Pràctica Tema 2. Detecció

Processament de Senyal Audiovisual i de Comunicacions (PSAVC)

Autors: Meritxell Lamarca, Montse Nájar, Marga Cabrera

PRÁCTICA: DETECCIÓN EFICIENTE DE COVID-19 MEDIANTE TÉCNICAS DE *GROUP TESTING*

1. Introducción

Group Testing (GT) o testeo en grupo es una técnica de cribado que se propuso en la Segunda Guerra Mundial para la identificación de los soldados que estaban infectados de sífilis [1]. Es una técnica que se emplea para la identificación de los individuos infectados (o positivos) dentro de una población donde la gran mayoría de individuos no están infectados (o son negativos). La idea es que cuando la tasa de infectados es baja existen métodos mucho más eficientes para el cribado de la población que el test uno a uno de todos los individuos.

GT explota el hecho de que cuando seleccionamos un subconjunto de individuos de la población y realizamos un único test conjunto de todos ellos para detectar una característica especial o defecto, si el resultado del test es negativo (y tiene una fiabilidad total) entonces tenemos la garantía de que todos los individuos que contribuyeron al test son todos ellos negativos. De este modo, con un único test se consigue etiquetar como no infectados todos los individuos del subconjunto que participó en el test. En el caso de la aplicación original en [1], la idea era hacer un único test a una muestra que combinara la sangre de varios soldados, de manera que el test daba positivo si alguno de los soldados implicados (no importaba cuántos de ellos) estaba infectado, y daba negativo si todos ellos estaban libres de la enfermedad. En general, GT requiere la realización de varios test sobre los mismos individuos, ya sea secuencialmente o en paralelo.

Las técnicas de GT se han empleado en entornos muy diversos como el chequeo de productos industriales, comunicaciones en acceso múltiple, adquisición de datos en redes de sensores, análisis sanguíneos, compresión de imágenes, escaneado de espectro, etc. En esta práctica se propone el estudio de una técnica de GT para el cribado de la población en un brote de la pandemia COVID-19 [2].

Centrémonos en el problema de la identificación de los individuos que son asintomáticos, pero están infectados por COVID-19. Como todos sabemos es muy importante la detección de estos individuos porque suponen un riesgo muy grande de propagar la enfermedad a personas sanas. Por otra parte, debemos tener en cuenta que en situaciones de pandemia puede haber restricciones en la disponibilidad de tests, por lo que es importante conseguir la detección empleando el mínimo número de test. Vamos a ver que la técnica de GT analizada en esta práctica nos puede mejorar simultáneamente las prestaciones tanto en términos de detección como en términos del número de pruebas a realizar cuando la tasa de individuos infectados es baja. De acuerdo con [2], la tasa de individuos asintomáticos que están infectados está típicamente por debajo del 20%.

2. Formulación del problema

2.1 Test básico

Tanto la estrategia clásica como los esquemas basados en GT emplean el mismo tipo de test básico para detectar COVID, la única diferencia entre los distintos métodos es si el test se aplica sobre el espécimen de un individuo o sobre la mezcla de los especímenes de un grupo de

individuos. Consideremos este **test básico (TB)**, denominemos las dos posibles hipótesis de este test

$\mathcal{H}_0^B$: la muestra testeada está libre del virus (es negativa)

$\mathcal{H}_1^B$: la muestra testeada incluye la contribución de uno (o varios) individuos infectados (es positiva)

El test básico entrega un resultado que no depende de si se ha aplicado sobre el espécimen de un individuo o sobre la mezcla de especímenes de los K individuos en un grupo. De acuerdo con [2] es razonable emplear para este test básico de COVID-19 la distribución siguiente para la función de test y :

$\mathcal{H}_0^B$: Individuo o grupo de individuos no infectado: $y \sim N(m_0, \sigma_0^2)$, m_0=5; $\sigma_0^2 = 1$

