Data mining, machine learning, and uncertainty reasoning

林偉川

Quest for natural laws

- Having powerful algorithms for concept formation is not enough
- Build a system which is capable not only of constructing new concepts, but also of describing their relations in terms of laws is more important

2

Quest for natural laws

- Several reasons support activities in this field:
  1. Huge DBs from many scientific fields are available, waiting for someone to analyze them
  2. Powerful techniques in machine learning and artificial intelligence have been developed so that one can hope for a kind of 'intelligent' analysis
  3. If intelligent automatic analyzers are not constructed, the search into artificial discovery may help to ellucidate some of the mysteries of human invention (inspiration, analogy, and abstraction)

3

一、 理想氣體物態方程式

$$\boxed{PV = nRT = NkT}$$

1.符號說明：

P：壓力。單位：atm，Nt/m^3　　V：體積。單位：ℓ（升），m^3

n：莫耳數。　　N：分子數。

R：理想氣體常數。　　k：波茲曼常數。

T：絕對溫度（K），且 $T=273+t$（攝氏溫度，℃）

2.常數：

在 S.T.P.（273K，1atm）下，1mole 理想氣體占體積 22.4 公升，

故 $R = \frac{PV}{nT} = \frac{1atm \cdot 22.4\ell}{1mole \cdot 273K} = \underline{0.082 \frac{atm-\ell}{mole-K}}$

$$= \frac{1.013\times10^5 \frac{Nt}{m^2} \cdot 22.4\times10^{-3} m^3}{1mole \cdot 273K} = \underline{\boxed{8.317} \frac{Joule}{mole-K}}$$

且 $\frac{R}{k} = \frac{n(\text{莫耳數})}{N(\text{分子數})} = N_0$(亞佛加厥常數)

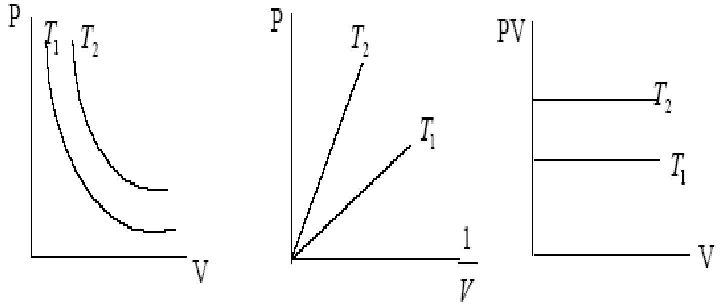
理想氣體常數＝波茲曼常數×亞佛加厥常數

$$\boxed{R = k \cdot N_0}$$

故 $k = \frac{R}{N_0} = \frac{8.317 \frac{J}{mole-K}}{6.02\times10^{23} \frac{\text{分子}}{mole}} = \underline{\boxed{1.38\times10^{-23}} \frac{Joule}{\text{分子}-k}}$

二、波以耳定律：理想氣體在定溫下，定量氣體的壓力（P）與其體積（V）成反比。

在 n、T 固定下　$P \cdot V =$ 定值　或　$P_1V_1 = P_2V_2$



5

在 n、T 不變下　P-V 圖：雙曲線；P-$\frac{1}{V}$ 圖：斜直線；PV-V 圖：水平直線。

若 n 不變，T 可變化下，以上三圖溫度關係皆為 $T_2 > T_1$。

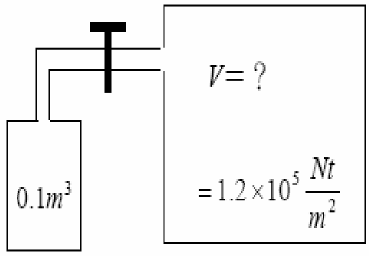
例：用一筒氦氣吹氣球，氦氣筒之容積為 0.1 立方米，原來之壓力為 $1.0\times10^7\ Nt/m^2$。每一汽球充氣後體積為 1.0×10^{-2} 立方米，壓力為 1.2×10^5 牛頓/米^2。用該氦氣筒最多約可吹出多少個這樣的氣球？

