AFIRMACIONES VERDADERAS

- MÉTODOS PARAMÉTRICOS Y NO PARAMÉTRICOS
    - El número de parámetros en un modelo no paramétrico es diferente de 0.
    - La suposición fundamental de los métodos no paramétricos es que el comportamiento del conjunto de datos bajo análisis presenta cambios suaves y por lo tanto se pueden hacer predicciones para una muestra nueva con base en el comportamiento de la vecindad de dicha muestra
    - Durante la etapa de validación los modelos NO paramétricos son en general más costosos computacionalmente que los modelos paramétricos.

- REGRESIÓN LOGÍSTICA Y REGRESIÓN POLINÓMICA
    - Desde el punto de vista del aprendizaje de máquina una regresión polinomial y una regresión lineal, pueden abordarse de la misma manera porque ambos son lineales en los parámetros del modelo
    - En una regresión polinomial, el criterio de minimización del error cuadrático medio es equivalente al criterio de máxima verosimilitud, asumiendo una función de densidad Gaussiana sobre los datos con media dada por la función polinomial de varianza constante.
    - El grado del polinomio en un problema de regresión es un parámetro que debe ser ajustado a través de una estrategia de validación cruzada
    - Una regresión logística es un modelo de tipo discriminativo.
    - Una regresión logística es un modelo paramétrico que puede ser usado únicamente para resolver problemas de clasificación.
    - El número de parámetros que deben ser ajustados durante el entrenamiento de una regresión logística crece con el grado del polinomio que se desea ajustar.
    - El algoritmo de entrenamiento de una regresión logística y de una regresión polinomial puede ser el mismo.
    - Los criterios de entrenamiento de una regresión logística y de una regresión polinomial son diferentes.
    - El número de parámetros que deben ser ajustados durante el entrenamiento de una regresión logística, crece con el grado del polinomio que se desea ajustar.
    - Las regresiones múltiple y logística son modelos paramétricos en los cuales el número de parámetros es igual al número de variables mas 1

- MEZCLAS DE FUNCIONES GAUSSIANAS
    - Un modelo de mezcla de funciones Gaussianas es un modelo paramétrico y no supervisado que tiene la capacidad de modelar diferentes funciones arbitrarias de densidad de probabilidad.
    - El modelo GMM se puede usar en problemas de clasificación y también en problemas de agrupamiento.
    - El número de componentes o funciones gaussianas que se usan en el modelo GMM (mezcla de funciones gaussianas) es independiente del número de clases en un problema de clasificación

- El algoritmo de Esperanza y Maximización corresponde a la aplicación del criterio de máxima verosimilitud.
- El método de agrupamiento K-means se puede considerar una simplificación del algoritmo EM.
- El método de agrupamiento K-means hace parte de los métodos de aprendizaje no supervisados
- El criterio de entrenamiento del método K-means se puede interpretar como la minimización de la dispersión intra-cluster.

- FUNCIONES DISCRIMINANTES GAUSSIANAS Y NAÏVE BAYES
  - El modelo de funciones discriminante Gaussianas es un caso particular de un modelo GMM.
  - Un modelo de funciones discriminantes Gaussianas es de tipo generativo
  - Un modelo de clasificación basado en funciones discriminantes Gaussianas con matrices de covarianza completa tiene menor capacidad de modelamiento que un modelo similar pero con matrices de covarianza esféricas.
  - En un modelo de mezcla de funciones gaussianas que usa matrices de covarianza diagonal las componentes pueden formar elipsoides cuyo eje principal es paralelo a alguno de los ejes coordenados
  - Un clasificador Naïve Bayes corresponde a un método generativo
  - Un clasificador de Naive Bayes es equivalente a un clasificador discriminante Gaussiano con matriz de covarianza diagonal.
  - Un clasificador Naïve Bayes asume que tanto las muestras como las variables son independientes

- MÉTODOS NO PARAMÉTRICOS BASADOS EN DISTANCIAS (PARZEN, k-NN)
  - Los métodos k-NN y ventana de Parzen no tienen fase de entrenamiento
  - El modelo de k-NN asigna una muestra a la clase más probable en el conjunto de vecinos de dicha muestra
  - El área definida como “vecindad” en el modelo de K-vecinos más cercanos, varía para cada muestra.
  - La función de distancia y el ancho de la ventana kernel son hiper parámetros del algoritmo k-NN
  - El tipo de medida de distancia en el modelo de k-NN se considera un hiperparámetro del modelo
  - La medida de distancia se puede considerar un hiperparámetro del modelo de ventana de Parzen
  - La estrategia de clasificación usando el modelo de ventana de Parzen se puede considerar un método generativo.
  - La estrategia de clasificación usando el modelo de ventana de Parzen es similar a los métodos generativos porque se deben separar las muestras por clase y modelar una fdp por clase

- ÁRBOLES DE DECISIONES, RANDOM FOREST, BOOSTING Y BAGGING
  - Un árbol de decisión es un modelo de tipo NO paramétrico.

