**Ejercicios Aprendizaje Automático**

**Ejercicio.**

**Explica la diferencia entre la aproximación tradicional y la aproximación del aprendizaje automático para resolver problemas.**

La diferencia es que por una parte la aproximación tradicional se programa a al programa para que resuelva ciertas tareas, por otra parte la aproximación pro aprendizaje automatic se programa sin saber todos las tareas a la que se enfrentara, de forma que pueda resolver tareas para las que no habia sido programada inicialmente sin la necesidad de reprogramarlo.

**Ejercicio.**

**En aprendizaje automático, ¿qué es la generalización?**

La generalización es la capacidad de abtraer un problema para sacar las caracteristicas communes con otras tareas que ya se Saben resoler para poder resolverlas de manera correcta.

**Ejercicio.**

**Explica la diferencia entre aprendizaje supervisado y aprendizaje no supervisado.**

La diferencia es que el aprendizaje supervisado es el que se hace por medio de un entrenamiento teniendo un “professor” que nos da un conjunto de ejemplos y sus caracteristicas para entrenarnos, el no supervidado se nos dan las caracteristicas sin clasificarnoslas y la maquina generaliza sola y une los grupos, sin que nadie le haya dicho nada.

**Ejercicio.**

**Queremos desarrollar una aplicación capaz de distinguir imágenes de perros y de gatos. Para construir dicha aplicación construimos un banco de datos que solo contiene imágenes de perros de razas grandes y gatos de razas pequeñas ¿estamos construyendo de manera correcta nuestro banco de datos? ¿por qué?**

**Ejercicio.**

**Para desarrollar la aplicación de perros y gatos descrita en el apartado anterior pensamos inicialmente en utilizar los siguientes descriptores para cada instancia de nuestro dataset: número de patas, número de ojos, ¿es mamífero?, ¿vuela?, y color de pelo. ¿Hemos hecho una buena elección de descriptores? ¿Por qué?**

**Ejercicio.**

**¿Por qué es habitual trabajar con descriptores numéricos?**

**Ejercicio.**

**¿Qué son los embeddings?**

**Ejercicio.**

**Disponemos de dos datasets de viviendas, en el primero de ellos cada vivienda se ha descrito mediante los siguientes descriptores: nº de habitaciones, nº de baños, localización geográfica y precio; en el segundo de los datasets, cada vivienda es descrita mediante nº de habitaciones, metros cuadrados, y años desde su construcción. ¿Podríamos juntar ambos datasets para tener un único dataset más grande?**

**Ejercicio.**

**Disponemos de un dataset de imágenes de bicicletas, motos y coches; y suponed que podemos obtener cualquier propiedad de las imágenes (e.g. color del vehículo, si tiene tubo de escape, etc). Dar tres descriptores que resultarían útiles en este dataset.**

**Ejercicio.**

**Dar ejemplos de descriptores binarios, numéricos, cadenas de caracteres y categóricos para describir a una persona.**

**Ejercicio.**

**Dadas los siguientes vectores de descriptores binarios A=(1,0,0,1,0) y B=(0,1,1,1,0) calcular los coeficientes de similaridad: simple-matching, coeficiente de Dice y coeficiente de Jaccard.**

**Ejercicio.**

**Dados los puntos (1,2,3) y (-2,4,5), calcular la distancia Euclídea y Manhattan entre ellos.**

**Ejercicio.**

**Dados los vectores de descriptores A=(verde,azul,amarillo,rojo) y B=(verde,morado,lila,rojo), calcular la similaridad entre ellos utilizando la medida de simple-matching.**