（A）940 個（B）820 個（C）600 個（D）480 個（E）260 個。

答：（B）

解：(1) 先利用波以耳定律，計算器瓶內壓力當降至與氣球相等壓力時的體積。

(2) 氣瓶內之氣體不可能全部用以充氣。



先將氣瓶內氣體放入右室，壓力大小爲 $1.2\times10^5\ Nt/m^2$

故右邊氣室之體積

$$V=\frac{1.0\times10^7\times0.1}{1.2\times10^5}-0.1(\text{氣瓶體積})$$

$$=8.23(m^3)$$

再將右室氣體用以吹氣球，氣球數 $=\frac{8.23}{1.0\times10^{-2}}\fallingdotseq 823$ 個

Ideal Gas Model

- Molecular Model for an Ideal Gas
  - http://www.phy.ntnu.edu.tw/java/idealGas/idealGas.html
  - http://hyperphysics.phy-astr.gsu.edu/hbase/kinetic/idegasc.html#c1

8

Quantitative Empirical Laws

- Quantitative empirical laws → rediscover the ideal gas. $PV=8.32NT$ where P is pressure, V is volume, N is gas amount, and T is temperature
- BACON start by suggesting a series of experiments that will provide the measurement data

Quantitative Empirical Laws

- The human operator carries them out and supplies the computer with the outcomes
- As enough data have been gathered, the system searches the space of mathematical functions with the objective of finding an equation consistent with the data
- One method of searching for the equation is to make one of the variables dependent while the others remain independent

Ideal Gas Model

- Try to find out the relations between
  - total number of molecules N ---- volume V
  - the pressure of the system P--- volume V
  - the velocity of the molecules v --- volume V
- An ideal gas can be characterized by three state variables: absolute pressure (P), volume (V), and absolute temperature (T). The ideal gas law : PV=nRT=NkT

n = number of moles

R = universal gas constant = 8.3145 J/mol K

N = number of molecules

k = Boltzmann constant = 1.38066×10^{-23} J/K

$k = R/N_A$ where

N_A = Avogadro's number = 6.0221×10^{23}

11

Ideal Gas Model

- The pressure that a gas exerts on the walls of its container is a consequence of the collisions of the gas molecules with the walls. In this model:
  - The molecules obey Newton's law of motion.
  - The molecules move in all direction with equal probability.
  - There is no interactions between molecules (no collisions between molecules).
  - The molecules undergo elastic collisions with the walls.

12

Quantitative Empirical Laws

- Let the system have a repertoire of typical law-forms such as

$$y = ax^2 + bx + c$$

$$\sin(y) = ax + b$$

$$y^{-1} = ax + b$$

- The principle consists in selecting the best law-form and tuning the parameters a,b, …, with the objective of finding an equation that best describes the observed data

13

BACON system policy

- Suppose the equation $y^{-1} = ax + b$ has been selected. At the beginning, the parameters a and b are initialized to the values 1, 0, and -1, so that the following combinations are considered as a set of initial states: [a=1, b=1], [a=1, b=0], [a=1, b=-1], [a=0, b=1], [a=0, b=0], etc.

➔ $y^{-1} = x + 1$ …

14

Sample data for the BACON system

quantity	temperature	pressure	volume
N=1	T=10	P=1000	V=2.36
.	.	P=2000	V=1.18
.	.	P=3000	V=0.78
.	T=20	P=1000	V=2.44
.	.	P=2000	V=1.22
.	.	P=3000	V=0.81
.	T=30	P=1000	V=. . .
.	.	P=2000	V=. . .
.	.	P=3000	V=. . .
. N=2 .	⋮		
. N=3 .	⋮		

BACON system policy

- Suppose the values in the above table have been measured. BACON will investigate them in the following steps:
  1. Find a function describing V=f(P) for the triplets of examples assigned in the table to each of the three temperature T=10, T=20, and T=30
  Suppose that $V^{-1} = \boxed{a}P + b$ with the following parameters provides the best fit:

