情報意味論 (第11回)
ベイズ学習

慶應義塾大学理工学部
櫻井 彰人

目次

- Bayes **定理**
- MAP と ML
- Bayes **最適分類器**, Gibbs **アルゴリズム**
- MDL
- **誤差関数**
- Naïve Bayes

ベイズ学習が対象とする課題

- **先験的知識やバイアスを表現する最良の方法は？**
- **決定木の枝狩りの正当化はどうやって？ 枝狩りはどのようにすればよいのか？**
- **神経回路網では、どうして2乗誤差の最小化を測るのか？もっと別の関数を使わなくてもよいのか？使うとしたら、いつか？**
- (Bayes-) **最適な分類器とは**
- Naïve Bayes: **属性数が多いときの課題克服**

Bayes の定理

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

例 (Mitchell Chap. 6.2)

ある患者が臨床検査を受けたところ結果が陽性であった. 当該検査は, 本当に病変があるときに陽性となる確率は 98% を誇る. また, 病変がないときに正しく陰性となる確率は 97% である. さらに, 全人口に対するこのガンをもつ率は .008 である.

P(cancer) = .008 P(¬cancer) = .992
P(+ | cancer) = .98 P(− | cancer) = .02
P(+ | ¬cancer) = .03 P(− | ¬cancer) = .97

P(+) = P(+ | c'r) P(c'r) + P(+ | ¬c'r) P(¬c'r) =.0376

$$P(\text{cancer} \mid +) = \frac{P(+ \mid \text{cancer})\,P(\text{cancer})}{P(+)} = \mathbf{.209}$$

例 (Mitchell Exercise 6.1)

2回目の検査を受け, その結果も陽性であったとしよう. ガンである事後確率はどうなるであろうか?

P(cancer) = .008 P(¬cancer) = .992
P(+ | cancer) = .98 P(− | cancer) = .02
P(+ | ¬cancer) = .03 P(− | ¬cancer) = .97

$P(+_1 +_2) = P(+_1 +_2 \mid c'r)\,P(c'r) + P(+_1 +_2 \mid \neg c'r)\,P(\neg c'r) = .00858$

$$P(\text{cancer} \mid +_1 +_2) = \frac{P(+_1 +_2 \mid \text{cancer})\,P(\text{cancer})}{P(+_1 +_2)} = \mathbf{.896}$$

有用な公式

乗法の公式 (実は、条件付確率の定義！) :

$P(A \wedge B) = P(A|B)\,P(B) = P(B|A)\,P(A)$

和事象:

$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

全確率の公式:

$P(B) = \sum_{i=1}^{n} P(B|A_i)\,P(A_i)$

仮説選択に関して教えてくれること

$$P(h|D) = \frac{P(D|h)\,P(h)}{P(D)}$$

P(h) = **仮説 h の事前確率**
P(D) = **訓練データ D の生起確率**
P(h|D) = **D が与えられたときの h の生起確率**
P(D|h) = **h が与えられたときの D の生起確率**

仮説選択

$$P(h|D) = \frac{P(D|h)\,P(h)}{P(D)}$$

データが所与のとき、必要とするのは、最もありうべき仮説であろう.
事後確率最大仮説 (Maximum a posteriori hypothesis) h_{MAP}:

$$h_{MAP} = \arg\max_{h \in H} P(h|D)$$
$$= \arg\max_{h \in H} \frac{P(D|h)\,P(h)}{P(D)}$$
$$= \arg\max_{h \in H} P(D|h)\,P(h)$$

仮説選択 (続)

全ての i, j **について** $P(h_i) = P(h_j)$ **と仮定すれば, より簡単化でき, 最尤** *Maximum Likelihood (ML)* **仮説 を選ぶことになる**

$$h_{MAP} = \arg\max_{h \in H} P(D|h)\,P(h)$$
$$h_{ML} = \arg\max_{h \in H} P(D|h)$$

力ずく MAP 仮説学習

1. **各仮説 h について, 事後確率を計算する:**

$$P(h|D) = \frac{P(D|h)\,P(h)}{P(D)}$$

2. **出力する仮説** h_{MAP} **は、その中で事後確率最大のもの, 引き分け時はランダムに選択:**

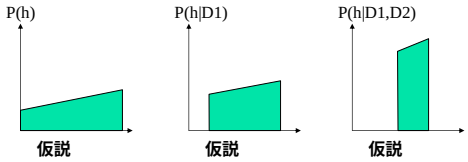
$$h_{MAP} = \arg\max_{h \in H} P(D|h)\,P(h)$$