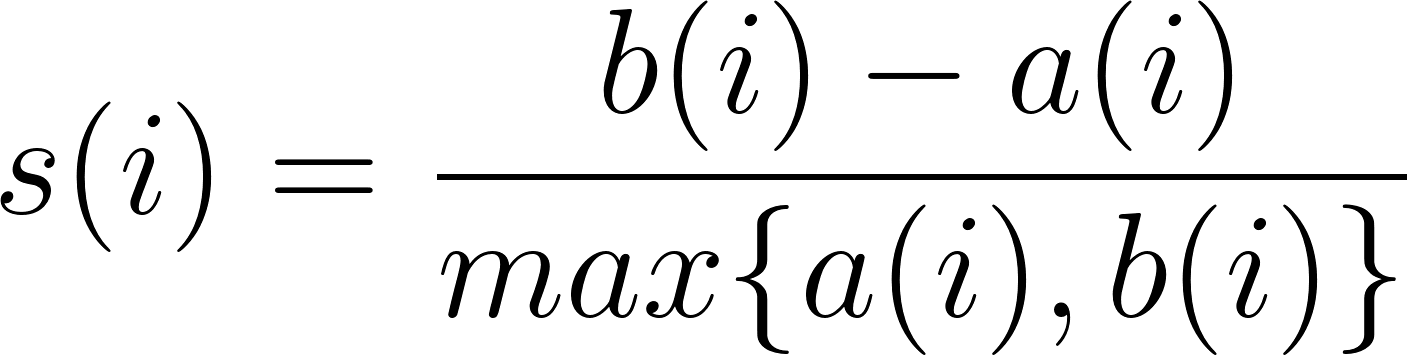
**Ejercicio.**

**Considerar el siguiente dataset: (2,2), (0,5), (3,3), (-1,6), (4,7). Aplicar el algoritmo K-means utilizando como centroides iniciales (0,0) y (4,0); en lugar de iterar el proceso hasta converger, repetir el proceso descrito en K-means 4 veces.**