$\mathcal{H}_1^B$: Individuo o grupo de individuos infectado: $y \sim N(m_1, \sigma_1^2)$, m_1=10; $\sigma_1^2 = 1.2$

La decisión en el TB se toma comparando esta función de test con un umbral γ que se elige para alcanzar el compromiso deseado entre la probabilidad de falsa alarma y la probabilidad de detección[1]:

$$\begin{matrix} \widehat{\mathcal{H}_1^B} \\ y \gtrless \gamma \\ \widehat{\mathcal{H}_0^B} \end{matrix}$$

2.2 Detectores propuestos

En nuestra aplicación de GT emplearemos el test básico para etiquetar todos los individuos de la población como "negativos" o "positivos". Si nos centramos en un individuo en concreto nuestro problema de detección es binario: las dos hipótesis entre las que hemos de decidir son

$\mathcal{H}_0$: el individuo no está infectado (es negativo)

$\mathcal{H}_1$: el individuo está infectado (es positivo)

En esta práctica vamos a comparar las prestaciones del detector clásico con el detector de GT propuesto. Consideremos una población de N individuos. Indiquemos con $\widehat{\mathcal{H}_0^C}, \widehat{\mathcal{H}_1^C}$ las decisiones del detector clásico y con $\widehat{\mathcal{H}_0^{GT}}, \widehat{\mathcal{H}_1^{GT}}$ las del detector de Group Testing o GT. Entonces,

- En el **detector clásico** se realiza un único test separado para cada individuo (es decir, un total de N tests):

  Se realiza el test básico para cada individuo por separado y el resultado de este test es la decisión adoptada. Es decir:

  …se decide $\widehat{\mathcal{H}_0^C}$ si el detector básico ha decidido $\widehat{\mathcal{H}_0^B}$

  …se decide $\widehat{\mathcal{H}_1^C}$ si el detector básico ha decidido $\widehat{\mathcal{H}_1^B}$

[1] En el entorno sanitario no se emplea normalmente la notación P_{FA} y P_D sino que la probabilidad $\Pr(\widehat{\mathcal{H}_0^B}|\mathcal{H}_0^B)$ se denomina de especificidad y la probabilidad $\Pr(\widehat{\mathcal{H}_1^B}|\mathcal{H}_1^B)$ se denomina sensibilidad. Por claridad, en esta práctica mantendremos la nomenclatura empleada en PSAVC.

- El **detector de GT** propuesto se realiza en dos fases:

    1. Se reparten los individuos a estudiar en N/K grupos de K miembros.

    2. En una **primera fase**, los especímenes de los K individuos se juntan y se someten a un test básico. En esta fase se realizan N/K tests básicos. Indiquemos como $\widehat{\mathcal{H}_0^{B1}}$ y $\widehat{\mathcal{H}_1^{B1}}$ las decisiones tomadas en este primer test básico.

    3. En el caso que el test conjunto dé negativo, se decide que los K individuos participantes en él son no infectados y el procedimiento finaliza. Es decir:

        Si el detector básico ha decidido $\widehat{\mathcal{H}_0^{B1}}$ entonces para cada uno de los K miembros del grupo se decide $\widehat{\mathcal{H}_0^{GT}}$.

    4. En el caso de que el test conjunto dé positivo, se realiza una **segunda fase** en la que cada uno de los K individuos participantes se somete a un test básico individual. La decisión tomada sobre cada individuo es el resultado de este segundo test. Es decir:

