確率と最適化

—ベイズ計算の基礎 (1)—

今回学ぶこと

- ベイズ則に基づく事後確率計算
- ２変数の場合の確率推論
- 多変数の場合の確率推論 (資料にて説明)

ある腫瘍(しゅよう)マーカーの検出性能

腫瘍マーカー: 「癌の進行とともに増加する生体因子のことで、主に血液中に遊離してくる因子を抗体を使用して検出する臨床検査のひとつである」(wikipedia)

- 腫瘍があるときにマーカー陽性になる確率 4/5
- 腫瘍があるときにマーカー陰性になる確率 1/5
- 腫瘍がないときにマーカー陽性になる確率 1/10
- 腫瘍がないときにマーカー陰性になる確率 9/10

偽陽性・偽陰性

- 本当は陽性なのに「陰性」の結果が出る→偽陰性
- 本当は陰性なのに「陽性」の結果が出る→偽陽性

臨床検査の場合、偽陰性を持つ検査は要注意である（腫瘍を見逃すと手遅れになる可能性があり得る)。

偽陰性を持つこの腫瘍マーカーは使えるのだろうか？どう使えばよいのか？

ベイズ的アプローチ：２変数の場合

手順

- 対象とする系を確率モデル化する。
- 観測値から、ベイズ則などを利用し事後確率を計算する。
- 事後確率に基づき、行動を決定する。

確率モデル化

- X: 腫瘍の有無を表す確率変数 ($X = 1$: 有、$X = 0$: 無)
- Y: マーカーの結果 ($Y = 1$: 陽性、$Y = 0$: 陰性)
- われわれの仮定：
  - $P_X(0) = 99/100, P_X(1) = 1/100$
  - $P_{Y|X}(0|0) = 9/10$
  - $P_{Y|X}(1|0) = 1/10$
  - $P_{Y|X}(0|1) = 1/5$
  - $P_{Y|X}(1|1) = 4/5$
- 観測結果：マーカーが陽性
- われわれの欲しい結果 (事後確率分布)：
  - $P_{X|Y}(0|1)$
  - $P_{X|Y}(1|1)$

われわれは何を知っていて何を知らないのか

- 事前分布 $P_X(x)$ を知っている。
- 条件付分布 $P_{Y|X}(y|x)$ を知っている。
- 事後分布 $P_{X|Y}(x|y)$ を知らない（これを計算したい）。

ベイズ則

$$P_{X|Y}(x|y) = \frac{P_X(x)P_{Y|X}(y|x)}{P_Y(y)}$$

確認ポイント：

- 自分で導けるか
- $P_Y(y)$ はどう求めればよいのか

問題

- $P_X(0) = 99/100, P_X(1) = 1/100$
- $P_{Y|X}(0|0) = 9/10$
- $P_{Y|X}(1|0) = 1/10$
- $P_{Y|X}(0|1) = 1/5$
- $P_{Y|X}(1|1) = 4/5$
- 観測結果：マーカーが陽性

この状況において

- $P_Y(1)$ を求めよ。
- 事後確率分布 $P_{X|Y}(0|1), P_{X|Y}(1|1)$ を求めよ。

$P_Y(1)$ を求める

$$
\begin{aligned}
P_Y(1) &= \sum_x P_{XY}(x,1) && (1)\\
&= \sum_x P_X(x)P_{Y|X}(1|x) && (2)\\
&= P_X(0)P_{Y|X}(1|0) + P_X(1)P_{Y|X}(1|1) && (3)\\
&= \frac{99}{100} \times \frac{1}{10} + \frac{1}{100} \times \frac{4}{5} && (4)\\
&= \frac{99+8}{1000} = \frac{107}{1000} && (5)
\end{aligned}
$$

ベイズ則を使う

$$P_{X|Y}(0|1) = \frac{P_X(0)P_{Y|X}(1|0)}{P_Y(1)} \quad (6)$$

$$= \frac{99}{100} \times \frac{1}{10} \times \frac{1000}{107} \quad (7)$$

$$= \frac{99}{107} \simeq 0.925234 \quad (8)$$

$$P_{X|Y}(1|1) = \frac{P_X(1)P_{Y|X}(1|1)}{P_Y(1)} \quad (9)$$

$$= \frac{1}{100} \times \frac{4}{5} \times \frac{1000}{107} \quad (10)$$

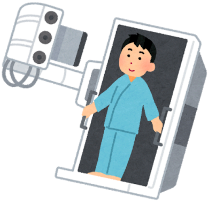
$$= \frac{8}{107} \simeq 0.0747664 \quad (11)$$

事前と事後

検査実施前: 腫瘍である確率：1%

検査実施後: 腫瘍である確率：7.4%

- false positive 確率がかなり高いので、直感的な値よりも低めに感じるかもしれない。
- 他の臨床検査に進む場合、この 7.4% を新たにこの人の事前確率として利用することもできる。



為替と株価

N 日の為替の状況と $N+1$ 日の T 社の株価の状況をわれわれは次のように確率モデル化した。

- X: 為替が上り調子 ($X=1$) か下り調子 ($X=0$) か
- Y: 株価が上り調子 ($Y=1$) か下り調子 ($Y=0$) か
- 同時分布 $P_{XY}(x,y)$

$x \backslash y$	0	1
0	0.4	0.1
1	0.1	0.4

- $X=1$ を観測した。
- 事後確率 $P_{Y|X}(1|1)$ を求めよ。

われわれは何を知っていて何をしらないか

- 同時分布 $P_{XY}(x, y)$ を知っている。
- 事後分布 $P_{Y|X}(y|x)$ を知らない（これを計算したい）。

事後確率の計算

$$P_X(0) = 0.5, \quad P_X(1) = 0.5$$

$$P_{Y|X}(1|1) = \frac{P_{XY}(1,1)}{P_X(1)} \tag{12}$$

$$= \frac{0.4}{0.5} = 0.8 \tag{13}$$

$$\tag{14}$$

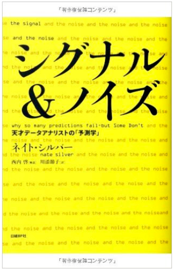
- ベイズ則を (直接) 使わない事後確率計算もあり得る
- 特に多変数の場合だと事後確率計算するときに周辺化計算を行う

まとめ

- ベイズ則に基づく事後確率計算
- ２変数の場合の確率推論
- 多変数の場合の確率推論 (資料にて説明)

いろいろな実例や関連する考え方をもっと知りたい人に

データサイエンティスト・アナリストに興味ある人は必読？
シグナル& ノイズ — 天才データアナリストの「予測学」
ネイト・シルバー, 日経 BP 社



- 「ビッグデータ」への期待と落とし穴
- マネーボールは何を語ったか
- ポーカーバブル

p.271: 「ベイズの定理を使うときには、世の中を確率的に見ることが要求される。たとえ確率の問題だと思いたくような問題でもだ。」