

情報意味論 (第6回)

ベイズ学習

慶應義塾大学工学部
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
- クラスの推定か確率の推定か
- MDL
- Naïve Bayes

Bayes の定理

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$



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

例 (Mitchell Chap. 6.2)

ある患者がガンの検査を受けたところ結果が陽性であった。
この患者には、本当に病変があるのだろうか？
なお、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。

$$P(\text{cancer}) = .008 \quad P(\neg \text{cancer}) = .992$$

$$P(+ | \text{cancer}) = .98 \quad P(- | \text{cancer}) = .02$$

$$P(+ | \neg \text{cancer}) = .03 \quad P(- | \neg \text{cancer}) = .97$$

$$P(+) = P(+ | \text{c}^+ r) P(\text{c}^+ r) + P(+ | \neg \text{c}^+ r) P(\neg \text{c}^+ r) = 0.0376$$

$$P(\text{cancer} | +) = \frac{P(+ | \text{cancer}) P(\text{cancer})}{P(+)} = .209$$

例 (Mitchell Exercise 6.1)

2回目の検査(2回は統計的に独立とする)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？

$$P(\text{cancer}) = .008 \quad P(\neg \text{cancer}) = .992$$

$$P(+ | \text{cancer}) = .98 \quad P(- | \text{cancer}) = .02$$

$$P(+ | \neg \text{cancer}) = .03 \quad P(- | \neg \text{cancer}) = .97$$

$$P(+_{1+2}) = P(+_{1+2} | \text{c}^+ r) P(\text{c}^+ r) + P(+_{1+2} | \neg \text{c}^+ r) P(\neg \text{c}^+ r) = 0.00858$$

$$P(\text{cancer} | +_{1+2}) = \frac{P(+_{1+2} | \text{cancer}) P(\text{cancer})}{P(+_{1+2})} = .896$$

良く使う公式

乗法の公式 (実は、条件付確率の定義！):

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

参考: 和事象に対しては

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

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

データ D を生成したらしい仮説 h を選択することができる！

大切な注: 条件付確率は因果関係(もしあれば)を反映するわけではない

目次

- Bayes 定理
- MAP と ML
- Bayes 最適分類器, Gibbs アルゴリズム
- クラスの推定か確率の推定か
- MDL
- Naïve Bayes

仮説選択

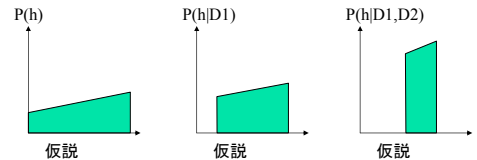
$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

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} :

$$\begin{aligned} h_{MAP} &= \arg \max_{h \in H} P(h|D) \\ &= \arg \max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg \max_{h \in H} P(D|h)P(h) \end{aligned}$$

ノイズがないときの事後確率の進展



かぞく 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:

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

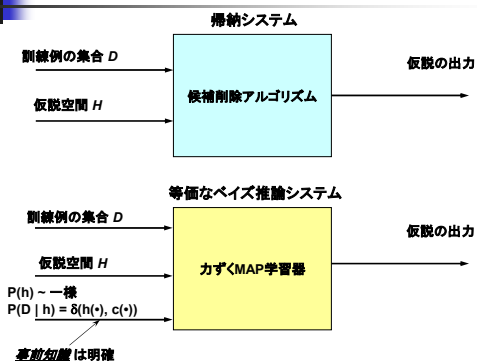
概念学習(FindS)との関係

- 通常概念学習の課題を考える
 - 事例集合 X , 仮説空間 H , 訓練事例 D
 - FindS は $VS_{H,D}$ から最も特殊な仮説を出力
- Bayes 規則が選ぶ MAP 仮説は？
- FindS は MAP 仮説を出力するか？