**Ejercicio.**

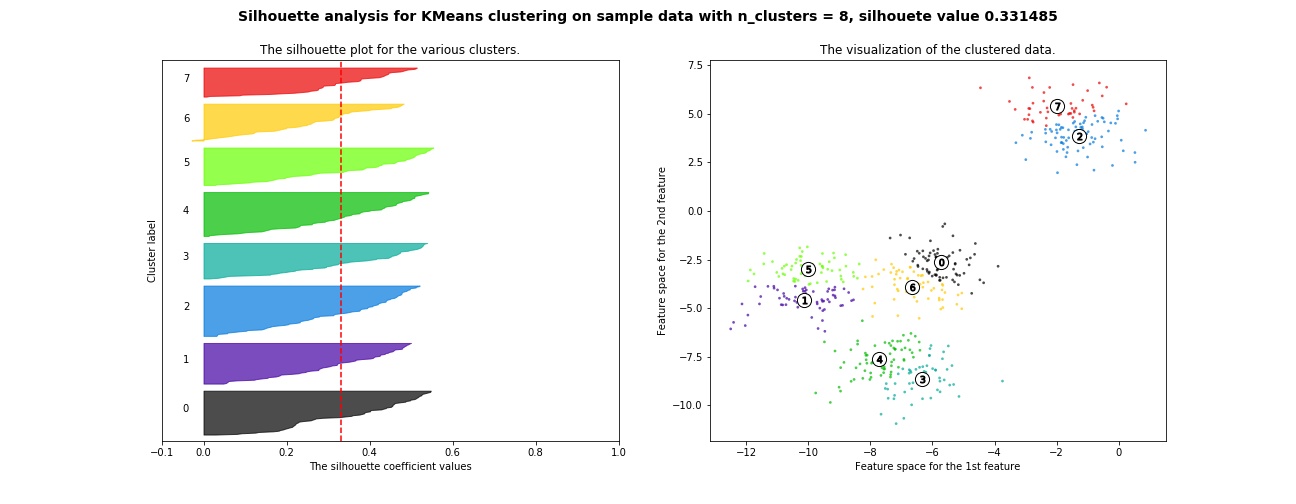
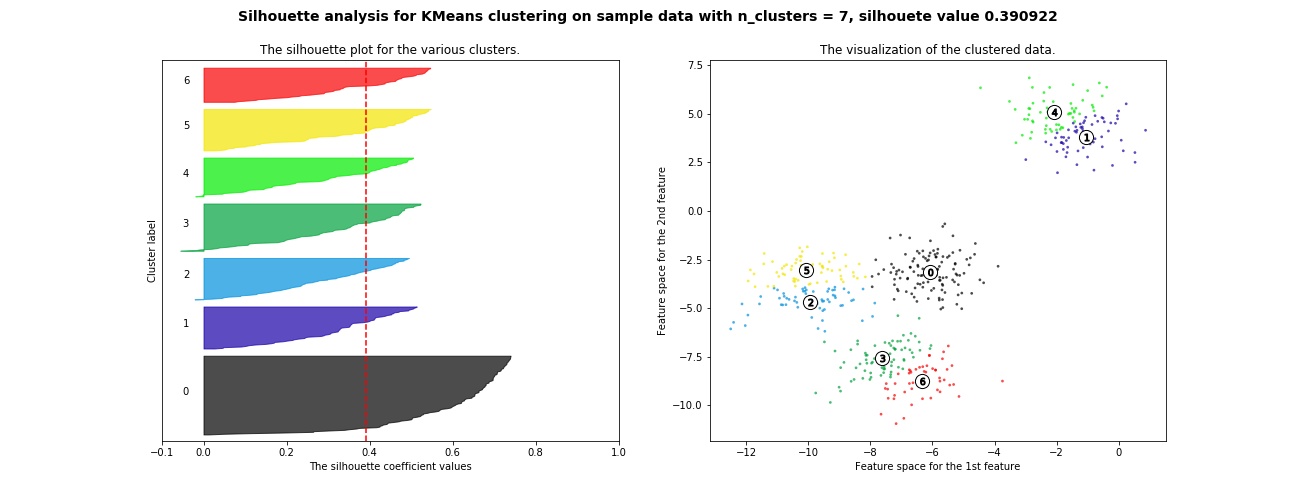
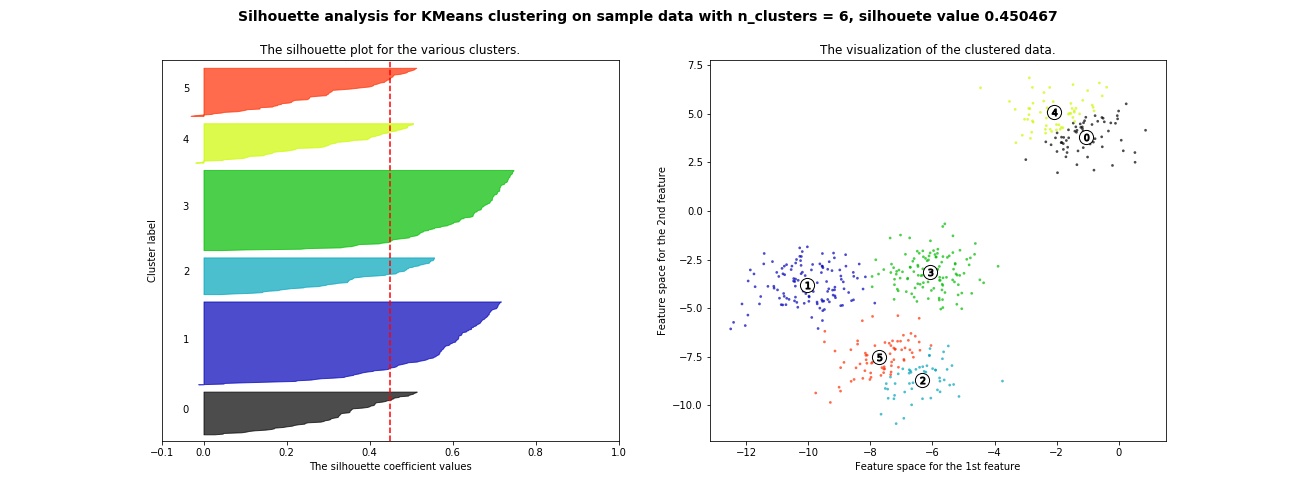
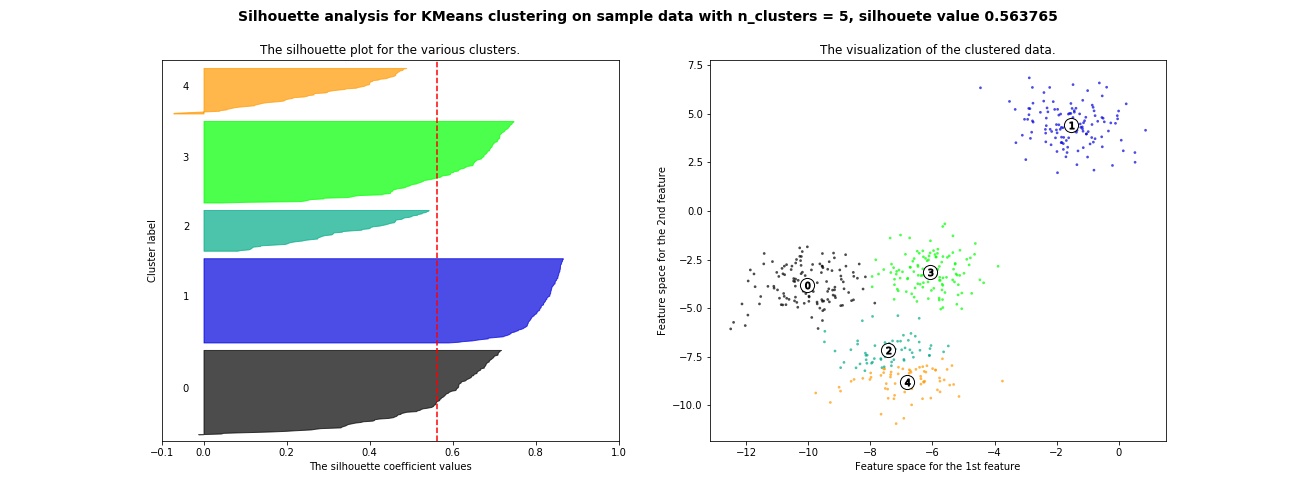
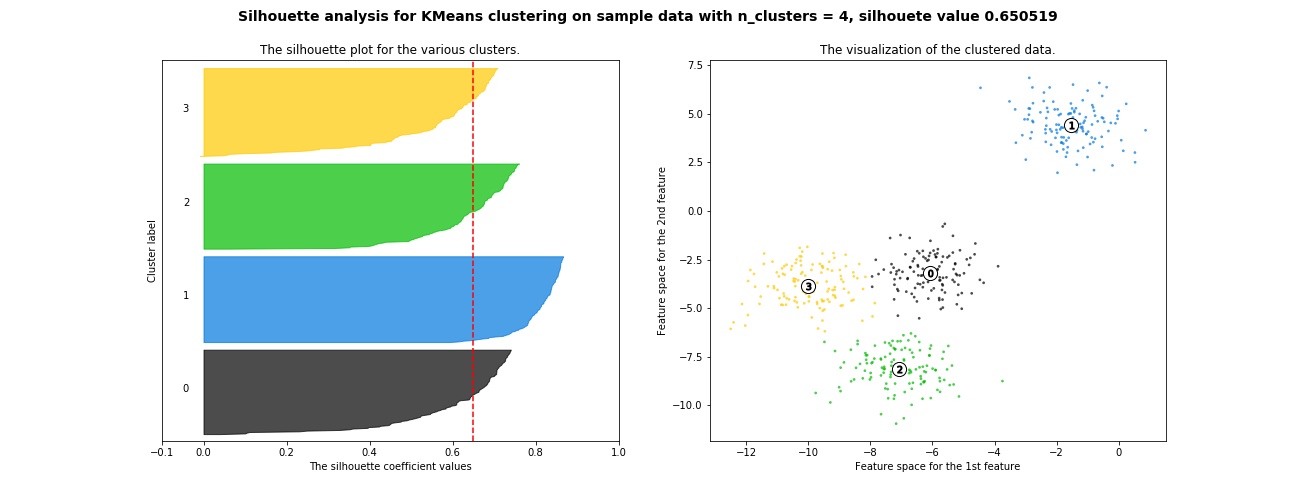
