Generación de clasificadores que predigan la necesidad de biopsia de cuello uterino en pacientes con sospecha de Cáncer Cervical

PF-5028 Mineria de datos

Paola Calderón
Carlos Solano

Agenda

- Origen de los datos
- Cancer Cervical
- Objetivos
- Metodología
- Demostración
- Q&A

Origen de los Datos

- Hospital Universitario de Caracas
- El conjunto de datos incluye historia médica de 858 mujeres. Algunos de los atributos no están disponibles por asuntos de privacidad.
- “Transfer Learning with Partial Observability Applied to Cervical Cancer Screening” - Iberian Conference on Pattern Recognition and Image Analysis. Springer International Publishing, 2017”
- Factores de riesgo de cáncer cervical que sugieran la necesidad de un examen de biopsia.

Cáncer Cervical

- 11,000 casos de cáncer cervical invasivo se reportan cada año en USA.
- Cada año mueren 300,000 mujeres alrededor del mundo.
- Las enfermedades de transmisión sexual han sido factores que aumentan el riesgo de cáncer en cualquier rango de edad.
- Factores como pobreza, múltiples parejas, historia familiar, gran cantidad de hijos y el fumado pueden ser riesgosos para el desarrollo del cáncer.

Objetivo General

- Generar clasificadores que predigan la necesidad de tomar una biopsia de cuello uterino en pacientes vulnerables a desarrollar cáncer de cuello uterino.

Objetivos Especificos

- Elegir los features predominantes y el número óptimo de features a partir del set de datos recopilado del Hospital Universitario de Caracas
- Evaluar diferentes algoritmos de clasificación para predecir si para una determinada paciente, dadas ciertas características, se requiere tomar una biopsia de cuello uterino.
- Implementar métodos de clasificación no supervisada para evaluar potenciales interacciones entre los features que hagan sospechar de la presencia de cáncer de cuello uterino.

Materiales y Métodos

- Pre-procesamiento de los Datos
- Decision Tree (SUP)
- MLP (ANN)
- ShuffleSplit (CV)
- kMeans (No-SUP)

Pre-procesamiento de Datos

- Análisis exploratorio de datos no relevantes.
- Uso de métodos para llenar valores faltantes de atributos con mayor importancia.
- Implementación de *ExtraTreesClassifier* y el método *feature_importances_* para identificar los valores con más impacto en el modelo.

Decision Tree

- El uso de reglas facilita la toma de decisiones para una clasificación binaria.
- Buen desempeño para múltiples escalas.
- Identificación de parámetros que ajusten mejor el modelo:
  - max_depth=3
  - min_samples_split=22

Multilayer perceptron

- Uso comparativo con respecto a los resultados del Decision Tree.
- Confirmación del comportamiento de los datos.
- Para implementaciones complejas, resolviendo los problemas de forma estocástica.

Cross Validation

- Shuffle-Split
- Scores de Cross-validation:
  - 0.97202797
  - 0.97435897
  - 0.95571096
  - 0.94172494
  - 0.96037296
  - 0.96270396
  - 0.93473193
  - 0.94871795
  - 0.97202797
  - 0.96270396

k-Means

- Identificación de comportamientos binarios en el conjunto de datos.
- Necesidad de analizar a profundidad los detalles de cada cluster identificado.
- Posible clasificación:
  - 0 -> no cancer
  - 1 -> cancer

Demostración

`Scikit-learn demo`

Conclusiones

- La representación de los clusters en el aprendizaje supervisado permite identificar los comportamientos existentes en el set de datos, con el fin de estudiarlos a más detalle e identificar a qué patología puede corresponder cada grupo.
- Se comprueba que el flujo de trabajo de un conjunto de datos va a depender del comportamiento que se obtenga de los diferentes clasificadores. No existe una sola forma de trabajar los datos, y el siguiente paso va a depender solamente del resultado del paso actual.