Învăţare automată

— Licenţă, anul III, 2016-2017, examenul parţial I —

Nume student: Suciu Ana Maria **Grupa:** B2

1. (Variabile aleatoare discrete: calcul mediei; formula lui Bayes; măsuri statistice folosite în clasificare; ipoteze MAP)

O anumită boală afectează una din 500 de persoane în medie. Identificarea persoanelor care au această boală se poate face cu ajutorul unei analize a sângelui, care costă 100 de dolari de persoană. Acestă analiză indică în cazul unui rezultat *pozitiv* faptul că se *poate* ca persoana respectivă să sufere de acea boală.

Testul/analiza are o *sensibilitate* (engl., *sensitivity* sau *recall*) perfectă — adică raportul dintre numărul instanţelor pozitive identificate ca atare de acel test şi numărul total de instanţe pozitive este 1 —, ceea ce înseamnă că pentru orice persoană care are boala respectivă, rezultatul testului este pozitiv cu probabilitate de 100%. Pe de altă parte, testul are o *specificitate* — raportul dintre numărul instanţelor negative identificate ca atare de acel test şi numărul total de instanţe negative — de 99%, adică o persoană care nu suferă de acea boală va avea cu probabilitate de 1% rezultatul testului pozitiv.

a. Se testează o persoană selectată în mod aleatoriu, iar rezultatul este pozitiv. Care este probabilitatea ca persoana respectivă să sufere de acea boală? Interpretaţi (în manieră calitativă) rezultatul obţinut.

b. Există şi un al doilea test, care costă 10.000 de dolari şi are atât sensibilitatea cât şi specificitatea de 100%. Dacă am cere ca toate persoanele detectate pozitiv la testul precedent să fie supuse acestui test mult mai scump, care ar fi costul mediu pentru testarea/analiza unui individ?

Sugestie: Definim următoarele variabile aleatoare:
B: ia valoarea 1/adevărat pentru persoanele care suferă de această boală şi 0/fals în caz contrar
T_1: rezultatul primului test, care poate fi + (în caz de boală) sau −
T_2: rezultatul celui de-al doilea test, care poate fi tot + sau −.
Folosind aceste variabile aleatoare, în vederea rezolvării problemei este bine ca mai întâi să formalizaţi enunţul sub forma următoare:

$$P(B) = \dots$$
$$P(T_1 = + \mid B) = \dots \qquad P(T_2 = + \mid B) = \dots$$
$$P(T_1 = + \mid \bar{B}) = \dots \qquad P(T_2 = + \mid \bar{B}) = \dots$$

$P(B) = \frac{1}{500}$ ✓ $\Rightarrow P(\bar{B}) = 1 - P(B) = \frac{500}{500} - \frac{1}{500} = \frac{499}{500}$

$P(T_1 = + \mid B) = 1$ ✓

$P(T_1 = + \mid \bar{B}) = \frac{1}{100}$ ✓

$P(T_2 = + \mid B) = 1$ ✓

$P(T_2 = + \mid \bar{B}) = 0$ ✓

$c_1 = 100\$$

$c_2 = 10.000\$$

a) $P(B \mid T_1 = +) = ?$ ← aplicăm formula lui Bayes

$$P(B \mid T_1 = +) = \frac{P(T_1 = + \mid B) \cdot P(B)}{P(T_1 = + \mid B) \cdot P(B) + P(T_1 = + \mid \bar{B}) \cdot P(\bar{B})} =$$

$$= \frac{1 \cdot \frac{1}{500}}{\frac{100}{1} \cdot \frac{1}{500} + \frac{1}{100} \cdot \frac{499}{500}} = \frac{1}{500} \cdot \frac{\frac{100}{50000}}{599} = \frac{100}{599} = 1{,}69$$ ✓

b) luăm o var aleatoare C, care reprezintă costul mediu

$$C = \begin{cases} c_1 & \text{- dacă persoanele au făcut doar } T_1 \\ c_1 + c_2 & \text{- dacă persoanele au făcut şi } T_1 \text{ şi } T_2 \end{cases}$$

$P(C = c_1) = P(T_1 = -)$ $P(C = c_1 + c_2) = P(T_1 = +)$

$$E[C] = \sum_{c \in C} c \cdot p(c) = c_1 \cdot P(C = c_1) + (c_1 + c_2) \cdot P(C = c_1 + c_2)$$

$= c_1 \cdot P(T_1 = -) + (c_1 + c_2) \cdot P(T_1 = +) =$

$= c_1 \cdot P(T_1 = -) + (c_1 + c_2) \cdot (1 - P(T_1 = -)) =$

$= P(T_1 = -)(c_1 - c_1 - c_2) + c_1 + c_2 =$

$= c_1 + c_2 - c_2 \cdot P(T_1 = -)$

formula prob. totale

$P(T_1 = +) = P(T_1 = + \mid B) \cdot P(B) + P(T_1 = + \mid \bar{B}) \cdot P(\bar{B})$

$= 1 \cdot \frac{1}{500} + \frac{1}{100} \cdot \frac{499}{500} = \frac{599}{50000}$

$E[C] = c_1 \cdot P(T_1 = -) + (c_1 + c_2) \cdot P(T_1 = +) =$

$= c_1 (1 - P(T_1 = +)) + (c_1 + c_2) \cdot P(T_1 = +) =$

$= c_1 + P(T_1 = +) \cdot (c_1 + c_2 - c_1) =$

$= c_1 + c_2 \cdot P(T_1 = +) =$

$= 100 + 10000 \cdot \frac{599}{50000} = 100 + \frac{599}{5} = 100 + 119.8 =$

$= 219.8 \simeq 220\$$

$\Rightarrow$ costul mediu $= 220\$$