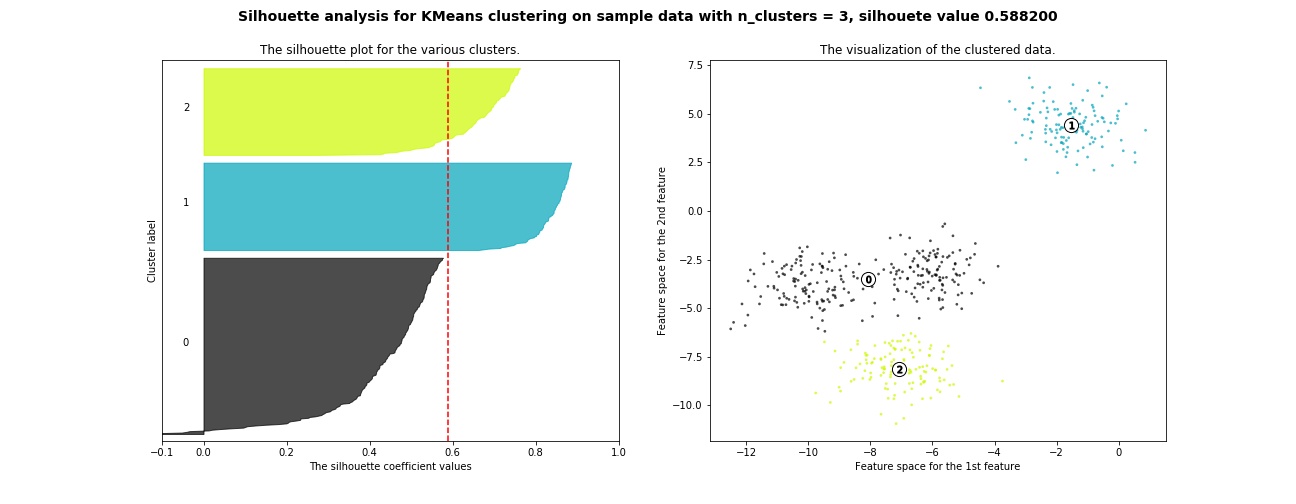
**¿Cuál es el mayor inconveniente del algoritmo K-means? ¿Qué técnicas podemos aplicar para abordarlo? ¿En qué consisten dichas técnicas?**

