# Trabajo práctico 1: Bayes Ingenuo

Ph. D. Saúl Calderón Ramírez
Instituto Tecnológico de Costa Rica,
Escuela de Ingeniería en Computación, Programa de Ciencias de Datos,
PAttern Recongition and MAchine Learning Group (PARMA-Group)

29 de septiembre de 2022

**Fecha de entrega:** Domingo 16 de Octubre
**Entrega**: Un archivo .zip con el código fuente LaTeX o Lyx, el pdf, y un notebook Jupyter, debidamente documentado, con una función definida por ejercicio. A través del TEC-digital.
**Modo de trabajo**: Grupos de 3 personas.

**Resumen**

En el presente trabajo práctico se introduce la implementación de redes bayesianas.

## 1. Implementación de la clasificación multi-clase con Bayes ingenuo (100 puntos)

1. Para el presente ejercicio, se implementará la clasificación de digitos manualmente escritos en 10 clases (dígitos decimales del 0 al 9). La Figura 1 muestra algunas observaciones del conjunto de datos. El código provisto lee las imágenes del conjunto de datos, y los transforma a matrices binarias de $28 \times 28 = 784$ pixeles. El objetivo de su equipo de desarrollo es utilizar el teorema de Bayes para construir un modelo conocido como Bayes ingenuo, el cual permita estimar la clase a la que pertenece una nueva observación.

2. En el material del curso, se discute el algoritmo de Bayes ingenuo, el cual tiene por objetivo estimar la **probabilidad posterior** de que una observación (imagen en este caso) $\overrightarrow{m} \in \mathbb{N}^{784}$ pertenezca a una clase $k$ como:

$$p\left(t = k|\overrightarrow{m}_i\right)$$

Para aproximarla, se utiliza el teorema de Bayes, el cual luego de desarrollar y simplificar la expresion de tal probabilidad posterior, se concluye

Figura 1: Muestra del conjunto de datos a utilizar.

que esta es proporcional a la multiplicación de la probabilidad a priori de $p\left(t=k\right)$ y la verosimilitud de un pixel (blanco: 1 o negro: 0) $p\left(m_i|t=k\right)$:

$$p\left(t=k|\overrightarrow{m}\right) \propto \prod_{i=0}^{D} p\left(m_i|t=k\right) p\left(t=k\right).$$

La verosimilitud del pixel $p\left(m_i|t=k\right)$ se implementa como la probablidad de que $p\left(m_i=0|t=k\right)$ en caso de que ese pixel $i$ de la observación a evaluar en el modelo sea negro (0),en caso contrario $p\left(m_i=1|t=k\right)$.

*a*) **(15 puntos)** Implemente el cálculo de las probabilidades a priori $p\left(t\right)$ para las $K = 10$ clases en el conjunto de datos de entrenamiento. Realice tal calculo dentro de la funcion *train_model.*

*b*) Para evaluar la verosimilitud $p\left(m_i|t\right)$, es necesario estimar las densidades $p\left(m_i=0|t\right)$ y $p\left(m_i=1|t\right), p\left(m_i=2|t\right), \ldots, p\left(m_i=255|t\right)$ para todos los pixeles $i = 1, \ldots 768$ pixeles.

1) **(10 puntos extra)Enfoque basado en histogramas**: Para ello se le sugiere crear la matriz *p_m_1_given_k* con 768 filas (una por pixel) y 10 columnas (una por clase), por lo que entonces cada columna corresponde a la densidad de cada pixel. **Para realizar este calculo solo se le permite usar un ciclo *for,* con una iteracion por clase $k$, como maximo**.
2) **(15 puntos) Enfoque basado en un modelo paramétrico:** Su equipo decide probar un enfoque paramétrico donde se ajuste un modelo paramétrico como una función de densidad Gaussiana (en este caso con parámetros $\mu$ y $\sigma$). Implemente la estimación de $p\left(m_i|t\right)$ usando un modelo Gaussiano para los pixeles en cada clase.

*a'* Explique los cambios en cuanto a la representación propuesta en el enfoque basado en histogramas. Qué ventaja tiene el enfoque basado en el modelo paramétrico en cuanto la representación?

*c*) **(10 puntos)** Implemente los dos puntos anteriores en la función *train_model* y retorne una lista con las dos matrices (*[[p_m_0_given_k, p_m_1_given_k]]*), junto con el arreglo de probabilidades a priori para todas las clases.

*d*) **(10 puntos)** Implemente la función *test_model(input_torch, p_m_pix_val_given_k, p_t_tensor, num_classes = 10)* la cual realice la estimación de a cual clase pertenece una observación contenida en el vector *input_torch,* para un modelo representado en *p_m_pix_val_given_k, p_t_tensor (*obtenido en el paso anterior*).* Para ello, el enfoque de Bayes ingenuo estima la función de densidad posterior como sigue:

$$p\left(t=k|\overrightarrow{m}\right) \propto \prod_{i=0}^{D} p\left(m_i|t=k\right) p\left(t=k\right).$$

La clase estimada a la que pertenece la observación $\overrightarrow{m}$ corresponde entonces a la clase $k$ con mayor probabilidad posterior $p\left(t=k|\overrightarrow{m}\right)$

*e*) **(10 puntos)** Implemente la función *test_model_batch(test_set, labels, p_m_pix_val_given_k, p_t_tensor)* la cual calcule y retorne la tasa de aciertos para un conjunto de observaciones, basado en la función anteriormente implementada *test_model.*

## 2. Prueba del modelo

1. **(10 puntos)** Entrene el modelo con el conjunto de observaciones contenido en la carpeta train, y reporte la tasa de aciertos al utilizar la función anteriormente implementada *test_model_batch* (se espera que la tasa de aciertos sea mayor a 80 %). Verifique y comente los resultados. Es posible que observe valores nulos en el resultado de evaluar la función posterior a través de la función *test_model* la cual implementa la ecuación:

$$p\left(t=k|\overrightarrow{m}\right) \propto \prod_{i=0}^{D} p\left(m_i|t=k\right) p\left(t=k\right).$$

Si observa valores de 0 o nulos en la evaluación de la función, argumente el porqué puede deberse este comportamiento. ¿Cómo se puede corregir el problema detectado, según las herramientas matemáticas estudiadas en clase? Implemente tal enfoque y compruebe los resultados.

*a*) Compruebe lo anterior usando el enfoque basado en la binarizacion (func. densidad binomial), histogramas y el basado en el modelo Gaussiano. Comente y compare los resultados.

2. **(20 puntos)** Particione los datos de forma aleatoria con 70 % de las observaciones para entrenamiento y 30 % para prueba (a partir de la carpeta *train*). Calcule la tasa de aciertos para 10 corridas, cada una con una partición de entrenamiento y otra de prueba distintas. Reporte los resultados de las corridas en una tabla, además de la media y desviación estándar de la tasa de aciertos para las 10 corridas. Para realizar las particiones puede usar la libreria *sklearn.*

    *a*) Compruebe lo anterior usando el enfoque basado en histogramas, binarizado y el basado en el modelo Gaussiano. Comente y compare los resultados, usando los .