ID3 に関して

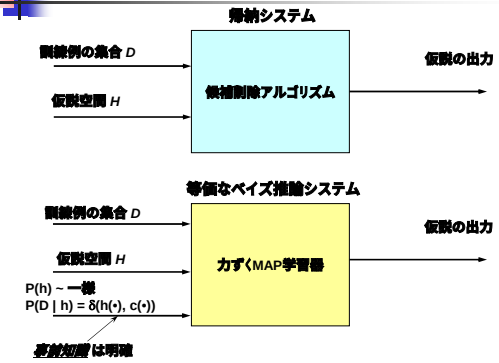
力ずく MAP 仮説学習、ただし

- **ノイズがないという仮定:**
  P(D|h)=1 if h **が** D **に矛盾しない、かつ**
  0 if D **に矛盾するとき**
- **より小さい木を選好するという仮定:**
  P(h) **には小さい木を選ぶバイアス込み**

ノイズがないときの事後確率の進展



一般には



ベイズ学習が対象とする課題

- 先験的知識やバイアスを表現する最良の方法は？
- 決定木の枝狩りの正当化はどうやって？ 枝狩りはどのようにすればよいのか？
- 神経回路網では、どうして2乗誤差の最小化を測るのか？もっと別の関数を使わなくてもよいのか？使うとしたら、いつか？
- (Bayes-) 最適な分類器とは
- Naïve Bayes: 属性数が多いときの課題克服

Bayes 最適な分類器

$$\arg\max_{c_j \in \{+,-\}} \sum_{h_i \in H} P(c_j \mid h_i) P(h_i \mid D)$$

注: Bayes 最適な分類器は H にあるとは限らない

例 (Mitchell Chap. 6.7)

$P(h_1 \mid D) = .4$ $P(- \mid h_1) = 0$ $P(+ \mid h_1) = 1$
$P(h_2 \mid D) = .3$ $P(- \mid h_2) = 1$ $P(+ \mid h_2) = 0$
$P(h_3 \mid D) = .3$ $P(- \mid h_3) = 1$ $P(+ \mid h_3) = 0$

それゆえ:

$$\sum_{h_i \in H} P(+ \mid h_i) P(h_i \mid D) = .4$$

$$\sum_{h_i \in H} P(- \mid h_i) P(h_i \mid D) = .6$$

そして:

$$\arg\max_{c_j \in \{+,-\}} \sum_{h_i \in H} P(c_j \mid h_i) P(h_i \mid D) = -$$

Gibbs 分類器 (Mitchell Chap. 6.8)

1. 仮説を P(h|D) に従ってランダムに選ぶ
2. 新事例をこれに従い分類する

驚くべきことに: もし仮説を事前分布 P(h) に従ってランダムに選ぶと,

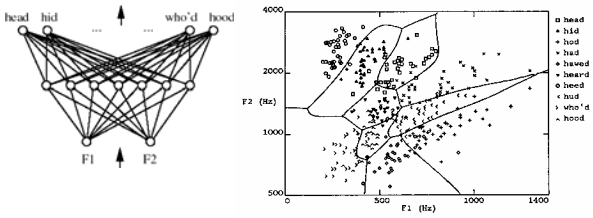
$E[error_{Gibbs}] \le 2E[error_{BayesOptimal}]$

(詳細は “Machine Learning”)

ベイズ学習が対象とする課題

- 先験的知識やバイアスを表現する最良の方法は？
- 決定木の枝狩りの正当化はどうやって？ 枝狩りはどのようにすればよいのか？
- 神経回路網では、どうして2乗誤差の最小化を測るのか？もっと別の関数を使わなくてもよいのか？使うとしたら、いつか？
- (Bayes-) 最適な分類器とは
- Naïve Bayes: 属性数が多いときの課題克服

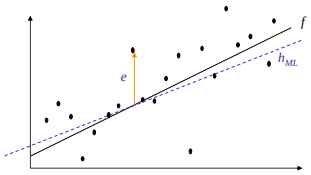
ニューラルネットでは



学習方法2通り

分類結果の出力: 確率　回帰結果の出力: 値

実数値の学習 (回帰)