**Ejercicio.**

**En la fórmula para calcular el valor de la silueta (**[****](https://www.codecogs.com/eqnedit.php?latex=s(i)%20%3D%20%5Cfrac%7Bb(i)-a(i)%7D%7Bmax%5C%7Ba(i)%2Cb(i)%5C%7D%7D%250)**), ¿que miden a(i) y b(i)?**

**Ejercicio.**

**Ante los siguientes diagramas de siluetas ¿qué número de clústeres recomendarías para el algoritmo de k-means?**

****

**Ejercicio.**

**Explica que mide la función de coste en el algoritmo de K-means.**

**Ejercicio.**

**¿Qué técnica se utiliza para evitar caer en mínimos locales en el algoritmo de K-means?**

**Ejercicio.**

**¿Qué ventajas tienen los algoritmos de clustering DBSCAN y Mean-shift con respecto al algoritmo k-means? ¿Qué problemas tienen estos algoritmos?**

**Ejercicio.**

**Explica los pasos de los algoritmos de clustering DBSCAN, Mean-shift y K-means.**

**Ejercicio.**

**Considerando los siguientes datos y utilizando la distancia Euclídea, construir los clústeres que se producen respectivamente al utilizar clustering jerárquico con enlace completo y con enlace de la media. Construye también los dendrogramas que se generan a partir de ellos.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **4 piernas** | **Huevo amniótico** | **Pelo** | **Desarrollo en placenta** | **Gran cerebro** |
| **Rana** | **1** | **0** | **0** | **0** | **0** |
| **Tortuga** | **1** | **1** | **0** | **0** | **0** |
| **Canguro** | **1** | **1** | **1** | **0** | **0** |
| **Ratón** | **1** | **1** | **1** | **1** | **0** |
| **Humano** | **1** | **1** | **1** | **1** | **1** |