- El índice gini se puede usar como medida de impureza de un espacio de características durante el entrenamiento de árboles de decisión.
- La ganancia de información es una medida de la bondad de una partición del espacio de características en un problema de clasificación
- En un árbol de regresión, cada nodo terminal tiene asociados un solo valor de predicción.
- Un árbol de regresión siempre requiere de un límite de crecimiento o de la aplicación de un método de poda.
- En el método Random Forest el conjunto de árboles generados para clasificación NO se podan.
- En el modelo conocido como árbol extremadamente aleatorio, además de escoger aleatoriamente un subconjunto de variables, también se escoge aleatoriamente el umbral de partición
- El bagging de árboles busca reducir la varianza en el error (variance) mientras que el boosting busca reducir el sesgo en el error (bias)

- REDES NEURONALES
  - La diferencia fundamental entre una RNA entrenada para un problema de regresión y una RNA entrenada para un problema de clasificación está en la función de activación en la capa de salida.
  - El bias o término independiente (w0) debe ser incluído en los perceptrones de todas las capas de una red neuronal artificial
  - El algoritmo de propagación hacia atrás puede ajustar los pesos de una red neuronal artificial con un número arbitrario de capas
  - El algoritmo de propagación hacia atrás es en realidad una extensión de los métodos de gradiente usando la regla de la cadena en la derivada del error
  - El algoritmo backpropagation consiste en propagar el error medido en la capa de salida, a las capas ocultas de la red neuronal usando la regla de la cadena de la derivada para ajustar los pesos de toda la red
  - Los perceptrones multi-capa (MLPs) son un tipo de RNA feed forward o de propagación hacia adelante
  - Los hiperparámetros de una red neuronal artificial tipo MLP (perceptrón multicapa) que deben ser ajustados a través del proceso de validación son:
    - Número de capas ocultas de la red neuronal
    - Función de activacion
  - Durante el entrenamiento de una red neuronal artificial tipo MLP (Perceptrón multicapa) se deben ajustar tanto los hiperparámetros de la red, como los hiperparámetros del algoritmo backpropagation
  - La función softmax es una generalización de la función sigmoidal para más de dos clases
  - Para más de dos clases y una codificación one-hot encoding de la clase a predecir, es recomendable usar una función de activación softmax.
  - El aprendizaje on-line difiere del aprendizaje tipo batch en el algoritmo de aprendizaje

    - Para un mismo número de épocas M, el esquema de entrenamiento on-line realiza un mayor número de actualizaciones de los pesos de una red neuronal artificial que el esquema tipo batch
    - El aprendizaje tipo on-line en un problema de clasificación podría no converger si en cada época del algoritmo las muestras se evalúan en órden, primero una clase y luego la siguiente
    - El aprendizaje on-line puede usarse para realizar entrenamiento de una red neuronal artificial que no ha sido entrenada previamente y también para ajustar una red que ya fue entrenada pero se requiere ajustar con muestras nuevas
    - El esquema de entrenamiento on-line es más susceptible a la tasa de aprendizaje que el esquema tipo batch o también mini-batch
    - El aprendizaje tipo on-line puede no llegar a encontrar un buen modelo si se usa una técnica de gradiente descendente común.
    - El aprendizaje tipo batch realiza la acumulación de los errores a partir de todas las muestras y posteriormente actualiza los parámetros del modelo
    - El aprendizaje tipo batch (por lotes) puede obtener resultados satisfactorios pero tiene mayores requerimientos en memoria que el aprendizaje tipo on-line
    - El aprendizaje tipo batch (por lotes) puede obtener resultados satisfactorios si se usa una técnica de gradiente descendente estocástico en lugar del gradiente descendente común.
    - El tamaño del mini-batch es un hiperparámetro del algoritmo de entrenamiento de una RNA

- MÁQUINAS DE SOPORTE VECTORIAL
    - Una máquina de soporte vectorial es un modelo de tipo no paramétrico
    - El *kernel* en una máquina de soporte vectorial es una función que mide la distancia (o similitud) entre dos muestras y es la que genera fronteras de decisión lineales o no lineales
    - La maximización del margen es un criterio derivado del concepto de minimización del riesgo estructural
    - En una máquina de soporte vectorial utilizada para clasificación, los vectores de soporte son las muestras más cercanas a la frontera y su margen
    - En una máquina de soporte vectorial utilizada para regresión, los vectores de soporte son las muestras más retiradas de la frontera y su márgen.

- El dendrograma es un tipo particular de gráfico que permite visualizar el resultado de la aplicación de una técnica de clustering jerárquico.
- Un sistema de clasificación basado en modelos generativos ajusta modelos de manera independiente por cada clase.
- Un sistema de clasificación basado en modelos discriminativos ajusta un sólo modelo que corresponde a la frontera de separación entre las clases.
- Controlar el particionamiento de las muestras durante la implementación de la metodología de validación.
- Aplicar una estrategia de sobre-muestreo sobre la clase minoritaria.

