MDL原理

オッカムの剃刀（Occam's razor）

- データマイニング・機械学習の仕事は、データを表現するモデル を探すことだと言える
  - 例: ガウス混合モデル, (等方正規分布の)混合（k-means法）.
  - Model vs Hypothesis
- では、正しいモデルとは何か? どうやって選ぶか？
- オッカムの剃刀: それ以外の条件が全て同じなら, 最も単純なモデルが最良である.
  - 人生訓としてもよかろう

Occam の剃刀

- 人口に膾炙しているのは
  - Entities should not be multiplied beyond necessity.
- Bertrand Russell によれば
  - It is vain to do with more what can be done with fewer.
- 最も普通の解釈
  - Among the theories that are consistent with the observed phenomena, one should select the simplest theory.

オッカムの剃刀とMDL

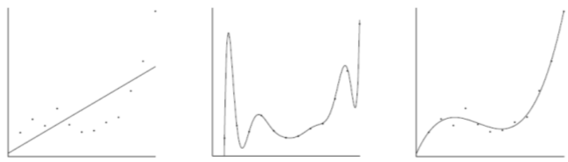
- 単純 なモデルとは何か?
- 最小記述長原理（Minimum Description Length Principle）: 全てのモデルは、データを（ロス無しで lossless） 符号化（コード化）すると考える. データの 最短符号コード （データを最短圧縮）を与えるモデルが最良.
  - 関連概念: Kolmogorov 複雑度=当該データを出力する最短のプログラムの長さ (を求める計算は計算不能).
  - 符号（コード）のコスト: AさんからBさんに当該符号（コード）を送信するコスト. 長さに比例.

Minimum Description Length (MDL)

- 記述長は二つの長さからなる
  - モデルの記述長
  - モデルが与えられたとして当該データを記述する長さ.
  - L(D) = L(M) + L(D|M)
- この2つの長さの間にはトレードオフがある
  - 非常に複雑なモデルを用いれば、当該データを短く記述することができるが、モデル自体の記述が大変（記述長が長くなる）
  - 非常に簡単なモデルを用いれば、モデルの記述は簡単だが、データの記述が大変に（記述長が長く）なる

例

- 回帰: データを記述する多項式を見つけよ
  - モデルの複雑さ vs. 当てはまりのよさ



Source: Grunwald et al. (2005) *Advances in Minimum Description Length: Theory and Applications.*

6

MDL と機械学習

- より短い符号（コード）がよいのはどうしてか?
  - より短い符号（コード）は、データ中の規則性をよく表している。
  - 規則性（の簡単なものは） パターンである
  - 規則性（パターン）は面白い、役立つ
- 例

でも、ランダムって何だ？

0000100001000010000100001000010000100001000100001000010000100001

- 短い記述が可能. 例えば, repeat 12 times 00001

0100111001010011011010100001110101111011011010101110010011100

- ランダム列, パターンなし, 圧縮できず

MDL とクラスタリング

- もしデータがクラスタリングできるなら、データの代わりに、クラスタを伝送すればよい
  - クラスタの記述をまず伝送する必要あり
  - そして、クラスタ毎にデータの記述を送る.
- もしクラスタリングが良ければ、伝送コストは低くなる
  - なぜか?
  - もしクラスタ内の全要素が実は同じものであったらどうか?
  - もしクラスタ内にごく少数の要素しかなかったらどうか?

クラスタ内が一様なクラスタはコード長短く記述できる。
しかし、たくさんあったら、その効果が無になる

MDL に係る課題

- 正しい、モデル族は何か?
  - これは（機械学習共通）、我々が得る解を決める
    - 例: 多項式、決定木、
    - クラスタリング
- 符号（コード）長とは何か?
  - それが、最適化する対象
  - 情報理論（符号理論）と関係あり

最小記述長(minimum description length)

- Occam’s razor: “最短仮説を選べ”

$$h_{MDL} = \arg\min_{h \in H} L_{C_1}(h) + L_{C_2}(D \mid h)$$

ex. 木を記述するビット数

h が所与のとき、D を記述するビット数

∝ 記述する符号の長さ

∝ 誤分類データの個数

このままでは、使えない。使うようにした方法がある
1. Rissanen による統計的MDL
2. Kolmogorov/Chaitin のプログラム複雑度に基づくMDL
であり、Lin & Vitanyi グループによるもの

ごく短いイントロ

情報理論

符号化（コード化）とは

- 次の列を考えてみよう

AAABBBAAACCCABACAABBAACCABAC

- 2文字を用いた文字列に符号化することを考えてみよう

50% A
25% B
25% C

A は他より大きい 50% を占めるので
他より短い表現とすべき

A → 0
B → 10
C → 11

これが最良であることは証明できる

符号化（コード化）

- Prefix Codes: どのコードも他のコードのprefix（語頭、接頭）ではない

  A → 0　　　一意に、即座にデコード可能
  B → 10
  C → 11

- 符号と分布: 符号から分布へ、分布から符号へ（多対1）のsh像がある
  - P は要素の集合 (例, {A,B,C}) から分布への写像であるとしよう. そうするとある（接頭、語頭）符号 C が存在して次の式を満たす
    $L_C(x) = -\lceil \log P(x) \rceil, x \in \{A, B, C\}$
  - 要素{A,B,C} からなる任意の（接頭、語頭）符号 C に対し次のような分布を定めることができる $P(x) = 2^{-C(x)}$