**Ejercicio.**

**Explica qué es un dendrograma y para qué se utiliza.**

**Ejercicio.**

**¿Cuáles son las razones para reducir la dimensionalidad de nuestros datos?**

**Ejercicio.**

**¿A qué nos referimos cuando hablamos de la maldición de la dimensionalidad?**

**Ejercicio.**

**Explica la diferencia entre la selección de descriptores y la extracción de descriptores como métodos para reducir la dimensionalidad de los datos.**

**Ejercicio.**

**Explica la diferencia entre la selección secuencial de descriptores hacia adelante y la selección secuencial de descriptores hacia atrás.**

**Ejercicio.**

**¿En qué nos basamos para elegir el número de componentes principales en el algoritmo de PCA?**

**Ejercicio.**

**¿Por qué decimos que el método t-SNE tiene una componente estocástica?**

**Ejercicio.**

**Suponiendo que los datos no son linealmente separables, ¿qué técnicas de reducción de la dimensionalidad sería conveniente usar?**

**Ejercicio.**

**Si queremos desarrollar un modelo de clasificación para predecir si un email es “spam” o “no es spam”. ¿Cuál de las siguientes afirmaciones es cierta?**

* **Las palabras en el asunto serán buenas etiquetas.**
* **Usaremos ejemplos no etiquetados para entrenar el modelo.**
* **Algunas etiquetas pueden ser engañosas.**
* **Los emails que no son marcados como “spam” o “no spam” son ejemplos no etiquetados.**

**Ejercicio.**

**En los problemas de aprendizaje supervisado, ¿por qué es importante que todas las categorías cuenten con un número uniforme de instancias?**

**Ejercicio.**

**Explica en qué tres conjuntos se suele partir un dataset en el aprendizaje supervisado y para qué se utiliza cada uno de esos conjuntos.**

**Ejercicio.**

**Al partir un dataset en conjunto de entrenamiento y de test, ¿por qué es importante que estos conjuntos sean independientes?**

**Ejercicio.**

**¿Por qué no podemos utilizar el conjunto de test para optimizar los hiperparámetros de los modelos de aprendizaje supervisado? ¿Cuál es el modo correcto de optimizar dichos hiperparámetros?**

**Ejercicio.**

**¿Cuáles son los hiperparámetros del algoritmo Knn?**

**Ejercicio.**

**¿Por qué el algoritmo Knn es tan rápido de entrenar?**

**Ejercicio.**

**Dado un dataset con 10 instancias, 8 de la clase A y 2 de la clase B y trabajando con el algoritmo de Knn, ¿sería razonable utilizar un k con valor 3? ¿y con valor 5?**

**Ejercicio.**

**En el algoritmo de Knn, ¿por qué no usamos valores pares para k?**

**Ejercicio.**

**En el algoritmo de Knn, ¿qué representa el valor de k?**

**Ejercicio.**

**¿Por qué no es adecuado utilizar el algoritmo de Knn cuando la dimensión de los vectores de descriptores crece?**

**Ejercicio.**

**Dadas las siguientes instancias (separadas por punto y coma):**

* **Clase A: (1,5); (2,4); (3,2); (4,3)**
* **Clase B: (0,3); (-2,3); (1,2); (-1,1)**

