PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA
DEPARTAMENTO DE CIENCIA DE LA COMPUTACIÓN
IIC2613 INTELIGENCIA ARTIFICIAL

12, Noviembre 2015

Interrogación 3

Tiempo: 90 minutos, SIN APUNTES

Nombre:_____

1. (*12 puntos*) **Comente o responda las siguientes afirmaciones o consultas, fundamentando brevemente cada respuesta.**

a. Se tiene una red de Bayes que relaciona 3 variables binarias A, B y C; tal que A es padre de B, y B es padre de C. ¿Cuántos párametros es necesario estimar para especificar la probabilidad conjunta de A, B y C?.

Solución: *1 parámetro para $p(A)$, 2 parámetros para $p(B|A)$, y 2 parámetros para $p(C|B)$. Total 5 parámetros.*

b. En el entrenamiento de un árbol de decisión, ¿Cuál es el mayor riesgo de NO agregar un post proceso de poda?.

Solución: *Sobreajuste.*

c. En el entrenamiento de un árbol de decisión, indique 1 tarea donde un set de validación es relevante.

Solución: *Podar el árbol.*

d. En el entrenamiento de un árbol de decisión, indique 1 tarea donde un set de test es relevante.

Solución: *Medir la capacidad de generalización del árbol, i.e., su rendimiento esperado para la clasificación de nuevas instancias.*

e. En el entrenamiento de un árbol de decisión, aumentar el tamaño del set de entrenamiento siempre redundará en un mejor modelo.

Solución: *No necesariamente, en primer lugar se requiere garantizar que los nuevos datos sean representativos del problema y no una muestra sesgada. Aún si este es el caso, puede ser que la herramienta de aprendizaje no sea la adecuada y nuevos datos no mejoren el rendimiento del modelo.*

f. En el entrenamiento de un árbol de decisión, siempre existe un atributo (dimensión) que está presente en todas las reglas de clasificación representadas por el árbol.

Solución: *El atributo ubicado en la raíz del árbol aparece en todas las reglas. De hecho, ésta es una fuente de inestabilidad en el uso de árboles decisión, dada su gran sensibilidad al correcto valor de este atributo.*

g. En general, en una red neuronal una estructura con más unidades (neuronas) en la capa oculta garantiza una convergencia a una solución óptima en un menor número de iteraciones (épocas).

Solución: *No necesariamente, de hecho al tener más neuronas el espacio de hipótesis es más complejo y por ende más difícil de explorar.*

h. Para una red neuronal, en situaciones de sobreajuste, el error en el set de validación será mayor que en el set de entrenamiento.

Solución: *Correcto, este hecho es el que justifica el uso de un set de validación para detectar situaciones de sobreajuste en una red neuronal.*

i. En general, en un problema de clasificación mientras más atributos (dimensiones) tenga el set de entrenamiento, mayor es la probabilidad de experimentar problemas de sobreajuste.

Solución: *No necesariamente, de hecho en muchos casos un mayor número de atributos implica un problema más complejo, por tanto más difícil de modelar y caer en situaciones de sobreajuste.*

j. Un algoritmo de aprendizaje de máquina puede ser útil en una situación en que se requiera predecir que clientes pueden dejar cierta compañía telefónica en los próximos meses.

Solución: *Correcto, los algoritmos de aprendizaje de máquina tienen utilidad en problemas de predicción.*

k. Un algoritmo de aprendizaje de máquina puede ser útil en una situación en que se requiera actualizar automáticamente la lista de operarios del departamento de ventas de cierta compañía.

Solución: *Falso, éste es un problema determinístico, donde no hay necesidad de predicciones, sino una consulta directa a la base de datos correspondiente.*

l. Un algoritmo de aprendizaje de máquina puede ser útil en una situación en que se requiera determinar el monto total de ventas del año anterior de cierta compañía.