概念学習との関係

- 事例集合 $\{x_1, \dots, x_m\}$ を固定
- 訓練集合 D は $\{c(x_1), \dots, c(x_m)\}$
- 次のような h を選ぶ
 - $P(D|h)=1$, h が D と整合していれば
 - $P(D|h)=0$, そうでなければ
- $P(h)=1/|H|$, すなわち一様分布とする
 - $P(h|D)=1/|VS_{H,D}|$, h が D と整合していれば
 - $P(h|D)=0$, そうでなければ

一般には



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

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

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

ML推定のもう一つの解釈

- 現実世界では、事前確率分布は、未知か、計算不能か、存在しないと思われる
 - 例えば、文書における単語の生起頻度の事前分布はあるのだろうか？ 年齢、社会的背景、人口分布で大きく異なりうる
- 事前確率分布が存在しないとしたら、尤度最大化は自然な考え

目次

- Bayes 定理
- MAP と ML
- Bayes 最適分類器, Gibbs アルゴリズム
- クラスの推定か確率の推定か
- MDL
- Naïve Bayes

未知事例の最もありうる分類

- これまで、事例 D のもとでの最もありうる仮説を求めてきた(例: h_{MAP})。
- 未知事例の最もありうる(確率が高い)分類はどうなるのであろうか？
 - $h_{MAP}(x)$ は最もありうる分類ではない！
 - 次の例で、 x のもっともありうる類別は？
 - 3仮説: $P(h_1|D)=0.4$, $P(h_2|D)=0.3$, $P(h_3|D)=0.3$
 - 新事例: $h_1(x)=+$, $h_2(x)=-$, $h_3(x)=-$

Bayes 最適な分類器

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

注: Bayes 最適な分類器は H に含まれるとは限らない

注: 実行可能か?

注: 論文にはうまくいくと報告されているのだが、試してみるとMAPやMLと変わらない場合がある。どのような場合にそうなるか、興味のあるところである

例 (Mitchell Chap. 6.7)

$$\begin{array}{lll} P(h_1 | D) = .4 & P(- | h_1) = 0 & P(+ | h_1) = 1 \\ P(h_2 | D) = .3 & P(- | h_2) = 1 & P(+ | h_2) = 0 \\ P(h_3 | D) = .3 & P(- | h_3) = 1 & P(+ | h_3) = 0 \end{array}$$

それゆえ: $\sum_{h_i \in H} P(+ | h_i) P(h_i | D) = .4$

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

そして: $\arg \max_{c_j \in \{+, -\}} \sum_{h_i \in H} P(c_j | h_i) P(h_i | D) = -$

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

慶賀: もし仮説を事前分布 $P(h)$ に従ってランダムに選ぶと,

$$E[\text{error}_{\text{Gibbs}}] \leq 2E[\text{error}_{\text{BayesOptimal}}]$$

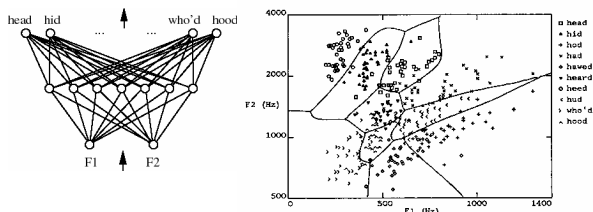
(詳細は "Machine Learning")

仮説の個数が多くて、ベイズ最適な分類器が計算できないときに有用

目次

- Bayes 定理
- MAP と ML
- Bayes 最適分類器, Gibbs アルゴリズム
- クラスの推定か確率の推定か
- MDL
- Naïve Bayes

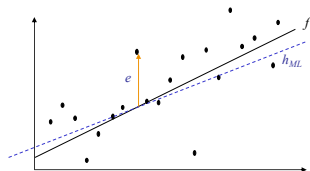
ニューラルネットでは



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

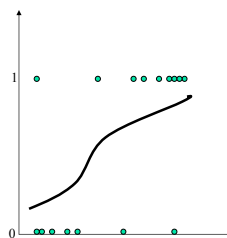
ならば (仮説):