T=10; a=0.000425; thus $V^{-1} = 0.000425P$
T=20; a=0.000410; thus $V^{-1} = 0.000410P$
T=30; a=0.000396; thus $V^{-1} = 0.000396P$

16

BACON system policy

2. Since the parameter values depend on the temperature T, the next task is to find the function relating a to T. Again, the best fit is achieved by the form $a^{-1} = cT + d$ with the values of parameters, c and d, depending on N:

N=1; c=8.32 and d=2271.4; thus $a^{-1} = 8.32T + 2271.4$

N=2; c=16.64 and d=4542.7; thus $a^{-1} = 16.64T + 4542.7$

N=3; c=24.96 and d=6814.1; thus $a^{-1} = 24.96T + 6814.1$

17

BACON system policy

3. Find function relating c to N and d to N. The best fit is achieved by c=eN and d=fN, with e=8.32 and f=2271.4. These parameters do not depend on any other variable $a = (8.32NT + 2271.4N)^{-1}$
4. Substituting the equation into those equations found in the previous steps, the system obtains: $V^{-1} = (8.32NT + 2271.4N)^{-1} P$ and this last expression can easily be transformed into:

$PV = 8.32NT + 2271.4N \rightarrow PV = 8.32N(T + 273)$

which is the standard form of the ideal gas law

18

Conclusion of BACON system

- The essence of BACON is to apply common search principles in the quest for an ideal form of quantitative law, rather than just find the best fitting parameters ➔ traditional regression technique

19

How to cope with vastness of the search space

- Machine learning is that the space of all possible descriptions is often so large that the search has to rely on heuristics, or becomes computationally intractable
- The danger of converging to local maxima of evaluation functions is in large spaces more serious
- 2 techniques to attack this problem: the use of analogy and the idea of storing the original examples instead of their generalized descriptions

20

Analogy providing search heuristics

- To find proper analogies is one of the secrets of intelligence. Much work has been devoted to analogy-based reasoning
- The general framework of analogy S stands for source, SC for source concept, T for target, TC for target concept
- The task is to derive TC from T in a way that is analogous to the way SC was derived from the S
- Thus, having target, the learner must find a proper source

General scheme of analogy

$$\begin{tikzcd} T \arrow[r] & TC \\ S \arrow[u] \arrow[r] & SC \arrow[u] \end{tikzcd}$$

Reasoning-by-Analogy Algorithm

1. Recognition. Given a target concept, find in the background theory a source S that is 'similar' to T. The similarity can be measured by syntactic distance, by the existence of common generalization of a pair of unifying substitutions, or by some hint supplied by the user
2. Elaboration. Find SC, together with the inference chain $\longmapsto_S$ leading to it from S. Note that, for each S, a collection of SC's usually exist
3. Evaluation. Among the SC's, find the one that best satisfies given criteria

23

Reasoning-by-Analogy Algorithm

4. Apply to T an inference chain $\longmapsto_T$ 'similar' to $\longmapsto_S$, thus obtaining TC. Access the utility of TC
5. If necessary, repeat iteratively steps 1 – 4 to find S, Sc, $\longmapsto_S$, and $\longmapsto_T$ that yield the most promising (useful) TC
6. Consolidation. Include TC together with the inference chain $\longmapsto_T$ into the background theory

24

Conclusion of using analogy

- The framework is too general, reasonable constraints are usually needed
- The source S can be explicitly supplied by the user telling the system what to do is analogous to other existed application
- The other possibility is that the user takes over the evaluation process and selects a proper SC for the source that has been suggested by the system

25

Instance-based learning

- The only reason for learning is the need to identify further examples, the learner can adopt an alternative policy ➔ instead of description, store typical examples
- This can preclude many troubles potentially entailed by the search through a prohibitively large space of generalization

26

Instance-based learning

- IBL system can store selected examples (described by attributes values) and use them according to “nearest-neighbor” principle ➔ the newly arrived example is assigned the class of the closest one among the stored examples
- A simple formula to calculate the similarity between the examples x and y is used (x_i and y_i are the respective values of the i-th attribute) :

