SC4000/CZ4041/CE4041: Machine Learning

Solutions to L2 Tutorial Questions

Kelly KE

School of Computer Science and Engineering,

NTU, Singapore

Question 1

Question 1: Suppose A, B and C are three variables of binary values (0 or 1). Given the probabilities $P(A = 1, B = 0) = 0.4$, $P(A = 0) = 0.3$, and $P(A = 1, B = 1, C = 1) = 0.1$, compute the following probabilities:

1. $P(B = 1|A = 1)$.
2. $P(C = 0|B = 1, A = 1)$.

Question 1.1

Question 1: Suppose A, B and C are three variables of binary values (0 or 1). Given the probabilities $P(A = 1, B = 0) = 0.4$, $P(A = 0) = 0.3$, and $P(A = 1, B = 1, C = 1) = 0.1$, compute the following probabilities:

1. $P(B = 1|A = 1)$.
2. $P(C = 0|B = 1, A = 1)$.

- $P(B = 1|A = 1) = \frac{P(B=1,A=1)}{P(A=1)}$
- $P(A = 1) = 1 - P(A = 0) = 1 - 0.3 = 0.7$
- $P(A = 1) = 0.7$
  $= P(A = 1, B = 0) + P(A = 1, B = 1)$
  $= 0.4 + P(A = 1, B = 1)$
- $P(A = 1, B = 1) = 0.7 - 0.4 = 0.3$
- Thus, $P(B = 1|A = 1) = \frac{0.3}{0.7} = \frac{3}{7}$

Question 1.2

Question 1: Suppose A, B and C are three variables of binary values (0 or 1). Given the probabilities $P(A = 1, B = 0) = 0.4$, $P(A = 0) = 0.3$, and $P(A = 1, B = 1, C = 1) = 0.1$, compute the following probabilities:

1. $P(B = 1|A = 1)$.
2. $P(C = 0|B = 1, A = 1)$.

- $P(C = 0|B = 1, A = 1)$
  $= \frac{P(C=0,B=1,A=1)}{P(B=1,A=1)} = \frac{P(C=0,B=1,A=1)}{0.3}$
- $P(B = 1, A = 1) = 0.3$
  $= P(C = 0, B = 1, A = 1) + P(C = 1, B = 1, A = 1)$
  $= P(C = 0, B = 1, A = 1) + 0.1$
- $P(C = 0, B = 1, A = 1) = 0.3 - 0.1 = 0.2$
- Thus, $P(C = 0|B = 1, A = 1) = \frac{0.2}{0.3} = \frac{2}{3}$

Question 2

Question 2: Suppose that if a person has lung cancer, his/her probability of having gene X is 0.9, and if a person does not have lung cancer, his/her probability of having gene X is 0.2. The probability of a person having lung cancer is 0.01. Now, we know that a patient A has gene X.

1. Use Bayesian decision theory with 0/1 loss to predict whether the patient A has lung cancer or not.

2. Consider that costs of misclassification are different. Assume that the cost for correct decisions is 0, the cost of misclassifying a person who does not have lung cancer to be a patient with lung cancer is 0.007, and the cost of misclassifying a person who has lung cancer to be a healthy person is 1. Please use Bayesian decision theory with the predefined loss to predict whether the patient A has lung cancer or not.

Question 2 (cont.)

- Let Y be the random variable that represents a person having lung cancer or not (0 or 1)
  - $Y = 0$: a person does not have lung cancer
  - $Y = 1$: a person has lung cancer
- Let Z be the random variable that represents a person having gene X or not (0 or 1)
  - $Z = 0$: a person does not have gene X
  - $Z = 1$: a person has gene X

Question 2.1

- If a person has lung cancer, his/her probability of having gene X is 0.9

$$P(Z = 1|Y = 1) = 0.9$$

$$\downarrow$$

$$P(Z = 0|Y = 1) = 0.1$$

- If a person does not have lung cancer, his/her probability of having gene X is 0.2

$$P(Z = 1|Y = 0) = 0.2$$

$$\downarrow$$

$$P(Z = 0|Y = 0) = 0.8$$

- The probability of a person having lung cancer is 0.01

$$P(Y = 1) = 0.01$$

$$\downarrow$$

$$P(Y = 0) = 0.99$$

Question 2.1 (cont.)