Solución: *Falso, éste es un problema determinístico, donde no hay necesidad de predicciones, sino una consulta directa a la base de datos correspondiente.*

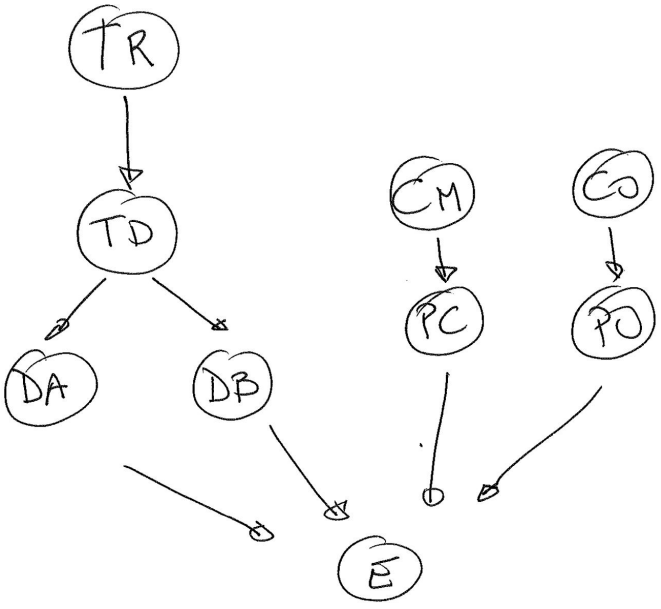
2. (*16 puntos*) Red de Bayes y Árbol de Decisión

a. (*8 pts*) En el contexto de una elección política entre 2 candidatos para las regiones A y B. En términos de uno de los candidatos se tienen las siguientes variables:

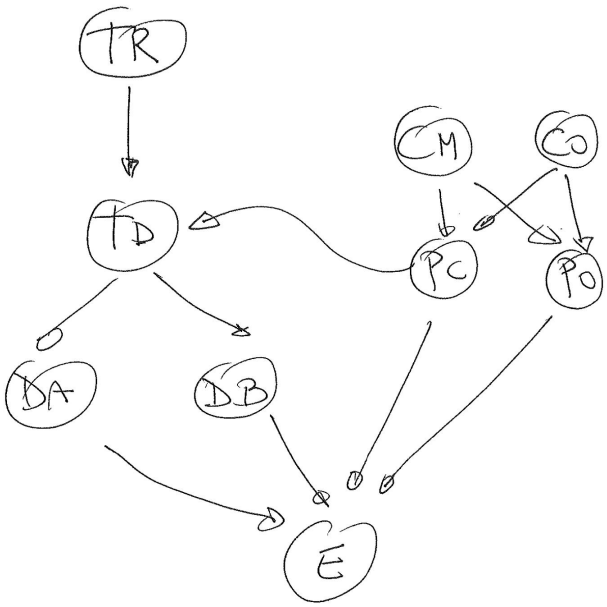
- TR: Tiempo dedicado a recolectar dineros para campaña.
- TD: Total de dinero recolectado para campaña.
- DA: Cantidad de dinero destinada a propaganda política en región A.
- DB: Cantidad de dinero destinada a propaganda política en regón B.
- CM: calidad del mensaje y carisma del candidato.
- PC: popularidad del candidato en su rol como político.
- CO: calidad del mensaje y carisma del oponente.
- PO: popularidad del oponente.
- E: Candidato es electo.

Dibuje una red de Bayes que modelo las relaciones entre estas variables e indique la factorización subyacente a esta red.

Solución: *Ver figuras.*



Algunos alumnos pueden haber considerado que el carisma del oponente (CO) puede influenciar la popularidad del candidato (PC), y que el carisma del candidato (CM) la popularidad del oponente (PO). Adicionalmente, algunos pueden haber considerado que la popularidad del candidato (PC) puede influenciar el total de dinero recaudado (TD). La siguiente figura agrega estas relaciones. Para efectos de la correción, ambas figuras se consideran como respuestas correctas.



b. (*8 pts*) Se desea construir un árbol de decisión para saber si Chile será capaz de ganar a Colombia esta noche en fútbol. La decisión estará basada en dos atributos binarios:

- A = Llueve.
- B = El estadio esta lleno.