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

実数値の学習(回帰) 続

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

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

$$\begin{aligned} h_{ML} &= \arg \max_{h \in H} \ln p(D | h) \\ &= \arg \max_{h \in H} \ln \prod_{i=1}^m P(d_i | h, x_i) P(x_i) \\ &= \arg \max_{h \in H} \sum_{i=1}^m \ln [P(d_i | h, x_i) P(x_i)] \\ &\stackrel{\text{cross entropy}}{=} \arg \max_{h \in H} \sum_{i=1}^m \ln (h(x_i)^{d_i} (1 - h(x_i))^{1-d_i} P(x_i)) \\ &= \arg \max_{h \in H} \sum_{i=1}^m d_i \ln h(x_i) + (1 - d_i) \ln(1 - h(x_i)) \end{aligned}$$

目次

- Bayes 定理
- MAP と ML
- Bayes 最適分類器, Gibbs アルゴリズム
- クラスの推定か確率の推定か
- MDL
- Naïve Bayes

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.

Occam の剃刀: 蛇足

- もっと以前から言われていたといわれる。表現も複数ある。
 - Wikipedia参照
 - 最近であれば "The philosophy of John Duns Scotus" の第8.2節
 - 古ければ, "The Myth of Occam's Razor" Mind, 27(107), 345-353 (1918)
- Albert Einstein: "Theories should be as simple as it is, but not simpler."
- 残差があるときは単に「単純に」とはいえない
- 「仮定」を入れても、それにより複数の現象が一つの理論で説明可能となるなら、これも単純化
 - 例: 未だ見えなかった分子による、現象の説明
- 単に「概念の個数」だけを数えるのでは、誤る。理論全体の長さも考えるべき
 - そして、残差の「長さ」も

最小記述長(minimum description length)

- Occam's razor: “最短仮説を選ぶ”

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

ex. 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

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

最小記述長 符号的解釈

- MDL: 次を最小化する仮説を選ぶ

$$\begin{aligned} h_{MAP} &= \arg \max_{h \in H} P(D|h)P(h) \\ &= \arg \min_{h \in H} -\log_2 P(D|h) - \log_2 P(h) \\ &= \arg \min_{h \in H} L_{C_2}(D|h) + L_{C_1}(h) \end{aligned}$$

蛇足: 確率と符号長

- 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 $U_{n>0}\{0,1\}^n$ への1-to-1 写像
- $L_C(x)$: 符号 C を用いた時の符号長(ビット)

- P : X 上で定義した確率分布

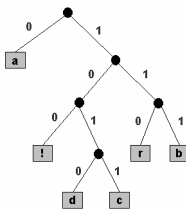
- $P(x)$: x の確率
- 観測値の系列 (通常は iid) $x_1, x_2, \dots, x_n, x^n$

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

蛇足: 確率と符号長: 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 |

蛇足: 確率と符号長: 最適符号

- ある符号 C の符号長の期待値

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

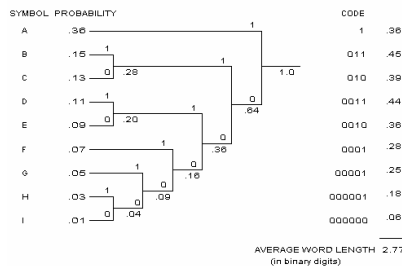
- 下界:

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

- 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、どう設計すればよいか?
 - Huffman 符号

蛇足: 確率と符号長: ハフマン符号



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

蛇足: 確率と符号長: 有限個数

- $\{1, 2, \dots, M\}$ 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 $1/M$
 - $\sim \log M$ ビット

蛇足: 確率と符号長: 無限集合なら

- 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 $\{1, \dots, 2^{\lceil \log k \rceil}\}$ の符号
 - 長さは合計 $\sim 2 \log k + 1$ ビット
 - 勿論、改善は可能...

