Artificial Intelligence -- 15-381
Supervised Machine Learning Methods

Jaime Carbonell
2-November-2000

OUTLINE:

- Typology of Machine Learning
- Data Bases (review/intro)
- Machine Learning (& Data Mining) task
- K-nearest-neighbors (kNN)
- Decision Tree review (time permitting)

Typology of Machine Learning Methods

- Learning by caching
    - What/when to cache
    - When to use/invalidate/update cache
- Learning from Examples
  (aka "Supervised" learning)
    - Labeled examples for training
    - Learn the mapping from examples to labels
    - E.g.: Naive Bayes, Decision Trees, ...
- "Speedup" Learning
    - Tuning search heuristics from experience
    - Inducing explicit control knowledge
    - Analogical learning (generalized instances)
- Optimization "policy" learning
    - Predicting continuous objective function
    - E.g. Regression, Reinforcement, ...
- New Pattern Discovery
  (aka "Unsupervised" Learning)
    - Finding meaningful correlations in data
    - E.g. association rules, clustering, ...

Data Bases in a Nutshell

A DB is set of tuples

$$
\begin{matrix}
t_{1,1}, & t_{1,2}, & \ldots, & t_{1,n} \\
t_{2,1}, & t_{2,2}, & \ldots, & t_{2,n} \\
\vdots & & & \vdots \\
t_{m,1}, & t_{m,2}, & \ldots, & t_{m,n}
\end{matrix}
$$

Where typically:

- m >> n, e.g. m = O(10^6), n = O(10^2)
- matrix $\mathbf{T}_{i,j}$ (a DB "table") is dense
- **T** does not usually fit in primary memory
- Relational algebra query system (e.g. SQL)
- Each $t_{i,j}$ is any scalar data type
  (real, integer, boolean, string,...)

Small tables => machine learning
Large tables => data mining

Data Bases in a Nutshell [2}

And we define:

- $\mathbf{t}_i$ = row: $<t_{i,1}, \ldots t_{i,n}>$
  (these are "tuples" in DB parlance)
- $\mathbf{t}_j$ = col: $<t_{1,j}, \ldots t_{m,j}>^T$
  (these are "value sets" in DB parlance)
- All t in $\mathbf{t}_j$ are the ***same*** data type
- All t in $\mathbf{t}_i$ can be different data types

Text Mining

- Rows in $T_{m,n}$ are document vectors
- n = vocabulary size = $O(10^5)$
- m = documents = $O(10^5)$
- $T_{m,n}$ is sparse
- Same data type for every cell $t_{i,j}$ in $T_{m,n}$

Supervised Machine Learning

Given:

- A data base table $\mathbf{T_{m,n}}$
- Predictor attributes: $\mathbf{t}^j$
- Predicted attributes: $\mathbf{t}^k$ (k # j)

Find Predictor Functions:

F_k: $\mathbf{t}^j$ --> $\mathbf{t}^k$, such that, for each k:

$$\mathbf{F_k} = \mathit{Arg}\mathbf{min}_{F_{l,k}} \mathbf{Error}[\mathbf{F_{l,k}}(\mathbf{t}^j), \mathbf{t}^k]_{L_2}$$

(or L_1, or L-infinity norm, ...)

Where typically:

- There is only one $\mathbf{t}^k$ of interest
  and therefore only one $F_k(\mathbf{t}^j)$
- $\mathbf{t}^k$ may be boolean
  => F_k is a binary classifier
- $\mathbf{t}^k$ may be nominal (finite set)
  => F_k is an n-ary classifier
- $\mathbf{t}^k$ may be a real number
  => $\mathbf{t}^k$ is a an approximating function
- $\mathbf{t}^k$ may be an arbitrary string
  => $\mathbf{t}^k$ is hard to formalize

Machine Learning and Data Mining APPLICATIONS

FINANCE:

- Credit-card & Loan Fraud Detection
- Time Series Investment Portfolio
- Credit Decisions & Collections

HEALTHCARE:

- Decision Support: optimal treatment choice
- Survivability Predictions
- medical facility utilization predictions

MANUFACTURING:

- Numerical Controller Optimizations
- Factory Scheduling optimization

MARKETING & SALES:

- Demographic Segmentation
- Marketing Strategy Effectiveness
- New Product Market Prediction
- Market-basket analysis

Simple Data Mining Example

Acct. numb.	Tot Income in K/yr	Job Now?	Num Delinq accts	Max Delinq cycles	Owns home?	Num Credit years	Final disp.
1001	25	Y	1	1	N	2	Y
1002	60	Y	3	2	Y	5	N
1003	?	N	0	0	N	2	N
1004	52	Y	1	2	N	9	Y
1005	75	Y	1	6	Y	3	Y
1006	29	Y	2	1	Y	1	N
1007	48	Y	6	4	Y	8	N
1008	80	Y	0	0	Y	0	Y
1009	31	Y	1	1	N	1	Y
1011	45	Y	?	0	?	7	Y
1012	59	?	2	4	N	2	N
1013	10	N	1	1	N	3	N
1014	51	Y	1	3	Y	1	Y
1015	65	N	1	2	N	8	Y
1016	20	N	0	0	N	0	N
1017	55	Y	2	3	N	2	N
1018	40	N	0	0	Y	1	Y
1019	80	Y	1	1	Y	0	Y
1021	18	Y	0	0	N	4	Y
1022	53	Y	3	2	Y	5	N
1023	0	N	1	1	Y	3	N
1024	90	N	1	3	Y	1	Y
1025	51	Y	1	2	N	7	Y
1026	20	N	4	1	N	1	N
1027	32	Y	2	2	N	2	N
1028	40	Y	1	1	Y	1	Y
1029	31	Y	0	0	N	1	Y
1031	45	Y	2	1	Y	4	Y
1032	90	?	3	4	?	?	N
1033	30	N	2	1	Y	2	N
1034	88	Y	1	2	Y	5	Y
1035	65	Y	1	4	N	5	Y
1036	12	N	1	1	N	1	N
1037	28	Y	3	3	Y	2	N
1038	66	?	0	0	?	?	Y
1039	50	Y	2	1	Y	1	Y
1041	?	Y	0	0	Y	8	Y
1042	51	N	3	4	Y	2	N
1043	20	N	0	0	N	2	N
1044	80	Y	1	3	Y	7	Y
1045	51	Y	1	2	N	4	Y
1046	22	?	?	?	N	0	N
1047	39	Y	3	2	?	4	N
1048	70	Y	0	0	?	1	Y
1049	40	Y	1	1	Y	1	Y