- このようにして定義された符号は、最短の平均符号長を有することが知られている

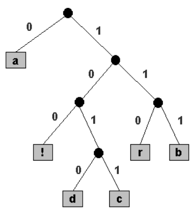
符号と確率

- 有限または可付番無限集合 X を考える
  - X の符号 $C(x)$ とは
    - X から $\cup_{n>0}\{0,1\}^n$ への1-to-1 写像
    - $L_C(x)$: 符号Cを用いた時の符合長（ビット）
  - P: X 上で定義した確率分布
    - $P(x)$: x の確率
    - 観測値の系列（通常は iid） $x_1, x_2, \ldots, x_n$: x^n

$$P(x^n) = \prod_{i=1}^{n} P(x_i)$$

接頭符号（語頭符号）

- 接頭符号: 瞬時復号可能な符号の例
  - どの符号も他の符号の語頭にはなっていない



a	0
b	111
c	1011
d	1010
r	110
!	100

http://www.cs.princeton.edu/courses/archive/spring04/cos126/

最適符号

- ある符号 C の符号長の期待値

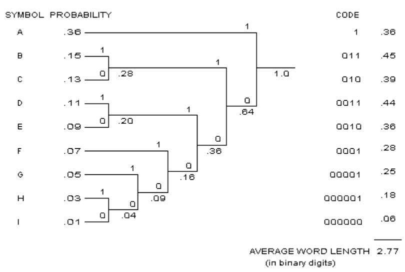
$$E_P(L_C(x)) = \sum_{x \in X} P(x) L_C(x)$$

  - 下界:

$$H(x) = -\sum_{x \in X} P(x) \log_2 P(x)$$

- 最適符号
  - 瞬時復号可能な符号の中で期待符号長が最小
  - 仮に分布 P が与えられた時、どう設計せればよいか?
    - Huffman 符号

ハフマン符号



http://star.itc.it/caprile/teaching/algebra-superiore-2001/

有限集合の符号

- {1, 2, …, M} の符号語を設計するには?
  - 一様分布を仮定すれば: それぞれの数に $1/M$
  - ~log$M$ ビット

無限集合の符号

- 正整数すべての符号を設計するには?
  - それぞれの k について
    - まず先頭に $\lceil \log k \rceil$ 個の0をおき
    - 次に一個の 1 をおき
    - そして k を符号化する。ただし $\{1, \ldots, 2^{\lceil \log k \rceil}\}$ の符号
    - 長さは合計 ~ 2log$k$ + 1 ビット
  - 勿論、改善は可能…

双対性(かな?)

- P を X 上の確率分布としよう。そうすると X に対する符号 C で次の条件を満たすものがある:

$$L_C(x) = \lceil -\log P(x) \rceil$$

- C を X 上の即時復号可能な符号とする. そうすると確率分布 P で次の条件を満たすものがある:

$$L_C(x) = -\log P(x)$$

$$L_C(x^n) = -\log P(x^n)$$

最小記述長 符号的解釈

- MDL: 次を最小化する仮説を選ぶ

$$h_{MAP} = \arg\max_{h \in H} P(D \mid h) P(h)$$
$$= \arg\min_{h \in H} -\log_2 P(D \mid h) - \log_2 P(h)$$
$$= \arg\min_{h \in H} L_{C2}(D \mid h) + L_{C_1}(h)$$

エントロピー

- 確率変数 X を考える. それは n この異なる値をとるとする

$$X = \{x_1, x_2, \ldots, x_n\}$$

その分布を $\mathrm{P(X)} = \{p_1, \ldots, p_n\}$ とする

- この分布から得られる符号 C がある. それは長さが $L_C(x_i) = -\lceil \log p_i \rceil$ であり平均符号長は次のようになる

$$-\sum_{i=1}^{n} p_i \lceil \log p_i \rceil$$

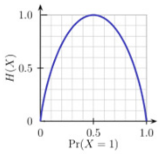
- これは、ほぼ、確率変数 X のエントロピー $H(X)$ である

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

- Shannon の定理: 分布 P(X) のエントロピーは、この分布に対応する符号の平均符号長の下界である
  - 分布 P(X) からサンプルした N 個の数字を符号化するとき, 実現可能な最短符号長は $N * H(X)$
  - 注意: ロス無し Lossless 符号化

エントロピー

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$



- その意味は?
- エントロピーは、分布の異なる側面を記述している:
  - 確率変数 X で表されるデータの圧縮可能性
    - Shannon の定理から得られる
  - その分布の 不確かさ (一様分布の時、エントロピーは最大値)
    - 確率変数の取る値をどのくらい正しく予想できるか、という量
  - その確率変数の持っている情報
    - その値を表現するのに用いたビット数(最小ビット数)がこの値の情報内容である.

情報理論で用いる物差し

- 条件付エントロピー H(Y|X): 確率変数 X を知った後でも残る確率変数 Y に関する不確かさ

$$H(Y|X) = -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)}$$

- 相互情報量 I(X,Y): Y (or X) を知ることによって減少する X (or Y) の不確かさ

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

情報理論で用いる物差し

- クロスエントロピー: 分布 Q の符号を用いて分布 P を符号化した時の平均符号長

$$-\sum_x P(x)\log Q(x)$$

- KL Divergence KL(P||Q):分布 Q の符号を用いて分布 P を符号化した時に、分布 P を用いて分布 P を符号化した時より長くなる、符号長の増加分

$$KL(P||Q) = -\sum_x P(x)\log Q(x) + \sum_x P(x)\log P(x)$$

  - 非対称. 従って、距離ではない
  - 扱いにくさ: Pが非零にも関わらず Q が零になる場合.

情報理論で用いる物差し

- Jensen-Shannon Divergence JS(P,Q): 二つの分布 P と Q の間の距離
  - KL-divergence の欠点に対応
- M = ½ (P+Q) を分布の平均とすれば

$$JS(P,Q) = \frac{1}{2}KL(P||M) + \frac{1}{2}KL(Q||M)$$

- Jensen-Shannon は距離である