実数値の学習 (回帰) 続

学習事例: $\langle x_i, d_i \rangle$ 但し

$d_i = f(x_i) + e_i$

e_i はノイズ = iid なる正規分布に従う確率変数で、平均=0 かつ分散は有限とする

iid=independent, identically distributed

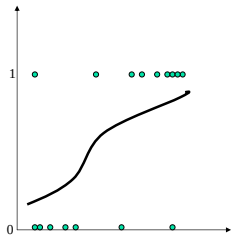
ならば (仮説):

$$h_{ML} = \arg\min_{h \in H} \sum_{i=1}^{m} (d_i - h(x_i))^2$$

実数値の学習 (回帰) 続

$$\begin{aligned} h_{ML} &= \arg\max_{h \in H} \ln p(D \mid h) \\ &= \arg\max_{h \in H} \ln \prod_{i=1}^{m} e^{-\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2} \\ &= \arg\max_{h \in H} \sum_{i=1}^{m} -\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2 \\ &= \arg\max_{h \in H} \sum_{i=1}^{m} -(d_i - h(x_i))^2 \\ &= \arg\min_{h \in H} \sum_{i=1}^{m} (d_i - h(x_i))^2 \end{aligned}$$

確率を予測するように学習する



確率を予測するように学習する

- 例: 生存確率を患者データから学習する

$$h_{ML} = \arg\max_{h \in H} \ln p(D \mid h)$$
$$= \arg\max_{h \in H} \ln \prod_{i=1}^{m} P(d_i \mid h, x_i) P(x_i)$$
$$= \arg\max_{h \in H} \sum_{i=1}^{m} \ln[P(d_i \mid h, x_i) P(x_i)]$$
$$\propto \arg\max_{h \in H} \sum_{i=1}^{m} \ln\left(h(x_i)^{d_i} (1 - h(x_i))^{1-d_i} P(x_i)\right)$$
$$= \arg\max_{h \in H} \sum_{i=1} d_i \ln h(x_i) + (1 - d_i) \ln(1 - h(x_i))$$

cross entropy

ベイズ学習が対象とする課題

- 先験的知識やバイアスを表現する最良の方法は？
- 決定木の枝狩りの正当化はどうやって？ 枝狩りはどのようにすればよいのか？
- 神経回路網では、どうして2乗誤差の最小化を測るのか？もっと別の関数を使わなくてもよいのか？使うとしたら、いつか？
- (Bayes-) 最適な分類器とは
- Naïve Bayes: 属性数が多いときの課題克服

最小記述長(minimum description length)

- Occam's razor: "最短仮説を選べ"
- MDL: 次を最小化する仮説を選ぶ

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

最小記述長(minimum description length)

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

木を記述するビット数
∝ 記述する符号の長さ

h が所与のとき、D を記述するビット数
∝ 誤分類データの個数

ベイズ学習が対象とする課題

- 先験的知識やバイアスを表現する最良の方法は？
- 決定木の枝狩りの正当化はどうやって？ 枝狩りはどのようにすればよいのか？
- 神経回路網では、どうして2乗誤差の最小化を測るのか？もっと別の関数を使わなくてもよいのか？使うとしたら、いつか？
- (Bayes-) 最適な分類器とは
- Naïve Bayes: 属性数が多いときの課題克服

Naïve Bayes 分類器

- 単純だが（だから？）よく知られた分類方法
- Bayes 定理 + 仮定 ***条件付独立***
  - 実際には成り立たないことが多い仮定
  - にもか関わらず，実際にはしばしばうまくいく
- 成功事例:
  - 文書分類
  - 診断

Bayes 定理を使うと

- 変数 x の属性$<a_1,\ldots,a_n>$が与えられたとき，x が属するクラス v を最尤推定するには?

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j \mid a_1, a_2, ..., a_n)$$
$$= \arg\max_{v_j \in V} \frac{P(a_1, a_2, ..., a_n \mid v_j) P(v_j)}{P(a_1, a_2, ..., a_n)}$$
$$= \arg\max_{v_j \in V} P(a_1, a_2, ..., a_n \mid v_j) P(v_j)$$

- 問題: 大量のデータが $P(a_1\ldots a_n|v_j)$ を評価するのに必要．パラメータ数が膨大 ($\Pi|A_i|$) (2値属性の場合、属性数が n なら 2^n 個)

Naïve Bayes 分類器

- Naïve Bayes **の仮定**: 属性同士は、属するクラスが所与なら独立
  - $P(a_1,\ldots,a_n|v_j) = P(a_1|v_j)\, P(a_2|v_j) \ldots P(a_n|v_j)$
  - ***条件付独立性*** (クラスが所与の時) とも
  - 推定すべきパラメータ数の削減:
    $\Pi|A_i|$ (=$O(2^n)$) $\rightarrow$ $\Sigma|A_i|$ (=O(n))
- この仮定のもと，v_{MAP} は