- We know that patient A has gene X, i.e., $Z = 1$
- To estimate $P(Y = 0|Z = 1)$ v.s. $P(Y = 1|Z = 1)$
- With the $0/1$ loss

Predict $Y = 1$ if $P(Y = 1|Z = 1) > P(Y = 0|Z = 1)$

Predict $Y = 0$ otherwise

OR $P(Z = 1|Y = 1)P(Y = 1) > P(Z = 1|Y = 0)P(Y = 0)$

Question 2.1 (cont.)

- We have
  - $P(Z = 1|Y = 1)P(Y = 1) = 0.9 \times 0.01 = 0.009$
  - $P(Z = 1|Y = 0)P(Y = 0) = 0.2 \times 0.99 = 0.198$
  - $P(Z = 1|Y = 0)P(Y = 0) > P(Z = 1|Y = 1)P(Y = 1)$

  - $P(Y = 0|Z = 1) > P(Y = 1|Z = 1)$
- Prediction: patient A does not have lung cancer

Question 2.2

Question 2: Suppose that if a person has lung cancer, his/her probability of having gene X is 0.9, and if a person does not have lung cancer, his/her probability of having gene X is 0.2. The probability of a person having lung cancer is 0.01. Now, we know that a patient A has gene X.

1. Use Bayesian decision theory with 0/1 loss to predict whether the patient A has lung cancer or not.
2. Consider that costs of misclassification are different. Assume that the cost for correct decisions is 0, the cost of misclassifying a person who does not have lung cancer to be a patient with lung cancer is 0.007, and the cost of misclassifying a person who has lung cancer to be a healthy person is 1. Please use Bayesian decision theory with the predefined loss to predict whether the patient A has lung cancer or not.

Question 2.2

- λ_{ik}: loss (or cost) of action a_i that predicts $Y = i$ while the true label is k
- The cost for correct decisions is 0

$$\lambda_{00} = 0, \text{ and } \lambda_{11} = 0$$

- Cost of misclassifying a person not having lung cancer to be a patient with lung cancer is 0.007

$$\lambda_{10} = 0.007$$

- Cost of misclassifying a person having lung cancer to be a healthy person is 1

$$\lambda_{01} = 1$$

Question 2.2 (cont.)

- Take action a^* if $a^* = \arg\min_{a_i} R(a_i|\boldsymbol{x})$, where $R(a_i|\boldsymbol{x}) = \sum_{c=0}^{C-1} \lambda_{ic} P(Y = c|\boldsymbol{x})$
- Expected risk of predicting A not having lung cancer:

$$R(a_0|Z=1) = \lambda_{01} \times P(Y=1|Z=1) + \lambda_{00} \times P(Y=0|Z=1)$$

$$= 1 \times \frac{P(Z=1|Y=1) \times P(Y=1)}{P(Z=1)} + 0$$

$$= 1 \times \frac{0.9 \times 0.01}{P(Z=1)} = \frac{0.009}{P(Z=1)}$$

Question 2.2 (cont.)

- Expected risk of predicting A having lung cancer:

$$R(a_1|Z=1) = \lambda_{11} \times P(Y=1|Z=1) + \lambda_{10} \times P(Y=0|Z=1)$$

$$= 0 + 0.007 \times \frac{P(Z=1|Y=0) \times P(Y=0)}{P(Z=1)}$$

$$= 0.007 \times \frac{0.198}{P(Z=1)}$$

$$= \frac{0.0014}{P(Z=1)} \quad < \quad R(a_0|Z=1) = \frac{0.009}{P(Z=1)}$$

- Prediction: patient A has lung cancer

Thank you!