Asuma que históricamente Chile gana el 70% del total de partidos que juega como local (el restante 30% lo empata o pierde). Además, Chile juega 20% de los partidos de local con lluvia y de ellos gana el 80%, mientras que cuando no llueve gana el 85%. Por otro lado, Chile juega de local 60% de sus partidos con estadio lleno y de ellos gana el 85%, mientras que cuando no hay estadio lleno gana sólo el 45% de las veces.

- a) Calcule la entropía de la variable binaria Chile ganará el partido esta noche.

  Solución: $H(ChileGana) = -\{0.7log_2(0.7) + 0.3log_2(0.3)\}$

- b) Calcule la ganancia de información del atributo A.

  Solución:

$$\begin{aligned} G(lluvia) &= -P(lluvia) * \{0.8log_2(0.8) + 0.2log_2(0.2)\} - P(noLluvia) * \{0.85log_2(0.85) + 0.15log_2(0.15)\} \\ G(lluvia) &= -0.2 * \{0.8log_2(0.8) + 0.2log_2(0.2)\} - 0.8 * \{0.85log_2(0.85) + 0.15log_2(0.15)\} \end{aligned}$$

  Como se comentó durante la interrogación, para efectos de este problema se ignora que algunas probabilidades marginales pueden potencialmente no sumar 1.

- c) Calcule la ganancia de información del atributo B.

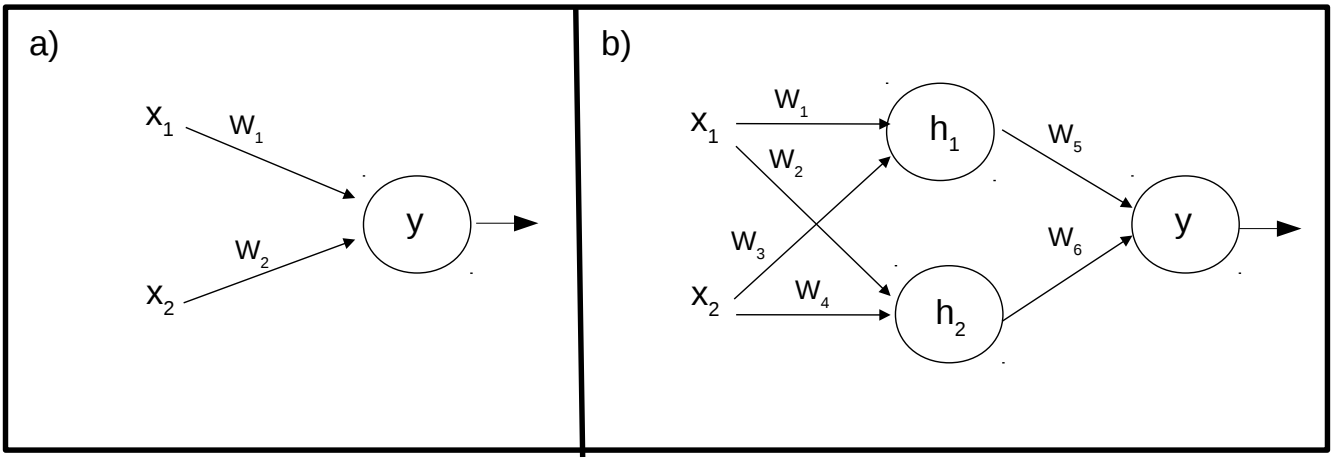
  Solución:

$$\begin{aligned} G(estadioLleno) &= -P(estadioLleno) * \{0.85log_2(0.85) + 0.15log_2(0.15)\} - P(estadioNoLleno) * \{0.45log_2(0.45) + 0.65log_2(0.65)\} \\ G(estadioLleno) &= -0.6 * \{0.85log_2(0.85) + 0.15log_2(0.15)\} - 0.4 * \{0.45log_2(0.45) + 0.55log_2(0.55)\} \end{aligned}$$

3. (*16 puntos*) **Redes Neuronales**

a. (*8 pts*) Las redes de la figura consisten de unidades del tipo perceptron lineal, es decir, la salida de cada unidad no incluye la función de activación no lineal. Matemáticamente, la salida y de cada unidad está dada por $y = \sum_i w_i x_i$, donde w_i y x_i corresponden respectivamente a los pesos y entradas que alimentan dicha unidad. Lo mismo para las unidades indicadas como h1 y h2, que también corresponden a perceptrones lineales.