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j)$$

Naïve Bayes: アルゴリズム

Naïve_Bayes_Learn(事例)

それぞれの目標クラス v_j

$\hat{P}(v_j) = P(v_j)$ の推定値

各属性 a の各属性値 a_i ごとに

$\hat{P}(a_i|v_j) = P(a_i|v_j)$ の推定値

Classify_New_Instance(x)

$$v_{NB} = \arg\max_{v_j \in V} \hat{P}(v_j) \prod_i \hat{P}(a_i \mid v_j)$$

Naïve Bayes: 推定

- どうやって $P(v_j)$ と $P(a_i|v_j)$ を推定するか?
  - 統計学が教える標準的な方法
    - サンプルの頻度から確率を推定する
    - $P(v)$ の推定値は $\text{count}(v) / N$
    - $P(A|B)$ の推定値は $\text{count}(A \wedge B) / \text{count}(B)$
  - 例: 100 事例. 内訳 70 + と 30 –
    - P(+)=0.7 かつ P(-)=0.3
    - 70 個の正例のなかに，35 個で a_1=SUNNY
    - P(a_1=SUNNY|+)=0.5

例

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

$P(Y) = 9/14,$
$P(sunny \mid Y) = 2/9,$
$P(cool \mid Y) = 3/9,$
$P(high \mid Y) = 3/9,$
$P(strong \mid Y) = 3/9$

Naïve Bayes: 例

- 昔懐かしい *PlayTennis* , と新事例
  <Outlk=sun, Temp=cool, Humid=high, Wind=strong>
- 計算したいのは:

$$v_{NB} = \arg\max_{v_j \in V} \hat{P}(v_j) \prod_i \hat{P}(a_i \mid v_j)$$

- $\hat{P}(Y)\hat{P}(sun \mid Y)\hat{P}(cool \mid Y)\hat{P}(high \mid Y)\hat{P}(strong \mid Y) = 0.005$
  $\hat{P}(N)\hat{P}(sun \mid N)\hat{P}(cool \mid N)\hat{P}(high \mid N)\hat{P}(strong \mid N) = 0.021$

$$\Rightarrow v_{NB} = N$$

Naïve Bayes: 微妙なところ

- もし仮定が成り立たなかったら?
  - i.e. $P(a_1,\ldots,a_n|v_j) \neq P(a_1|v_j)\, P(a_2|v_j)\ldots P(a_n|v_j)$
- それでも、下記の〔弱い〕条件が成り立つ限り、予測値は Bayes 予測値と等価:

$$\arg\max_{v_j \in V} P(a_1 \mid v_j)P(a_2 \mid v_j)...P(a_n \mid v_j)P(v_j)$$
$$= \arg\max_{v_j \in V} P(a_1, a_2, ..., a_n \mid v_j)P(v_j)$$

- しかし、予測に伴う *確率* は 0 や 1 に極めて近い非現実的な値になりうる

Naïve Bayes: 微妙なところ

- もしも、属性値 a_i が観測されないクラス v_j があったら?
  - 推定値 $P(a_i|v_j)=0$ なぜなら $\text{count}(a_i \wedge v_j) = 0$?
  - 影響は甚大: これが 0 だと積は 0!
- 解: Laplace correction を用いる
  - $\hat{P}(a_i \mid v_j) = \dfrac{n_c + mp}{n + m}$
  - n 訓練例数. 但し $v = v_j$
  - n_c 訓練例数. 但し $v = v_j$ かつ $a = a_j$
  - p 先験確率の推定 $\hat{P}(a_i|v_j)$ (通常は一様分布)
  - m "仮想" 事例数

文書分類の学習

- 適用事例:
  - どのニュースが興味あるかを学習する
  - あるニュースがどのニュースグループのものかを判定できるように学習する
  - web ページをトピックで分類することを学習する
- Naïve Bayes がうまくいく
  - どうやって Naïve Bayes を用いるか?
  - キー: どう事例を表現するか? 属性は何か?