Calculate the similarity

- where f is calculated by the following formula for numeric / symbolic and Boolean attributes

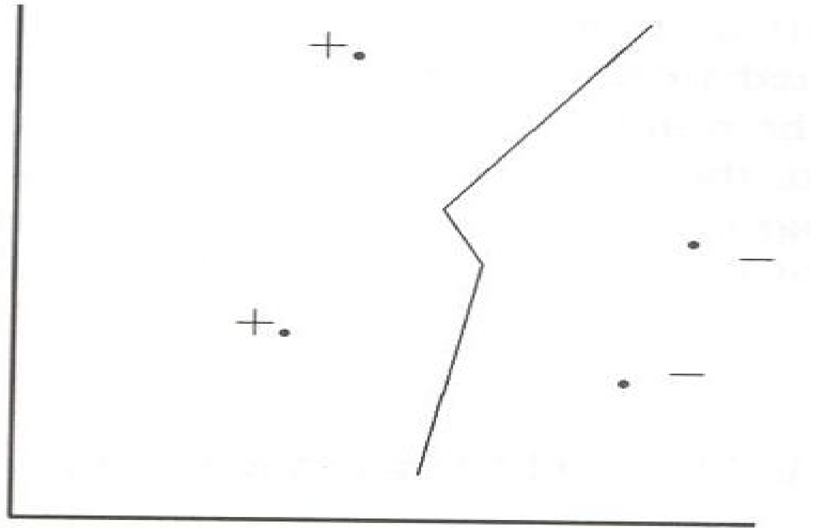
$$similarity(x, y) = -\sqrt{\sum_{i=1}^{n} f(x_i, y_i)}$$

$$f(x_i, y_i) = (x_i - y_i)^2$$

$$f(x_i, y_i) = \begin{cases} 1, & for x_i \neq y_i \\ 0, & for x_i \approx y_i \end{cases}$$

- 4 examples described by 2 numeric variables are depicted, together with the discrimination function separating the space of positive examples from the space of negative examples

Positive/negative example defining the space



29

IBL policy and algorithm

- The learner assumes the availability of a feedback that will immediately inform the learner about the success or failure of each single classification attempt

30

IBL policy and algorithm

- IBL algorithm involves the following steps：
  1. Define the set of representatives containing the first example
  2. Read a new example x
  3. ∀y in the set of representatives, determine similarity(x,y)
  4. Label x with the class of the closest example in the set of representatives
  5. Find out from the feedback whether the classification was correct
  6. Include x in the set of representatives, go to 2.

Shortcomings of IBL

- 2 shortcomings degrade the utility of this elementary version: excessive storage requirements caused by the fact that all examples are stored; and sensitivity to noise

Rectification of IBL

- The rectification consists of a selective storage of examples by a "wait-and-see" strategy that is summarized by the following principles:
  1. Whenever a new instance has been classified, the 'significance-score' of each of the previous instances is updated and the instance is stored
  2. Instances with good scores are used for classification; bad scores are deleted (noise)
  3. Mediocre instances are retained as potential candidates and not used for classification

33

IBL's classification

- In the classification phase, the new arrival is assigned the class of the nearest good instance if a good instance existed. Otherwise, the new arrival is assigned the class of the nearest mediocre instance
- The system increments the score of those mediocre that are closer to the new arrival than the closest good instance. If no good instance is available, the system updates mediocres inside a randomly chosen hypersphere around the new arrival ➔ locality characteristics in the OS

34

IBL's classification

- A score is considered as good whenever the classification accuracy achieved by this instance is higher than the frequency of the example's class
- The classification accuracy of class ⊕ is the percentage of correctly recognized positive example in the set of all examples

35

Conclusion of IBL system

- Instance-based learning has been reported to achieve a significant recognition power in attribute-value domains, especially when the number of examples is large and the attributes describing them are properly chosen
- The robustness against noise is satisfactory
- The power of the system degrades if the descriptions of the examples contain irrelevant attributes and/or if the number of examples available to the learning procedure is small

36

Homework

- Find another example of Instance based learning and described them as your report!!

37