Demostrar que bajo este tipo de unidad, las 2 redes de la figura tienen el mismo espacio de hipótesis.



Solución: *Todas las neuronas son lineales, por tanto, el espacio de hipótesis de ambas redes queda dado por combinaciones lineales de las entradas. Dado que una combinación lineal de funciones lineales, sigue siendo una función lineal, las redes son equivalentes. Aca una demostración para el caso de las redes de la figura:*
Para la red 1 tenemos:

$$y = \hat{w_1} x_1 + \hat{w_2} x_2$$

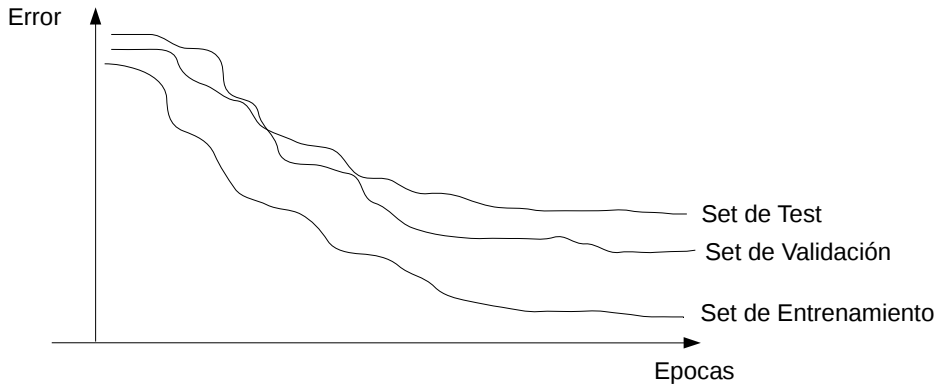
Para la red 2 tenemos:

$$y = w_5(x_1 w_1 + x_2 w_3) + w_6(x_1 w_2 + x_2 w_4)$$
$$y = x_1(w_5 w_1 + w_6 w_2) + x_2(w_5 w_3 + w_6 w_4)$$

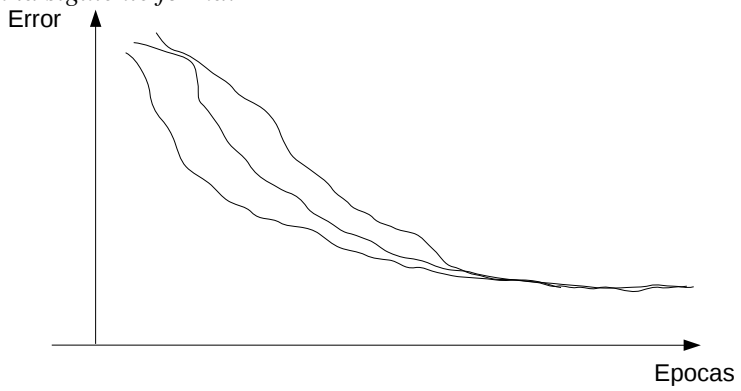
Por tanto, basta igualar los parámetros para demostrar que las redes son equivalentes, i.e.:

$$\hat{w_1} = w_5 w_1 + w_6 w_2$$
$$\hat{w_2} = w_5 w_3 + w_6 w_4$$

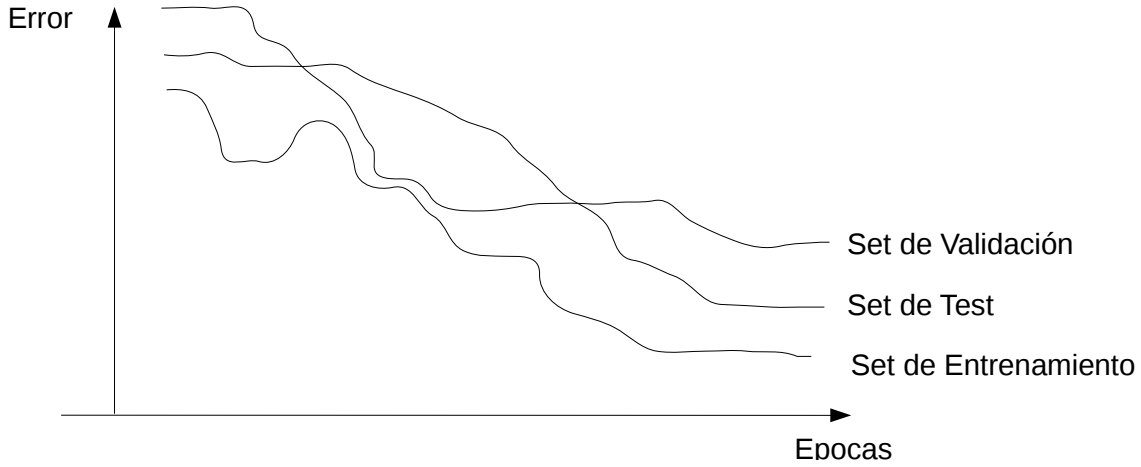
b. (*4 pts*) Las siguientes curvas muestran la evolución de la función de error en los sets de entrenamiento, validación y test, al entrenar cierta red usando el algoritmo de backpropagation. Considerando hipotéticamente, que es posible repetir el proceso de entrenamiento de esta red, pero usando un conjunto ilimitado de datos y épocas para entrenar, ¿Cómo sería la forma esperada de estas 3 curvas?. Utilice la figura en blanco para dibujar la forma aproximada de ellas.



Solución: *Al contar con datos infinitos de entrenamiento, no existe sobreajuste. Adicionalmente, como se cuenta con infinitas épocas de entrenamiento, los errores de los 3 sets convergen a un mismo valor. Por tanto, las curvas tienen la siguiente forma:*



c. (*4 pts*) Las siguientes curvas muestran la evolución de la función de error en sets de entrenamiento, validación y test, al entrenar cierta red usando el algoritmo de backpropagation. Alguna(s) de las curva(s) merece ciertas dudas respecto a si fue obtenida correctamente?, Justifique su respuesta.



Solución: *Hay 2 características de estas curvas que llaman la atención:*

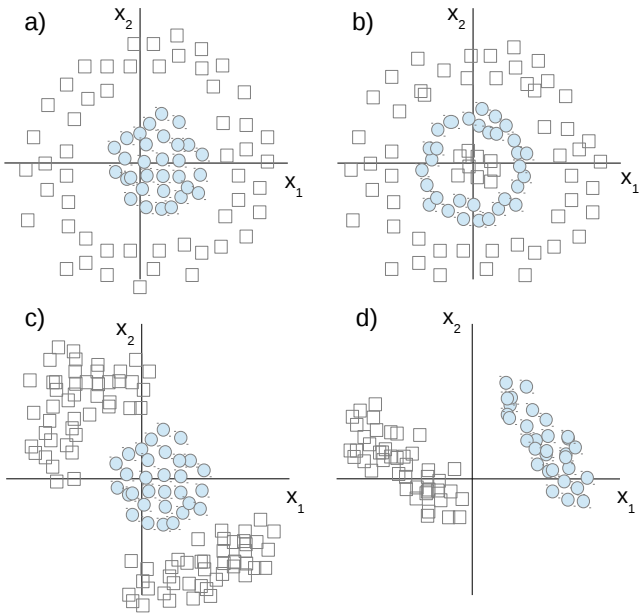
- La curva correspondiente al set de entrenamiento tiene subidas muy pronunciadas, lo cual indica que probablemente hay un error en el método de descenso de gradiente, que debiera siempre ir disminuyendo el error. En algunas ocasiones excepcionales, en que se utiliza método incremental es posible que la curva de error aumente no sea estrictamente descreciente, pero cualquier aumento debiera ser muy leve.

