Problema toma de decisión por bayes.

Para esta asignación utilizarán el conjunto de datos encontrados en https://www.kaggle.com/sulianova/cardiovascular-disease-dataset.

Entregables

Como entrega se espera recibir:

1. Un documento "paper" escrito con el formato IEEE y con las partes esperadas para reporte de una investigación a saber: introducción con objetivos, métodos, resultados y análisis. Recuerden que tendrán al menos tres resultados de clasificación en el trabajo completo.
2. Anexos, que entendieran necesarios. Esto puede incluir códigos o cálculos a mano, pero en ningún caso justificaría que el documento principal se encuentre incompleto por hacer referencia a los anexos.
3. Formatos: el paper deberá ser entregado en pdf, los códigos deberán ser de Matlab y entregados en archivos scripts (.m) de Matlab.

Conceptos teóricos:

Esta asignación consiste en utilizar la teoría de decisión de bayes (Cap3 del libro) para la clasificación de pacientes cardíacos a partir del conocimiento de algunas de sus carácterísiticas. Además, se espera que apliquen los conceptos de los capítulos 4 y 5 para modelar las distribuciones de las diferentes clases usando la distribución normal.

Manejo de la data

Trabajarán con toda la data. Dejando una porción que consideren adecuada para training y el resto será dejada para validación. Recuerde que debe organizar aleatoriamente previamente los casos, pues no sabemos si tienen un orden específico en la lista original. Asumirán, además, que el dataset es una muestra representativa de la población general para los fines cardíacos y por tanto lo pueden utilizar para estimación de los Priors.

Asignación 1 (de 2): Aplicación de Bayes univariable con caráterísticas continuas.

Se quiere determinar qué personas tienen o no una enfermedad cardiovascular en base a solo conocer una sola característica. Para ello modelaremos la probabilidad de que tenga un problema cardiovascular usando distribución gausiana alrededor de la característica escogida. Para estos fines probaremos usando el peso, la estatura y la edad.

Procedimientos estadísticos

Esta es una guía del procedimiento estadístico que esperamos que realicen:

1. Establecer los priors $P(C_k)$. A partir de la data misma
2. Elegir las características x, que se usarán en el modelo (en este caso ya elegidas por el problema)
3. Modelar los likelihoods $p(x|C_k)$. Fíjense que esto es simplemente sacar los parámetros para la distribución de cada una de las clases. Y como están usando un feature a la vez pues lo repetirán 3 veces.
4. Modelar el evidence $p(x)$. Este paso no es necesario para clasificación. Sin embargo, dejamos la referencia aquí, pues sí es necesario en caso de que quieran modelar la probabilidad como tal o porque quieran calcular dicha probabilidad para estimar el riesgo asociado a la decisión.

5. Poner todo junto para formar la función discriminante. Aquí pueden pensar en usar directo la fórmula de Bayes: $P(C_k|x) = \frac{P(C_k)\, p(x|C_k)}{p(x)}$. Pero, esperamos que la usen de manera simplificada usando logaritmo natural (ver slide 10 de presentación del cap 4 y la teoría del libro).

Aunque se pueden utilizar códigos para realizar las operaciones básicas estadísticas que necesitarán, se espera que el trabajo contenga los planteamientos matemáticos y resultados (y no en forma de códigos). Ejemplo de esto sería que estén:

…las ecuaciones:

$$P(C_k|x) = \frac{P(C_k)\, p(x|C_k)}{p(x)}$$

….La explicación de los parámetros o elementos de la ecuación o de cómo se sacaron:

$P(C_k) = \frac{\sum xyj}{NN}$ es el promedio de la variable tal.

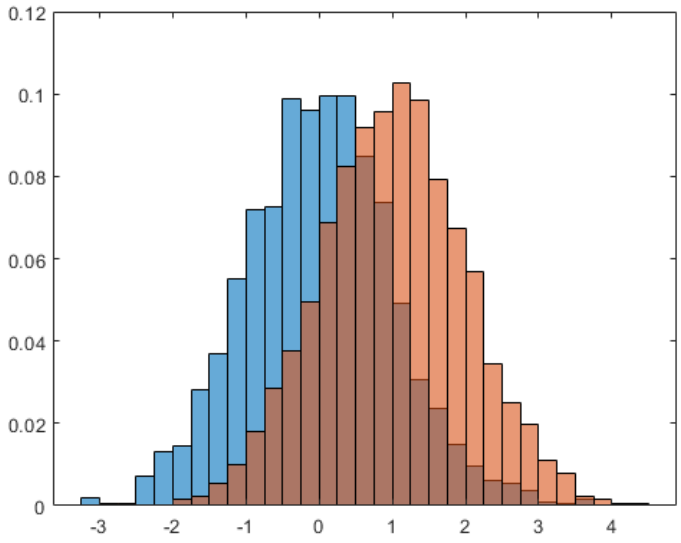
Y valores que se usan para dichos parámetros cuando aplique. Las cuales pueden estar dados en tablas o en el mismo texto. Como ejemplo:

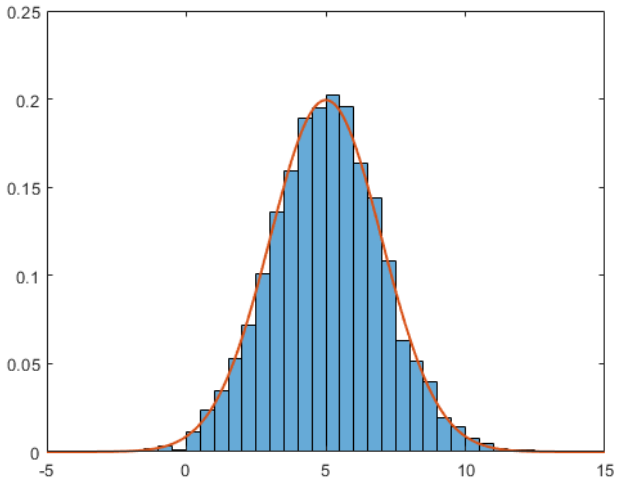
Clase K	$P(C_k)$
1	0.12
2	2.3
3	4.4

Valores de $P(C_k)$ para las diferentes clases

Gráficos esperados:

Para cada característica o feature usada, se espera que presenten el gráfico comparativo de los histogramas de las clases y la distribución normal estimada. Para los histogramas, pueden ver el help de Matlab\ comando histogram\ plot multiple histograms (https://la.mathworks.com/help/matlab/ref/matlab.graphics.chart.primitive.histogram.html#d123e574595) donde está el ejemplo de cómo hacer los siguientes dos gráficos que combinan lo que se quiere aquí:





Resultados:

Se espera que entre los resultados se presente el desempeño del algoritmo para lograr la clasificación. Deben, para esto explicar la métrica de error usada. Además del error general, se espera que se presenten las tablas de confusión (Pueden buscar: Confusion matrix) para cada uno de los features usados.

Análisis:

Debe incluir sus comentarios. Estos deben ser críticos, comparando los diferentes métodos, features, etc usados. Además, deberán mirar factores de complejidad de los algoritmos y efectividad (esto es: tiempo de desarrollo, tiempo de entrenamiento, tiempo de ejecución, efectividad de aciertos, etc).

Asignación 2: Bayes multivariable

Re aplicarán lo anterior pero esta vez usando los diferentes features a la vez para la clasificación. Pero igual tendrán tres opciones que explorar:

Usando el peso y altura como features.

En este caso queremos deberán modelar las distribuciones para cada clase usando la distribución normal multivariable. Ver (presentación 5, slide 7). Pero que aplicarán usando la forma de discriminante simplificado por el uso de ln( ) presentada en el slide14 “Different S_i”.

Se espera que aparte de la descripción del procedimiento, estén los valores resultantes para las matrices de covarianzas, los vectores de medias y priors.

En cuánto a gráficos, se espera que grafiquen las distribuciones para ambas clases en forma 3 dimensional (si es posible en un mismo gráfico) y así ustedes puedan tener mejor opinión acerca del problema. Les recomendamos ver: Comandos como Surf, meshgrid, hist3 les serán de importancia aquí. Pueden entrar en Distribución normal multivariante

Resultados y análisis: se entieden igual a la sección de una única variable.

Usando el peso, altura y edad como features independientes (Naive Bayes)

En este caso harán la clasificación usando Naive bayes y las tres variables. Para esto pueden referirse al capítulo 5, slide11 y 18 sobre “independent inputs…”

Nota: Para los casos de más de dos variables no aplican los gráficos de distribución. Pero si deben incluir el resto de información.

Usando el peso, altura y edad como features correlacionados.

Finalmente usarán las tres variables usando la distribución completa (como mismo se hizo en el punto de dos features).