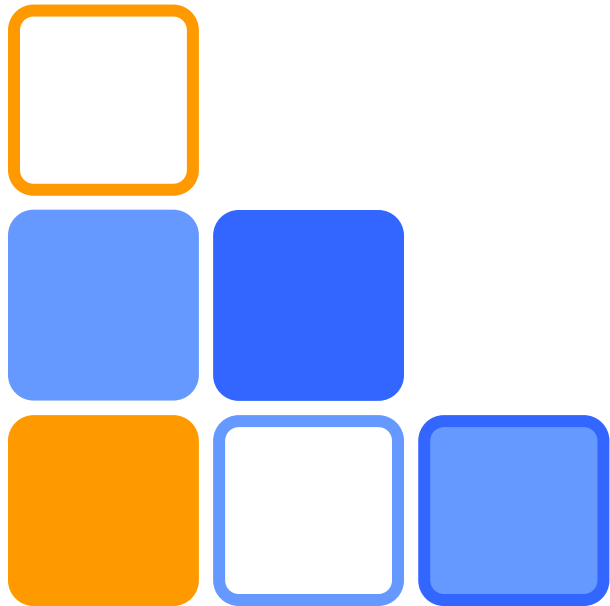


薬学情報処理演習 第7回

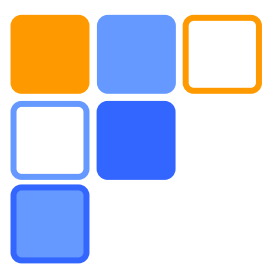
機械学習(2)

ナイーブベイズ分類



奥菌 透

コロイド・高分子物性学

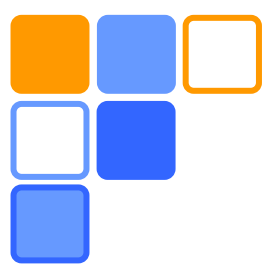


ベイズフィルター

- ベイズの定理を応用した分類アルゴリズム
- 応用例: スпамメールの排除
- メールに含まれる単語から以下を分類する
 - 通常メール
 - スпамメール
- 検出される単語の確率

検出語	迷惑メール(spam)	通常メール(ham)
秘密	0.7	0.1
無料	0.7	0.3
化学	0.1	0.4
物理	0.2	0.5

受信メール中の迷惑メール:通常メール=6:4とする。← 事前確率



ベイズの定理との対応関係

□ ベイズの定理

$$p(H_i|D_j) = \frac{p(D_j|H_i)p(H_i)}{p(D_j)}$$

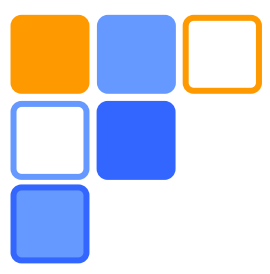
D_j の原因の確率
(事後確率)

尤度 事前確率

□ 仮定 $H_1(\text{spam}), H_2(\text{ham})$

□ データ

データ		尤度	$H_1(\text{spam}), H_2(\text{ham})$	
D_1	受信メールに「秘密」		0.7	0.1
D_2	受信メールに「無料」		0.7	0.3
D_3	受信メールに「化学」		0.1	0.4
D_4	受信メールに「物理」		0.2	0.5



ベイズ更新その1

- ベイズの式から ($i = 1, 2; j = 1, 2, 3, 4$)

$$p(H_i|D_j)p(D_j) = p(D_j|H_i)p(H_i)$$

- $i = 1, 2$ の式を辺々割り算

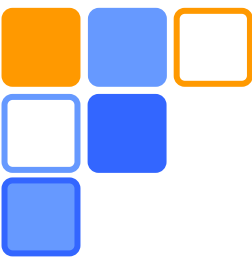
$$\frac{p(H_1|D_j)}{p(H_2|D_j)} = \frac{p(D_j|H_1)p(H_1)}{p(D_j|H_2)p(H_2)}$$