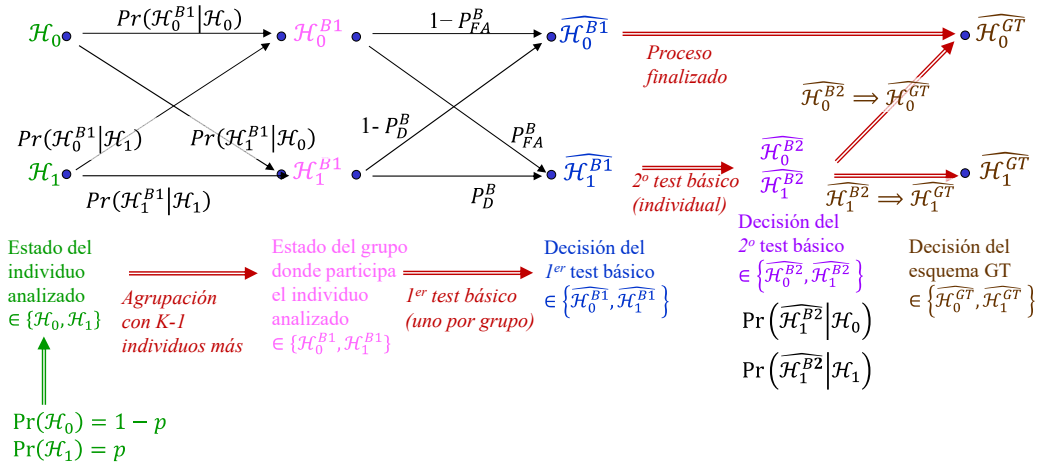
        Si el detector básico ha decidido $\widehat{\mathcal{H}_1^{B1}}$ entonces se realiza un test básico adicional para cada uno de los K miembros del grupo. Indiquemos como $\widehat{\mathcal{H}_0^{B2}}$ y $\widehat{\mathcal{H}_1^{B2}}$ las decisiones tomadas en este segundo test básico. Entonces,

        … se decide $\widehat{\mathcal{H}_0^{GT}}$ si el detector básico ha decidido $\widehat{\mathcal{H}_0^{B2}}$

        … se decide $\widehat{\mathcal{H}_1^{GT}}$ si el detector básico ha decidido $\widehat{\mathcal{H}_1^{B2}}$

        Obsérvese que el número de TB realizados en esta segunda fase es variable, puesto que sólo se realizan los K test adicionales cuando la primera fase ha decidido que en el grupo había algún individuo positivo.

La siguiente figura representa un diagrama del esquema de GT y las probabilidades implicadas:



La estrategia global del procedimiento de GT es eliminar en una primera fase del test la mayoría de los individuos no infectados y centralizar la fase de test individual en aquellos individuos que pueden ser positivos. A modo de ejemplo, si la población tuviera N=1000 individuos, se hicieran grupos de $K = 5$ y la probabilidad de que el test básico en la primera fase decidiera "positivo"

fuera de 0.1 entonces el número de tests realizados en la primera fase sería de N/K=200 y el número promedio de tests en la segunda fase sería de 200·0.1· K=100. Es decir, se realizarían un total de 300 TB en vez de los 1000 que serían necesarios con el cribado clásico individual. Evidentemente, hay que asegurar que se puede realizar esta reducción del número de tests sin que ello perjudique las prestaciones del detector en términos de la ROC (*receiver operation characteristic).* En esta práctica veremos que ello es posible cuando $\Pr(\mathcal{H}_1)$ es pequeña.

3. Estudio previo

Vamos a considerar que la probabilidad de que un individuo asintomático sea positivo es $p = \Pr(\mathcal{H}_1)$. Emplee los parámetros de la distribución del test básico, así como p y el tamaño del grupo K para evaluar las prestaciones del detector basado en GT:

3.1. Halle la expresión de la probabilidad de falsa alarma y detección en el test básico, $P_{FA}^B = \Pr\left(\widehat{\mathcal{H}_1^B}\middle|\mathcal{H}_0^B\right)$ y $P_D^B = \Pr\left(\widehat{\mathcal{H}_1^B}\middle|\mathcal{H}_1^B\right)$, en función del valor del umbral γ. Halle la expresión de P_D^B en función de P_{FA}^B y de los parámetros del modelo $m_0, \sigma_0^2, m_1, \sigma_1^2$. Estos valores serán también los del detector clásico: $P_{FA}^C = \Pr\left(\widehat{\mathcal{H}_1^C}\middle|\mathcal{H}_0\right) = P_{FA}^B$ y $P_D^C = \Pr\left(\widehat{\mathcal{H}_1^C}\middle|\mathcal{H}_1\right) = P_D^B$.