蛇足: 確率と符号長: 双対性(?)

- 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:

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

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

再掲: 最小記述長 符号的解釈

- MDL: 次を最小化する仮説を選ぶ

$$\begin{aligned} h_{MAP} &= \arg \max_{h \in H} P(D|h)P(h) \\ &= \arg \min_{h \in H} -\log_2 P(D|h) - \log_2 P(h) \\ &= \arg \min_{h \in H} L_{C_2}(D|h) + L_{C_1}(h) \end{aligned}$$

目次

- Bayes 定理
- MAP と ML
- Bayes 最適分類器, Gibbs アルゴリズム
- クラスの推定か確率の推定か
- MDL
- 誤差関数
- Naïve Bayes

Naïve Bayes 分類器

- 単純だが(だから?)よく知られた分類方法
 - 単純な割には高精度
 - 単純なだけに、高速
- 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
 - それにも関わらず、実際にはしばしばうまくいく
- 成功事例:
 - 文書分類
 - 診断

Bayes 定理を使う場合の課題

- 変数 x の属性 $\langle a_1, \dots, a_n \rangle$ 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?

$$\begin{aligned} v_{MAP} &= \arg \max_{v_j \in V} P(v_j | a_1, a_2, \dots, a_n) \\ &= \arg \max_{v_j \in V} \frac{P(a_1, a_2, \dots, a_n | v_j) P(v_j)}{P(a_1, a_2, \dots, a_n)} \\ &= \arg \max_{v_j \in V} P(a_1, a_2, \dots, a_n | v_j) P(v_j) \end{aligned}$$

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

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
 - $P(a_1, \dots, a_n | v_j) = P(a_1 | v_j) P(a_2 | v_j) \dots P(a_n | v_j)$
 - **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:
 $\prod |A_i| (=O(2^n)) \rightarrow \sum |A_i| (=O(n))$
- この仮定のもと, v_{MAP} は

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

Naïve Bayes: アルゴリズム

Naïve_Bayes_Learn(事例集合)

それぞれの目標クラス v_j

$P^*(v_j) = P(v_j)$ の推定値

各属性 a の各属性値 a_i ごとに

$P^*(a_i | v_j) = P(a_i | v_j)$ の推定値

Classify_New_Instance(x)

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

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 $a_1=\text{SUNNY}$
 - $P(a_1=\text{SUNNY}|+)=0.5$

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 |

$$P(Y) = 9/14,$$

$$P(\text{sunny} | Y) = 2/9,$$

$$P(\text{cool} | Y) = 3/9,$$

$$P(\text{high} | Y) = 3/9,$$

$$P(\text{strong} | Y) = 3/9$$

Naïve Bayes: 例

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

$$v_{NB} = \arg \max_{v_j \in V} \hat{P}(v_j) \prod_i \hat{P}(a_i | v_j)$$
- $$\hat{P}(Y)\hat{P}(\text{sun} | Y)\hat{P}(\text{cool} | Y)\hat{P}(\text{high} | Y)\hat{P}(\text{strong} | Y) = 0.005$$

$$\hat{P}(N)\hat{P}(\text{sun} | N)\hat{P}(\text{cool} | N)\hat{P}(\text{high} | N)\hat{P}(\text{strong} | N) = 0.021$$

$$\Rightarrow v_{NB} = No$$

Naive Bayes: 条件付独立は必須か？

- もし仮定が成り立たなかったら？
 - i.e. if $P(a_1, \dots, a_n | v_j) \neq P(a_1 | v_j) P(a_2 | v_j) \dots P(a_n | v_j)$
- 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:

$$\arg \max_{v_j \in V} P(a_1 | v_j) P(a_2 | v_j) \dots P(a_n | v_j) P(v_j)$$

$$= \arg \max_{v_j \in V} P(a_1, a_2, \dots, a_n | v_j) P(v_j)$$
- しかし、予測時に求める **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: 困ること

