Tema 6. Clasificadores Bayesianos. Ejercicios

Pedro Larrañaga, Iñaki Inza, Abdelmalik Moujahid
Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad del País Vasco–Euskal Herriko Unibertsitatea

1. Supongamos que el 20 por ciento de la población masculina con edades superiores a 60 años tiene cáncer de próstata. Para tratar de diagnosticarlo nos fijamos en los niveles de expresión de dos genes concretos que denotamos por G_1 y G_2. Ambas variables se van a considerar discretizadas a dos valores.

   Se conoce que la probabilidad de que un individuo con cáncer de próstata tenga un valor alto de G_1 es de 0.75, mientras que para individuos sin cáncer de próstata el valor alto de G_1 tan solo se da con probabilidad 0.20.

   Por lo que respecta a G_2, se tiene que entre los que presentan cáncer de próstata, el valor alto de G_2 se da con probabilidad 0.10, y entre los que no presentan cáncer de próstata dicho valor alto se da con probabilidad 0.60.

   Efectuar el diagnóstico para un individuo con valor de expresión alta en G_1 y baja en G_2. Explicar las premisas sobre las que se basa el paradigma utilizado para llevar a cabo el diagnóstico.

2. Se conoce que el 40 por ciento de los alumnos aprueban la asignatura A en la primera convocatoria.

   Supongamos que en el modelo que hemos construido para predecir si un alumno va a aprobar la asignatura A en su primera convocatoria tenemos dos variable predictivas que se suponen independientes dado el valor de la clase. Las dos variables predictivas son *aprobar en primera convocatoria la asignatura* B y *aprobar en primera convocatoria la asignatura* C. Ambas asignaturas B y C deben de ser cursadas en un curso anterior a la asignatura A siendo además prerrequisito de la misma.

   Entre los que aprueban en primera convocatoria la asignatura A el 75 por ciento habia aprobado tambien en primera convocatoria la asignatura B, mientras que entre los que suspenden en primera convocatoria la asignatura A tan sólo el 10 por ciento habia aprobado en primera convocatoria la asignatura B.

   Por lo que respecta a la asignatura C, se tiene que entre los que aprueban en primera convocatoria la asignatura A el 55 por ciento habia aprobado tambien en primera convocatoria la asignatura C, mientras que entre los que suspenden en primera convocatoria la asignatura A tan sólo el 20 por ciento habia aprobado en primera convocatoria la asignatura C.

   Efectuar el diagnóstico para un alumno que aprobó la asignatura B en primera convocatoria, pero necesitó mas de una convocatoria para superar la asignatura C.

3. Supongamos que se conoce que la probabilidad a priori de recidiva para un enfermo que sufre por vez primera un ataque de corazón es de 0.30.

Supongamos que en el modelo que tenemos construido para tratar de predecir la recidiva, tenemos dos variables predictoras que se suponen independientes dado el valor de la clase. Las dos variables predictivas son la *realización de ejercicio moderado de manera asidua* y la *edad del paciente*. Mientras que la primera es discreta con dos posibles valores (si, no), la segunda es continua y está para cada valor de la clase distribuida normalmente.

Entre los que no tienen recidiva, el 80 por ciento efectúa ejercicio moderado de manera asidua, mientras que entre los que tienen recidiva dicha probabilidad se reduce al 10 por ciento.

La edad de los pacientes que no tienen recidiva sigue un modelo normal con esperanza matemática de 60 años y desviación típica de 10 años, mientras que en el caso de los pacientes con recidiva los parámetros del modelo normal son 75 años de esperanza matemática y 15 de desviación típica.

Efectuar el diagnóstico para un paciente de 58 años que no efectúa ejercicio moderado de manera asidua.

4. Supongamos que en una determinada carrera universitaria el 30 por ciento de los individuos que la comienzan son capaces de terminarla en 6 o menos años.

   Además se conoce que el 75 por ciento de los individuos que la terminan en 6 o menos años tuvieron una nota de selectividad superior o igual a 8 puntos. Por otra parte tan sólo el 20 por ciento de los individuos que necesitaron más de 6 años para terminarla entraron en la Universidad con una nota de selectividad superior o igual a 8 puntos.

   Si tenemos en cuenta el número de créditos obtenidos durante el primer año de estancia en la Universidad, obtenemos que el 85 por ciento de los individuos que necesitaron 6 o menos años para terminar la carrera, fueron capaces de conseguir al menos 30 créditos, mientras que tan sólo el 25 por ciento de los individuos que necesitaron más de 6 años para finalizar sus estudios universitarios, fueron capaces de conseguir al menos 30 créditos.