- Llama la atención que las curvas de error en set de validación y test sean tan dispares. Dado que ambos conjuntos corresponden a datos no vistos por la red, uno esperaría que fueran más similares.

Ambas situaciones indican potenciales problemas en la implementación de la red.

4. (*16 puntos*) Naive Bayes

Las siguientes figuras muestran los set de entrenamiento para diferentes problemas de clasificación binaria. Como se muestra en las figuras, cada registro (dato) de entrenamiento consiste de 2 atributos x_1 e x_2, y un rótulo de clase binario indicado con los símbolos □ y ⊙, respectivamente. En cada uno de los set de entrenamiento, el número de registros en cada clase es el mismo. Se decide solucionar los problemas utilizando clasificadores del tipo Naive Bayes, donde las funciones de probabilidad de los atributos dado la clase serán modeladas por funciones Gaussianas, y la de la clase con función multinomial.



a. (*4 pts*) Escriba las expresiones genéricas que usará el algoritmo de Naive Bayes para realizar la clasificación, es decir, las expresiones con las cuales el algoritmo de Naive Bayes aproximará: $\underset{c_i \in \{\square, \odot\}}{\text{argmax}} \; P(c_i|x_1, x_2)$.

obs: Para referirse a una función gausiana sobre una variable x basta con indicar $G(x)$ o $P_G(x)$, no es necesario escribir explícitamente la ecuación de dicha distribución.

Solución: -

$$\underset{c_i \in \{\square, \odot\}}{\text{argmax}} \; P(c_i|x_1, x_2) \propto P(x_1|c_i)P(x_2|c_i)P(c_i) \tag{1}$$

b. (*4 pts*) ¿Algún(os) término(s) de las expresiones anteriores resulta(n) ser no relevante(s) para la decisión final? Justifique.

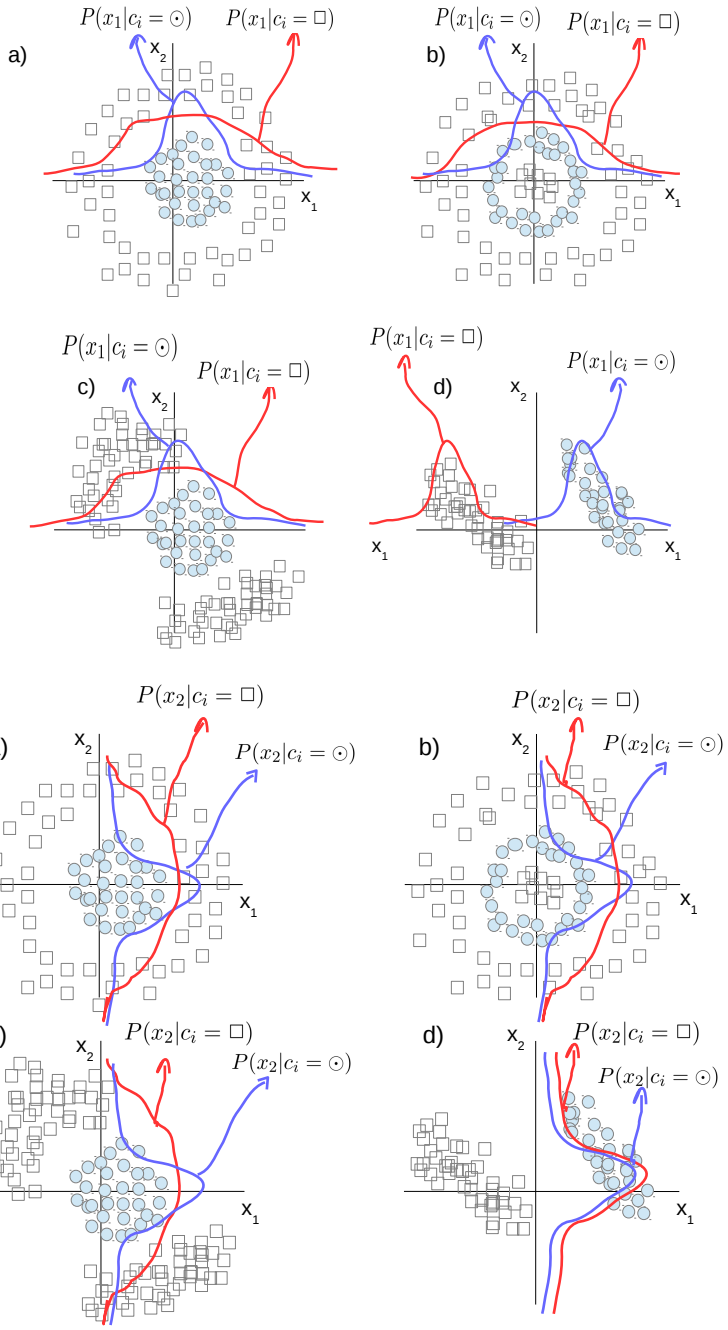
Solución: *Dado que el número de registros de cada clase es el mismo, se tiene que:*

