \log x = (\log x_1,...,\log x_n)^T$$

Spam classification

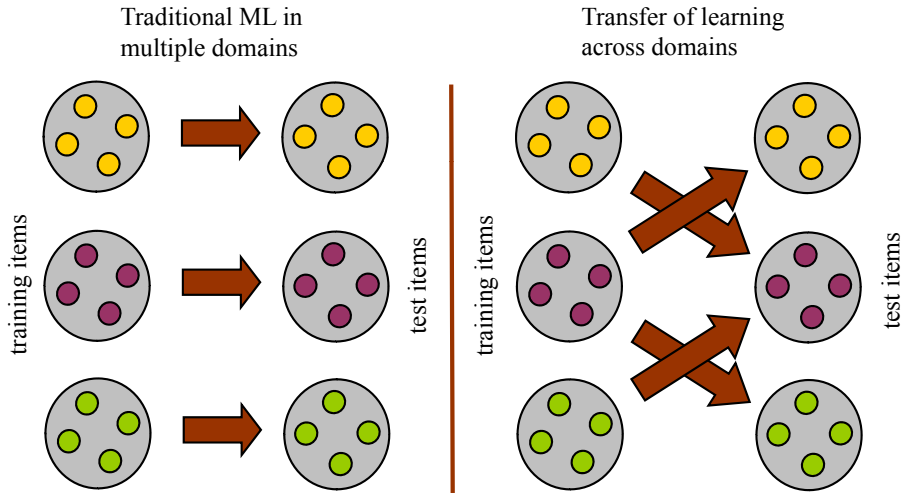
Method	DLN	LDA	SVM
Accuracy	**0.5937**	0.4984	0.4945

Predicting crime

DLN	LDA	SVM
0.2442	0.1035	0.2261

(Than Quang Khoat, Ho Tu Bao, 2010)

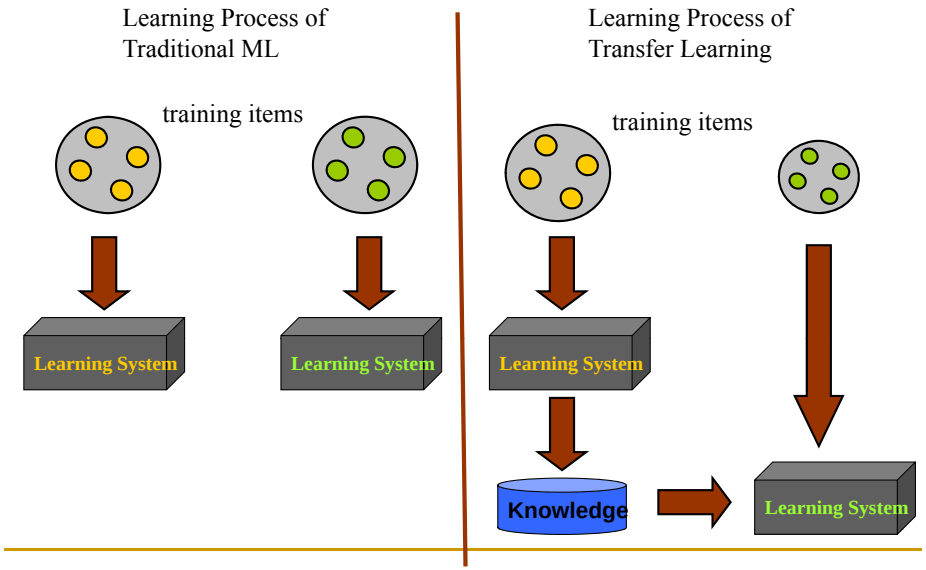
Traditional ML vs. TL *(P. Langley 06)*



Humans can learn in many domains.

Humans can also transfer from one domain to other domains.

Traditional ML vs. TL



Notation

Domain:

It consists of two components: A feature space $\mathcal{X}$, a marginal distribution

$$\mathcal{P}(X), \text{ where } X = \{x_1, x_2, ..., x_n\} \in \mathcal{X}$$

In general, if two domains are different, then they may have different feature spaces or different marginal distributions.

Task:

Given a specific domain and label space $\mathcal{Y}$, for each x_i in the domain, to predict its corresponding label y_i, where $y_i \in \mathcal{Y}$

In general, if two tasks are different, then they may have different label spaces or different conditional distributions

$$\mathcal{P}(Y|X), \text{ where } Y = \{y_1, ..., y_n\} \text{ and } y_i \in \mathcal{Y}$$

Notation

For simplicity, we only consider at most two domains and two tasks.

Source domain:

$\mathcal{P}(X_S)$, where $X_S = \{x_{S_1}, x_{S_2}, ..., x_{S_{n_S}}\} \in \mathcal{X}_S$

Task in the source domain:

$\mathcal{P}(Y_S|X_S)$, where $Y_S = \{y_{S_1}, y_{S_2}, ..., y_{S_{n_S}}\}$ and $y_{S_i} \in \mathcal{Y}_S$

Target domain:

$\mathcal{P}(X_T)$, where $X_T = \{x_{T_1}, x_{T_2}, ..., x_{T_{n_T}}\} \in \mathcal{X}_T$

Task in the target domain

$\mathcal{P}(Y_T|X_T)$, where $Y_T = \{y_{T_1}, y_{T_2}, ..., y_{T_{n_T}}\}$ and $y_{T_i} \in \mathcal{Y}_T$