- もしも、あるクラス v_j で属性値 a_i が観測されなかったら？
 - 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
 - $$\hat{P}(a_i | v_j) = \frac{n_c + mp}{n + m}$$
 - 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(の推定値) $P^*(a_i | v_j)$ (通常は一様分布)
 - m “仮想” 事例数(しばしば、当該属性 a の属性値の個数を用いる)
 $m=1$ とする方法がある。その方が結果がよいことがある

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 が結構うまくいく
 - どうやって Naive Bayes を用いるか?
 - ポイント: どう事例(すなわち、1 文書)を表現するか? 属性は何か?

事例の表現 or 属性

- 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
 - $\text{doc} = (a_1=w_1, a_i=w_k, \dots, a_n=w_n)$
 - 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
 - ある文書 $\text{doc} = (a_1=w_1, a_i=w_k, \dots, a_n=w_n)$ について
 - $P(a_i=w_k | v_j) = P(a_n=w_n | v_j) = P(w_k | v_j) \forall i, m$
 - $P(\text{doc} | v_j) = P(a_1=w_1, a_2=w_2, \dots, a_n=w_n | v_j)$

$$= P(w_1 | v_j)^{\text{TF}(w_1)} P(w_2 | v_j)^{\text{TF}(w_2)} \dots P(w_n | v_j)^{\text{TF}(w_n)}$$
 ただし $\text{TF}(w)$ は単語 w の doc における出現度数(term frequency)

Naïve Bayes による文書分類

- ある doc = $(a_1=w_1, \dots, a_i=w_k, \dots, a_n=w_n)$ につき

$$v_{NB} = \arg \max_{v_j \in V} P(v_j) \prod_{i=1}^{|\text{doc}|} P(a_i | v_j)$$

$$= \arg \max_{v_j \in V} P(v_j) \prod_{w_k \in V} P(w_k | v_j)^{TF(w_k, \text{doc})}$$

ただし、 $TF(w_k, \text{doc}) = \text{doc}$ 中の w_k の出現度数

- なお下記の推定値を使用; ただし、 $n_j = \text{クラス } v_j \text{ 中の全単語出現度数}$, $n_{k,j} = \text{クラス } v_j \text{ 中の単語 } w_k \text{ 出現度数}$

$$P(w_k | v_j) = \frac{n_{k,j} + 1}{n_j + |Voc|}$$

アルゴリズム

procedure learn_naive_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 (stop word は除く) の集合

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:

$N_j = \text{クラス } j \text{ の文書の数}$

$N = \text{文書の総数}$

$P(v_j) = N_j / N$

$n_{kj} = \text{クラス } j \text{ の全文書中の単語 } w_k \text{ の出現数}$

$n_j = \text{クラス } j \text{ 中の単語出現数}$

$P(w_k | v_j) = (n_{kj} + 1) / (n_j + |Voc|)$

procedure classify_naive_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 $\arg \max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j) = \arg \max_{v_j \in V} P(v_j) \prod_{w_k \in Voc} P(w_k | v_j)^{TF(w_k, A)}$

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

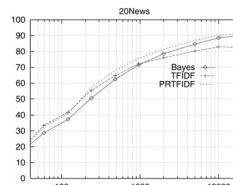
| | |
|-----------------------------------|--------------------|
| comp.graphics | misc.forsale |
| comp.os.ms-windows.misc | rec.autos |
| comp.sys.ibm.pc.hardware | rec.motorcycles |
| comp.sys.mac.hardware | rec.sport.baseball |
| comp.windows.x & rec.sport.hockey | rec.sport.hockey |
| alt.atheism | 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

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

- 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:
 - $P(h) = \text{description-length}(H)$ としての MAP
- Naive Bayes: 乱暴な仮説だが実用的