**Clasificar con el algoritmo Knn usando la distancia Euclídea y valor de k=3 las instancias (2,2); (0,0) y (1,3).**

**Ejercicio.**

**Explica cuál es la idea intuitiva de los árboles de decisión.**

**Ejercicio.**

**¿Por qué se dice que los árboles de decisión son un algoritmo de caja blanca?**

**Ejercicio.**

**Considerando el siguiente árbol de decisión**

****

**Completa la siguiente tabla donde la columna Examen indica la decisión que se toma basándonos en el árbol de decisión anterior:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Aprovechamiento** | **Asiste** | **Examen** |
| **Instancia 1** | **Deficiente** | **Sí** |  |
| **Instancia 2** | **Bueno** | **No** |  |
| **Instancia 3** | **Excelente** | **No** |  |

**Ejercicio.**

**Explica el papel que juega el ratio de aprendizaje en el método del descenso de gradiente.**

**Ejercicio.**

**Explica en qué consiste el método de descenso de gradiente.**

**Ejercicio.**

**En regresión logística, ¿qué característica debe tener una función para reemplazar a la función sigmoide?**

**Ejercicio.**

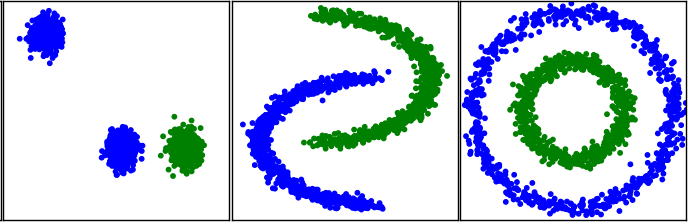
**En regresión logística, si la función sigmoide nos devuelve el valor 0.6 para una instancia, ¿qué etiqueta asignaremos a dicha instancia?**

**Ejercicio.**

**Explica cómo la regresión logística puede ser aplicada para abordar problemas de clasificación múltiple.**

**Ejercicio.**

**Dados los siguientes gráficos donde los puntos verdes representan una clase y los puntos azules otra, ¿cuáles son linealmente separables?**

****

**Ejercicio.**

**¿Qué son los support vectors?**

**Ejercicio.**

**Intuitivamente, ¿en qué consiste el truco del kernel en el algoritmo SVM?**

**Ejercicio.**

**Explica la diferencia entre las estrategias uno-contra-todos y uno-contra-uno aplicadas en SVMs para clasificación múltiple.**

**Ejercicio.**

**¿Es posible utilizar el perceptrón para clasificar de manera correcta conjuntos de datos que no son linealmente separables? ¿Por qué?**

**Ejercicio.**

**El perceptrón en su formulación original, ¿sirve para resolver problemas de clasificación binarios o multi-clase?**

**Ejercicio.**

**En el aprendizaje de las redes neuronales, ¿qué es una época?**

**Ejercicio.**

**¿Cuándo está asegurada la terminación en el algoritmo de aprendizaje del perceptrón?**

**Ejercicio.**

**En caso de que el conjunto de entrenamiento no sea linealmente separable, ¿cuándo termina el proceso de aprendizaje del perceptrón?**

**Ejercicio.**

**Explica por qué es fundamental elegir un ratio de entrenamiento que no sea excesivamente grande o excesivamente pequeño en el entrenamiento de una red neuronal.**

**Ejercicio.**

**¿Qué técnica utiliza el método de backpropagation para encontrar los pesos que minimizan el error durante el proceso de entrenamiento?**