   Suponemos que los resultados proporcionados por las dos pruebas anteriores son, una vez conocido el status del individuo (≤ 6 años, > 6 años) independientes.

   Conocidos los datos de dos individuos, Juan y Maria, al finalizar su primer año de estancia universitaria:

   - Juan: nota de selectividad 6.8, 25 créditos en su primer año
   - Maria: nota de selectividad 8.1, 35 créditos en su primer año

   se quiere calcular, para cada uno de ellos, la probabilidad de terminar la carrera en 6 o menos años.

   Explicar las características teóricas fundamentales del paradigma clasificatorio utilizado.

5. Para determinar si un individuo está enfermo o sano en relación con una determinada enfermedad, a dicho individuo se le efectúan dos pruebas.

   Los resultados de la primera prueba se expresan en términos de positivo o negativo. Se sabe que si el individuo está enfermo, la probabilidad de que el resultado de la primera prueba sea positivo es de 0.95, mientras que si el individuo está

sano, un resultado positivo en la primera prueba tiene asociada una probabilidad de 0.02.

La segunda prueba tiene como posibles resultados: alto, normal y bajo. Se tiene que si el individuo está enfermo la probabilidad de que el resultado de esta segunda prueba sea alto es de 0.85, mientras que dicha probabilidad se reduce hasta 0.10 para resultados normales. Si el individuo está sano, la segunda prueba tiene un resultado alto con probabilidad 0.20 y normal con probabilidad 0.70.

Los resultados proporcionados por las dos pruebas anteriores son, una vez conocido el status del individuo (enfermo o sano), independientes. Además la probabilidad de padecer la enfermedad anterior en la población estudiada es de 0.01.

Supongamos que a un paciente susceptible de padecer la anterior enfermedad se le efectúan las dos pruebas citadas anteriormente, obteniéndose que en la primera de ellas el resultado es positivo, mientras que en la segunda el resultado es alto.

Se solicita llevar a cabo un diagnóstico de dicho individuo, justificando el paradigma utilizado.

6. Cierto motor puede tener una avería eléctrica con probabilidad 10^{-3} o mecánica, en este caso con probabilidad 10^{-5}. El hecho de que se produzca un tipo de avería hace que no se produzca una avería de otro tipo.

   Cuando hay avería eléctrica se enciende un piloto luminoso el 95 por ciento de las veces; cuando la avería es de tipo mecánico el piloto luminoso se enciende el 99 por ciento de las veces, y cuando no hay avería el piloto luminosos se enciende –causando una falsa alarma– en un caso por millón.

   Cuando no hay avería, la temperatura está elevada en el 17 por ciento de los casos y reducida el 5 por ciento de las veces. En el resto de los casos, la temperatura se encuentra en los límites de normalidad. Cuando hay avería eléctrica está elevada en el 90 por ciento de los casos y reducida en el 1 por ciento. Finalmente cuando hay avería mecánica, la temperatura está elevada el 10 por ciento de los casos y reducida el 40 por ciento de las veces.

   El funcionamiento del piloto es independiente de la temperatura para cada diagnóstico.

   Si se enciende el piloto y la temperatura está por debajo de su valor normal, se pide efectuar el diagnóstico del motor.

7. Dadas 5 variables aleatorias predictoras discretas: X_1, X_2, X_3, X_4, X_5 y una variable a predecir discreta: C, y teniendo en cuenta que se verifican las siguientes inecuaciones en relación con la cantidad de información mutua entre cada variable predictora y la clase:

$$I(X_3, C) > I(X_1, C) > I(X_4, C) > I(X_5, C) > I(X_2, C)$$

   así como las siguientes relaciones de desigualdad entre las cantidades de información mutua entre parejas de variables aleatorias predictoras condicionadas a la variable clase:

$$I(X_3, X_4|C) > I(X_2, X_5|C) > I(X_1, X_3|C) > I(X_1, X_2|C) > I(X_2, X_4|C)$$

$$> I(X_2, X_3|C) > I(X_1, X_4|C) > I(X_4, X_5|C) > I(X_1, X_5|C) > I(X_3, X_5|C)$$

Obtener la estructura k-DB siguiendo el algoritmo propuesto por Sahami (1996).

Una vez que se obtenga dicha estructura establecer la factorización de la distrubción de probabilidad condicionada: $p(c|x_1, x_2, x_3, x_4, x_5)$ a partir de la cual se llevarían a cabo las clasificaciones de los distintos patrones nuevos.