表現

- 属性 = 単語の出現位置
  - i.e. 属性 i は文書中の第 i 番目の単語位置
  - 属性値 = その位置に現れる単語
    - doc = $(a_1=w_1, a_i=w_k, \ldots, a_n=w_n)$
    - 注: 他の表現方法もある; e.g. 属性 = 特定の単語, 属性値 = 文書中のその単語の出現頻度
  - 更なる仮定: ある特定の単語がある確率は、その位置とは独立
    - $P(a_i=w_k|v_j) = P(a_m=w_k|v_j) = P(w_k|v_j)\ \forall i,m$
    - $P(\text{doc}|v_j)=P(a_1=w_1, a_2=w_2, \ldots, a_n=w_n|v_j)$
      $= P(w_1|v_j)^{TF(w1)} P(w_2|v_j)^{TF(w2)} \ldots P(w_n|v_j)^{TF(wn)}$
      ただし TF(w) は単語 w の出現頻度(term frequency)

Naïve Bayes による文書分類

- $v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j)$
  $= \arg\max_{v_j \in V} P(v_j) \prod_{w_k \in Voc} P(w_k \mid v_j)^{TF(w_k)}$
- $P(w_k \mid v_j) = \dfrac{n_{k,j} + 1}{n_j + |Voc|}$

アルゴリズム

procedure learn_naïve_bayes_text(*E*: 文書集合, V: クラス集合)
- Voc = E に現れる全ての単語とトークン
- E 中の w_k と V 中の v_j すべてについて，$P(v_j)$ と $P(w_k|v_j)$ を推定する:
  - N_j = クラス j の文書の数
  - N = 文書の総数
  - $P(v_j) = N_j/N$
  - n_{kj} = クラス j の全文書中の単語 w_k の出現数
  - n_j = クラス j 中の単語数
  - $P(w_k|v_j) = (n_{kj}+1)/(n_j+|Voc|)$

procedure classify_naïve_bayes_text(A: 文書)
- A から，Voc にない単語とトークンすべてを除去
- return $\text{argmax}_{vj \in V} P(v_j) \prod_i P(a_i|v_j)$

Twenty News Groups (Joachims 1996)

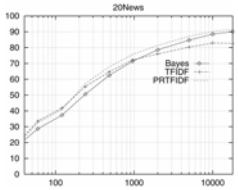
- 各グループ1000の訓練文書
- 新規の文書を、もとのnewsgroupに割振る

comp.graphics	misc.forsale
comp.os.ms-windows.misc	rec.autos
comp.sys.ibm.pc.hardware	rec.motorcycles
comp.sys.mac.hardware	rec.sport.baseball
comp.windows.x & rec.sport.hockey	rec.sport.hockey
alt.atheism	sci.space
soc.religion.christian	sci.crypt
talk.religion.misc	sci.electronics
talk.politics.mideast	sci.med
talk.politics.misc	
talk.politics.guns	

T. Joachims. *A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization.* In Proceedings of the 14th International Conference on Machine Learning, Nashville, TN, 1997, pp.143–151.

Twenty News Groups (Joachims 1996)

- Naive Bayes: 89% 分類正解率
  - 頻出単語上位100 個 (the and of ...) は除去
    - このように文法機能を担う単語や、文書を類別するのに有効でない単語を stop words として除去するのが普通
  - 頻度が3回に満たない単語は除去
  - 残った単語は、約 38,500 語



精度対訓練データ数 (1/3はテスト用にとりおいた)

NewsWeeder (Lang 1995)

- 目標概念 “usenet articles that I find interesting” を学習する
- ユーザはネットニュースを読むときに、興味深さの点数をつける
- 点数のついた文書を訓練例とする
- 点数を自動的につけた文書のうち上位 10% に興味深い文書が含まれる割合は、ユーザが普通に読む文書集合に含まれる割合の 3〜4倍高かった

Lang, K (1995). *NewsWeeder: Learning to Filter News.* Proceedings of the 12th International Conference on Machine Learning, 331-339, Lake Tahoe, CA.

まとめ: Bayes 学習

$$P(h \mid D) = \frac{P(D \mid h)\,P(h)}{P(D)}$$

- 学習アルゴリズムの俯瞰像:
  - ML: P(D|h) の最大化
  - MAP: $P(h|D) \propto P(D|h)\,P(h)$ の最大化
  - Bayes 最適分類器: $P(c|D) = \int P(c|h)P(h|D)\,dh$
- Gaussian ノイズ下の回帰:
  ⇔ 二乗誤差の最小化
- 二値事象の確率の学習
  ⇔ cross-entropy の最小化
- Occam's Razor:
  - P(h) = description-length(H) としての MAP
- Naive Bayes: 乱暴な仮説だが実用的