**Ejercicio.**

**¿Cuáles son los hiperparámetros de las redes neuronales?**

**Ejercicio.**

**Explica por qué son importantes las funciones de activación en las redes neuronales.**

**Ejercicio.**

**¿Qué dice el “no free lunch theorem”?**

**Ejercicio.**

**Describe qué son los verdaderos positivos, los verdaderos negativos, los falsos positivos y los falsos negativos.**

**Ejercicio.**

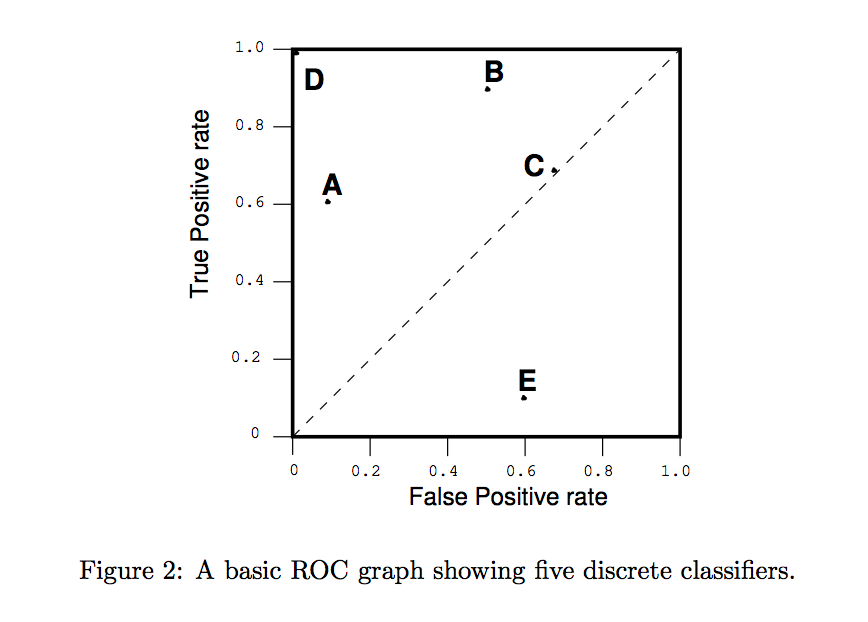
**Tenemos un clasificador capaz de determinar si una imagen contiene un gato o no. De las 100 imágenes que disponemos, 50 de ellas contienen un gato y las otras 50 no. De las 50 imágenes con gato, el clasificador dice que hay gato en 30 de ellas, y en las otras 20 dice que no. De las 50 imágenes sin gato, el clasificador dice que hay gato en 10 de ellas, y en las otras 40 dice que no. A partir de esta información dar los valores de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos; y calcular los siguientes ratios: sensitivity (definido como TP/(TP+FN)), specificity (definido como TN/(FP+TN)) y accuracy (definido como (TP+TN)/(TP+FP+FN+TN)).**

**Ejercicio.**

**¿Por qué nos interesa que una matriz de confusión sea diagonal?**

**Ejercicio.**

**En la siguiente imagen de un espacio ROC ¿cuál es el mejor clasificador? ¿y el peor? ¿cuál es el equivalente a predecir de manera aleatoria?**

****

**Ejercicio.**

**Resume en qué consiste el proceso de evaluación y por qué es importante separar el dataset inicial en conjuntos de entrenamiento, validación y test.**

**Ejercicio.**

**¿Qué problema resuelve la estratificación? y ¿cómo lo resuelve?**

**Ejercicio.**

**¿Qué problema resuelve la validación cruzada? y ¿cómo lo resuelve?**

**Ejercicio.**

**¿Por qué el proceso de validación cruzada es un proceso costoso?**

**Ejercicio.**

**Explica cuándo ocurre el subajuste y cuándo el sobreajuste.**

**Ejercicio.**

**Explica qué es el sesgo y qué es la varianza.**

**Ejercicio.**

**¿Qué indica un sesgo alto? ¿y una varianza alta? Indica algoritmos que tengan un sesgo bajo y una varianza alta.**

**Ejercicio.**

**Utilizando las curvas de validación (o aprendizaje) ¿qué indica que se obtengan errores similares para un rango amplio de tamaños de conjunto de entrenamiento? ¿y que haya una diferencia considerable entre el error de entrenamiento el de validación cruzada?**

**Ejercicio.**

**¿Cuáles son las soluciones para reducir la varianza y el sesgo?**

**Ejercicio.**

**¿De qué partes consta la búsqueda en el espacio de hiperparámetros?**

**Ejercicio.**

**Explica la diferencias entre buscar hiperparámetros utilizando la técnica de GridSearch y la de RandomSearch.**

**Ejercicio.**

**¿Para qué sirven la normalización y la estandarización?**

**Ejercicio.**

**Explica la metodología CRISP-DM.**

**Ejercicio.**

**Explica la metodología OSEMN.**