3.2. Centrémonos en un individuo en concreto. Halle la probabilidad de detección y de falsa alarma en método de GT, denominados $P_{FA}^{GT} = \Pr\left(\widehat{\mathcal{H}_1^{GT}}\middle|\mathcal{H}_0\right)$ y $P_D^{GT} = \Pr\left(\widehat{\mathcal{H}_1^{GT}}\middle|\mathcal{H}_1\right)$ en términos de P_{FA}^B, P_D^B, p y K. Para ello siga los pasos:

a. Halle la probabilidad de que el grupo de K individuos donde está nuestro individuo de interés incluya alguno que está infectado en cada una de las hipótesis, es decir $\Pr(\mathcal{H}_1^{B1}|\mathcal{H}_0)$ y $\Pr(\mathcal{H}_1^{B1}|\mathcal{H}_1)$.

b. Halle la probabilidad de que el primer test básico dé un resultado positivo en cada una de las hipótesis, es decir $\Pr\left(\widehat{\mathcal{H}_1^{B1}}\middle|\mathcal{H}_0\right)$ y $\Pr\left(\widehat{\mathcal{H}_1^{B1}}\middle|\mathcal{H}_1\right)$.

c. Considerando que el primer test básico ha sido positivo y se ha realizado la segunda fase del procedimiento de GT, halle la probabilidad de que el segundo test básico dé un resultado positivo en cada una de las hipótesis, es decir $\Pr\left(\widehat{\mathcal{H}_1^{B2}}\middle|\mathcal{H}_0\right)$ y $\Pr\left(\widehat{\mathcal{H}_1^{B2}}\middle|\mathcal{H}_1\right)$.

d. Halle la probabilidad de que globalmente un individuo sea etiquetado como infectado en cada una de las hipótesis, es decir $P_{FA}^{GT} = \Pr\left(\widehat{\mathcal{H}_1^{GT}}\middle|\mathcal{H}_0\right)$ y $P_D^{GT} = \Pr\left(\widehat{\mathcal{H}_1^{GT}}\middle|\mathcal{H}_1\right)$. Nótese que para que un individuo sea etiquetado como positivo se requiere que se realicen las dos fases y se decida "infectado" tanto en el primer test básico como en el segundo.

3.3. Calcule el número medio de tests que se realizan para cada individuo en el esquema de GT en función de P_{FA}^B, P_D^B, p y K. Nótese que este número de tests no depende de la decisión tomada globalmente ($\widehat{\mathcal{H}_0^{GT}}$ o $\widehat{\mathcal{H}_1^{GT}}$) sino sólo de la decisión tomada en la primera fase ($\widehat{\mathcal{H}_0^{B1}}$ o $\widehat{\mathcal{H}_1^{B1}}$).

4. Trabajo de laboratorio

4.1. Escriba un programa de Matlab que tenga como parámetros de entrada los valores de p y K y realice las siguientes tareas

Detector clásico:

- Definir un vector con valores de P_{FA}^{B} (es decir, de P_{FA}^{C}) que cubran el margen entre 10^{-5} y 1.

[comandos Matlab `qfunc, qfuncinv,logspace`]
[Nota: Es necesario el Communications Toolbox]

- Calcular el valor de, P_{D}^{C} para cada valor de P_{FA}^{C}, y guardarlos en un vector.
- Representar en una figura (figura 1) la ROC del detector clásico.