- 事前確率 $p_0(H_1) = 0.6, \quad p_0(H_2) = 0.4$

- データ D_1 を得た後の事後確率 $p_1(H_1|D_1), p_1(H_2|D_1)$

$$\frac{p_1(H_1|D_1)}{p_1(H_2|D_1)} = \frac{p(D_1|H_1)p_0(H_1)}{p(D_1|H_2)p_0(H_2)}$$

各データが独立と仮定。尤度は最初の表の値を利用



ベイズ更新その2

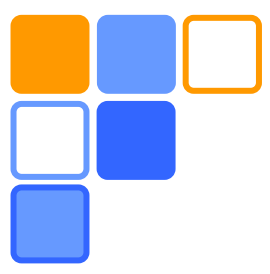
□ D_2 を得た後の事後確率の比

$$\frac{p_2(H_1|D_2)}{p_2(H_2|D_2)} = \frac{p(D_2|H_1)p_1(H_1|D_1)}{p(D_2|H_2)p_1(H_2|D_1)}$$

□ 同様に $\frac{p_3(H_1|D_4)}{p_3(H_2|D_4)} = \frac{p(D_4|H_1)p_2(H_1|D_2)}{p(D_4|H_2)p_2(H_2|D_2)}$

□ 最終的に

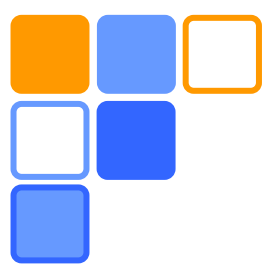
$$\frac{p_3(H_1|D_4)}{p_3(H_2|D_4)} = \frac{p_0(H_1)p(D_1|H_1)p(D_2|H_1)p(D_4|H_1)}{p_0(H_2)p(D_1|H_2)p(D_2|H_2)p(D_4|H_2)}$$



迷惑メールの判別

	H1 (spam)	H2 (ham)
事前確率	0.6	0.4
秘密D1	0.7	0.1
無料D2	0.7	0.3
化学D3	0.2	0.5
最後の事後確率比	$0.6 \times 0.7 \times 0.7 \times 0.2$	$0.4 \times 0.1 \times 0.3 \times 0.5$

$$\frac{p_3(H_1|D_4)}{p_3(H_2|D_4)} > 1 \quad \text{「迷惑メール」と判定}$$

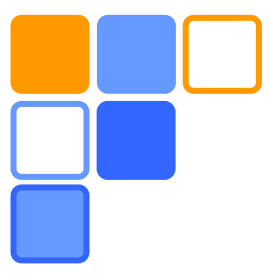


エクセルによる計算(1)

□ 事前確率と尤度の設定

D	H1(spam)	H2(ham)
秘密	0.7	0.1
無料	0.7	0.3
化学	0.1	0.4
物理	0.2	0.5

A	B	C
	H1(spam)	H2(ham)
事前確率	0.6	0.4



エクセルによる計算(2)

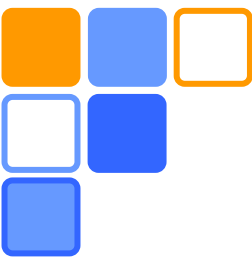
□ データ入力とナイーブベイズ計算

A	B	C
D	H1(spam)	H2(ham)
秘密	0.7	0.1
無料	0.7	0.3
化学	0.2	0.5

□ ナイーブベイズ分類

20

A	B	C
同時確率	=product(.....)	=product(.....)
判定結果	=if(B20>=C20, "spam", "ham")	



参考文献

- C.M.ビショップ 「パターン認識と機械学習 上
ベイズ理論による統計的予測」(丸善出版 2012)
- 涌井良幸、涌井貞美 「Excelでわかる機械学習超入門ーAIのモデルとアルゴリズムがわかる」(技術評論社 2019)
- 富谷昭夫 「これならわかる機械学習入門」(講談社 2021)
- 江崎貴裕 「データ分析のための数理モデル入門
本質をとらえた分析のために」(ソシム株式会社 2020)
- <https://phy-lum.com/opendata/machine-learning-dataset.html>(機械学習のためのデータセット)