K Nearest Neighbors [kNN]

Ingredients

- Predictor attributes:
  (boolean, nominal, numeric)
- Target attribute:
  (boolean, nominal)
- Similarity function $Sim(R_i,R_j)$
  on predictor attributes in rows i & j.
- Parameter k (number of neighbors)
- New instance to classify
  (predictors known, target unknown)
- Voting function, e.g.
  $S = \Sigma Sim(R_i^+,R_{new}) - \beta\Sigma Sim(R_i^-,R_{new})$, where:
  R_i^+ belong to class +, and in k-neighborhood
  R_i^- belong to class -, and in k-neighborhood
- Decision function:
  If S > threshold, R_{new} belongs to class +
  Otherwise R_{new} belongs to class -

Algorithm

1. Let $K = \mathit{Arg}\max^k[Sim(R_j,R_{new})]$
2. Divide K into K^+ and K^-
3. Calculate S for R_{new} [see above] usually **b** = 1
4. Vote for class membership
   usually S-threshold = 0

Trend Detection in DM

Example: Sales Prediction

2000 Q1 sales = 4.0M,
2000 Q2 sales = 3.5M
2000 Q3 sales = 3.0M
2000 Q4 sales = ??

Now if we knew *last year*:
1999 Q1 sales = 3.5M,
1999 Q2 sales = 3.1M
1999 Q3 sales = 2,8M
1999 Q4 sales = 4.5

And if we knew *previous year*:
1998 Q1 sales = 3.2M,
1998 Q2 sales = 2.9M
1998 Q3 sales = 2.5M
1998 Q4 sales = 3.7M

What will 2000 Q4 sales be?
What if Christmas 2000 was cancelled?

What will 2001 Q4 sales be?

Trend Detection in DM [2]

Methods

- Numerical series extrapolation
- Cyclical curve fitting
    - Find period of cycle
    - Fit curve for each period
      (often with L_2 or $L_{infinity}$ norm)
    - Find translation (series extrapolation)
    - Extrapolate to estimate desire values
- Preclassify data first
  (e.g. "recession" and "expansion" years)
- Combine with "standard" data mining

Thorny Problems

- How to use external knowledge* to make up for limitations in the data?
- How to make longer-range extrapolations?
- How to cope with corrupted data?
    - Random point errors (easy)
    - Systematic error (hard)
    - Malicious errors (impossible)

Methods for Supervised DM

Classifiers

- Linear Separators (regression)
- Naive Bayes (NB)
- Decision Trees (DTs)
- k-Nearest Neighbor (kNN)
- Decision rule induction
- Support Vector Machines (SVMs)
- Neural Networks (NNs) ...

Points of Comparison

- Hard vs Soft decisions
  (e.g. DTs and rules vs kNN, NB)
- Human-interpretable decision rules
  (best: rules, worst: NNs, SVMs)
- Training data needed (less is better)
  (best: kNNs, worst: NNs)
- Graceful data-error tolerance
  (best: NNs, kNNs, worst: rules)

Symbolic Rule Induction

General idea

- Labeled instances are DB tuples
- Rules are generalized tuples
- Generalization occurs at term in tuple
- Generalize on new E^+ not predicted
- Specialize on new E^- not predicted
- Ignore predicted E^+ or E^-

Example term generalizations

- Constant => disjunction
  e.g. if small portion value set seen
- Constant => least-common-generalizer class
  e.g. if large portion of value set seen
- Number (or ordinal) => range
  e.g. if dense sequential sampling

Example term specializations

- class => disjunction of subclasses
- Range => disjunction of sub-ranges

Symbolic Rule Induction Example

Age	Gender	Temp	b-cult	c-cult	loc	Skin	disease
65	M	101	+	.23	USA	normal	strep
25	M	102	+	.00	CAN	normal	strep
65	M	102	-	.78	BRA	rash	dengue
36	F	99	-	.19	USA	normal	*none*
11	F	103	+	.23	USA	flush	strep
88	F	98	+	.21	CAN	normal	*none*
39	F	100	+	.10	BRA	normal	strep
12	M	101	+	.00	BRA	normal	strep
15	F	101	+	.66	BRA	flush	dengue
20	F	98	+	.00	USA	rash	*none*
81	M	98	-	.99	BRA	rash	ec-12
87	F	100	-	.89	USA	rash	ec-12
12	F	102	+	??	CAN	normal	strep
14	F	101	+	.33	USA	normal	
67	M	102	+	.77	BRA	rash	

Candidate Rules:

```
IF age = [12,65]
   gender = *any*
   temp = [100,103]
   b-cult = +
   c-cult = [.00,.23]
   loc = *any*
   skin = (normal,flush)
THEN: strep

IF age = (15,65)
   gender = *any*
   temp = [101,102]
   b-cult = *any*
   c-cult = [.66,.78]
   loc = BRA
   skin = rash
THEN: dengue
```

Disclaimer: *not* real medical data or rules.