$$P(c_i = \square) = P(c_i = \odot) = 0.5 \tag{2}$$

Por tanto, este término no afecta la decisión entre las clase □ y ⊙.

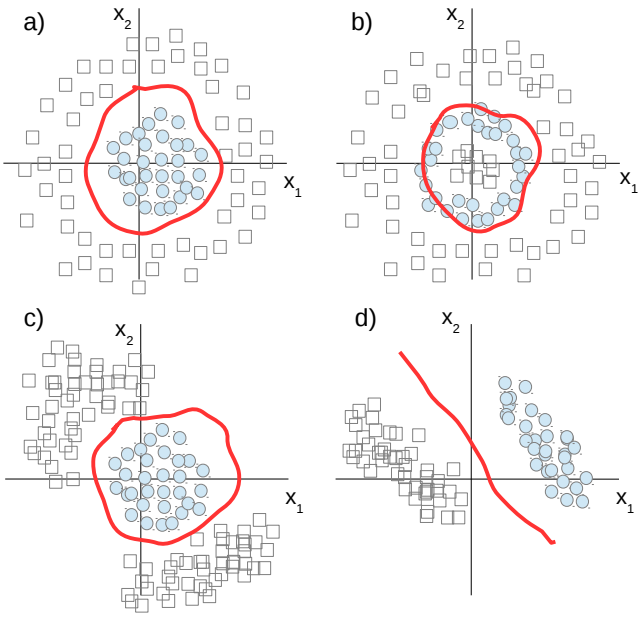
c. (*4 pts*) Usando la figura indicada arriba, dibuje aproximadamente para cada caso la posición de cada una de las funciones gausianas relevantes al problema.

Solución: -



d. (*4 pts*) Usando la figura indicada a continuación, dibuje aproximadamente para cada caso la superficie de decisión utilizada por el clasificador de Naive Bayes. ¿En que casos (a-d) el supuesto gausiano lleva a obtener un buen clasificador?

Solución: *Para cada problema de clasificación, la superficie de decisión está dada por la clase que maximiza el producto de los 2 términos relacionados con los atributos, i.e., el producto* $p(x_1|c_i)p(x_2|c_i)$. *Analizando los gráficos en la pregunta anterior, se obtienen las superficies de decisión de la siguiente figura. El mejor resultado es para el caso d), donde claramente el clasificador es capaz de separar las clases correctamente y con un alto margen. Esto esta dado por la buena separación aportada por el atribuo* x_1. *Los casos a) y c) son adecuados aunque, dependiendo de la ubicación precisa de los registros, el margen de separación entre las clases puede ser bastante ajustado. El peor caso es b), donde la superficie de decisión no es apropiada, especialmente para los registros de la clase* $\square$ *ubicados en el centro del sistema de coordendas.*


a)
x2
x1
b)
x2
x1
c)
x2
x1
d)
x2
x1