[comandos Matlab `figure, plot`]

Detector de Group Testing:

Para tamaños de grupo K=2,3 y 4:

- Calcular los valores de P_{FA}^{GT}, P_{D}^{GT} para cada valor de P_{FA}^{B} y guardarlos en los vectores respectivos.
- Representar en la misma figura 1 anterior la ROC del detector de GT.
- Represente en otra figura (figura 2) la probabilidad de detección alcanzada en el esquema GT vs el número de test realizados por individuo. A fin de poder comparar con el detector clásico, añada a esta misma figura 2 la curva de la probabilidad de detección alcanzada vs número de test realizados por individuo para el detector clásico.

[comandos Matlab `'.*', figure, hold on, plot`]

Operaciones finales

- Añada a la figura 1 y a la figura 2 la leyenda para identificar cada curva y el título de los ejes

[comandos Matlab `legend, xlabel,ylabel,grid on`]

4.2. Ejecute el programa con p =0.01, 0.1 y 0.2 Analice los resultados obtenidos en términos de:

- ROC del detector de GT en comparación con detector clásico.
- Número de test requeridos por el detector de GT en comparación con el detector clásico.

4.3. Se desea obtener un detector con una probabilidad de falsa alarma no superior a 0.005. en el caso de $p = 0.1$, ¿cuál es el valor del umbral γ que se debería emplear en el detector clásico y en el detector GT con $K = 2$?

[comandos Matlab `qfuncinv, sqrt`]

[Nota: sea un vector "`vector`" que tiene números *crecientes*. Para encontrar el valor de este vector que es más próximo a un número "`num`" y que al mismo tiempo no es mayor

que éste se puede hacer con la instrucción siguiente. El valor más próximo se guarda en "`valor`" y su posición en el vector se guarda en la variable "`indice`"

```
[temp]  = find(vector<=num);
indice = temp(end);
valor   = vector(indice);
```

4.4. Al final del fichero "GT_MonteCarlo.m" encontrará la función `realizacion_funcion_test`. Complete el código de esta función para que genere una realización del resultado de la función de test de acuerdo con el estado de la muestra a analizar (0 para 'no infectado' y 1 para 'infectado) y con la estadística del test básico descrita por $m_0, \sigma_0^2, m_1, \sigma_1^2$.

[comandos Matlab `function, if/else/end, randn, sqrt`]

4.5. Al principio del fichero "GT_MonteCarlo.m" encontrará el código Matlab para simular por MonteCarlo los detectores clásicos y GT y evaluar sus prestaciones con $p = 0.1$ y $K = 2$. Complete el software proporcionado para que el programa realice las siguientes tareas

- Genere una realización de los estados infectados/no infectados de una población de N individuos y los guarde en un vector de tamaño $1 \times N$ como 0 (no infectado) y 1 (infectado).

Detector clásico:

- Para cada individuo genere una realización de la función de test y tome una decisión empleando el valor del umbral hallado en el apartado **4.3.**
- Estime las prestaciones de este detector: evalúe las tasas experimentales de falsa alarma y detección y compárelas con los valores teóricos obtenidos en el apartado **4.2.**

Detector Group Testing:

- Agrupe los individuos en grupos de K=2. Simule la operación del detector de GT empleando el valor del umbral hallado en el apartado **4.3.**
- Estime las prestaciones de este detector: evalúe las tasas experimentales de falsa alarma y detección, así como el número de tests empleados**.**

Comparación final de resultados:

- Añada los valores obtenidos de las tasas experimentales de falsa alarma y detección y de número de tests a las figuras 1 y 2 generadas en el apartado **4.2**.

4.6. Compare los resultados obtenidos por MonteCarlo con los resultados teóricos.

5. Referencias

[1] R. Dorfman: "The detection of defective members of large populations". *Ann. Math.Statist*. (Diciembre de 1943), vol. 14(4): pp. 436–440.

[2] F. Huang, P. Guo y Y. Wang, "Optimal group testing strategy for the mass screening of SARS-CoV-2", *Omega*, vol. 112, Octubre 2022, 102689