- Asignar durante el entrenamiento pesos diferentes a los errores en que incurre el modelo para cada clase.
- Un método de submuestreo inteligente elimina muestras redundantes y datos atípicos de la clase mayoritaria.
- La curva ROC es un indicador del desempeño esperado de un sistema de clasificación.
- SMOTE es un método de sobremuestreo inteligente.
- El dendograma es un tipo particular de gráfico que permite visualizar el resultado de la aplicación de una técnica de clustering jerárquico
- Un sistema de clasificación basado en modelos discriminativos ajusta un sólo modelo que corresponde a la frontera de separación entre las clases
- Un sistema de clasificación basado en modelos generativos ajusta modelos de manera independiente por cada clase
- En un problema de clasificación biclase en el que cada clase se modela a partir de una función de densidad Gaussiana con media dada por el conjunto de muestras de cada clase, y con varianza constante para todas las clases, la frontera de decisión tiene la forma de una recta.
- Los modelos generativos pueden generar muestras artificiales, los discriminativos no.
- La curva ROC se obtiene a partir de graficar 1 - Especificidad vs Sensibilidad para diferentes umbrales de decisión
- Entrenar modelos capaces de operar bajo condiciones de alta variabilidad intra-clase
- Seleccionar los mejores hiperparámetros con base en la media geométrica entre la sensibilidad y la especificidad
- El umbral EER se estima sobre la función de distribución acumulada de los scores de salida del modelo, mientras que el umbral MCP se estima sobre la función de densidad
- El algoritmo k-means se puede entender con un caso particular del algoritmo EM cuando se asume que la covarianza de todas las componentes es la misma y que todas las componentes tienen el mismo peso en el modelo.

AFIRMACIONES FALSAS

- REGRESIÓN LOGÍSTICA Y REGRESIÓN POLINÓMICA
    - Una regresión logística es un modelo de aprendizaje que puede ser usado tanto en problemas de regresión como de clasificación
    - ➢ Una regresión logística es un modelo paramétrico que puede ser usado para resolver problemas tanto de regresión como de clasificación
    - ➢ El criterio de entrenamiento para los modelos de regresión logística y de regresión polinomial o múltiple es el mismo.
    - ➢ La ganancia de información es una medida de la bondad de una partición del espacio de características en un problema de regresión.

- FUNCIONES DISCRIMINANTES GAUSSIANAS

- ➢ El criterio de entrenamiento de un modelo basado en funciones discriminantes gaussianas es la maximización de la discriminación entre clases.
- ➢ Los modelos de funciones discriminantes Gaussianas y modelos de mezcla de Gaussianas asumen que las muestras son i.i.d. mientras que la regresión polinomial NO.
- ➢ Un modelo de clasificación basado en funciones discriminantes gaussianas con matrices de covarianza completa tiene menor capacidad de modelamiento que un modelo similar pero con matrices de covarianza esféricas.
- ➢ El criterio de entrenamiento de un modelo basado en funciones discriminantes Gaussianas es la minimización del error de clasificación.
- ➢ Un clasificador Naïve Bayes corresponde a un método discriminativo.
- ➢ En un problema de clasificación biclase en el que cada clase se modela a partir de una función de densidad Gaussiana con media dada por el conjunto de muestras de cada clase, y con varianza constante para todas las clases, la frontera de decisión tiene la forma de una parábola.
- ➢ El nombre “Naïve (ingenuo) del clasificador Naïve Bayes se le da debido a que asume independencia entre las muestras del conjunto de datos de entrenamiento

- MÉTODOS NO PARAMÉTRICOS BASADOS EN DISTANCIAS (PARZEN, k-NN)

  - ➢ El criterio de entrenamiento del método k-means se puede interpretar como la maximización de la dispersión intra-cluster.
  - ➢ El número de parámetros en un modelo no paramétrico es 0

- ÁRBOLES DE DECISIONES, RANDOM FOREST, BOOSTING Y BAGGING
  - ➢ Bagging es una técnica de combinación de modelos que sólo se puede aplicar a árboles de decisión
  - ➢ En el método conocido como Random Forest se escoge aleatoriamente un subconjunto de de variables para ser analizadas en cada árbol.

- MEZCLAS DE FUNCIONES GAUSSIANAS
  - ➢ El modelo GMM es un modelo discriminativo
  - ➢

- UNSUPERVISED LEARNING (K-means)
  - ➢ El método de agrupamiento K-means se puede usar para resolver problemas supervisados.

- REDES NEURONALES ARTIFICIALES (BACKPROPAGATION)
  - ○ En el aprendizaje tipo on-line se pueden usar indistintamente el gradiente descendente común o el gradiente descendente estocástico
  - ○ El aprendizaje on-line difiere del aprendizaje tipo batch en la función de costo o criterio usado para el ajuste de los pesos de una red neuronal artificial

- ➢ El aprendizaje on-line puede usarse únicamente cuando el sistema se encuentra ya en fase de producción
- ➢ El bias o término independiente sólo es necesario incluirlo en los perceptrones de la capa de entrada de una red neuronal artificial
- ➢ El algoritmo de propagación hacia atrás puede ajustar los pesos de una red neuronal artificial con un número máximo de tres capas
- ➢ El algoritmo de propagación hacia atrás garantiza encontrar los pesos óptimos que minimizan la función de costo en una red tipo MLP

- REDES NEURONALES ARTIFICIALES (BACKPROPAGATION)
  La forma de determinar la convergencia del algoritmo k-means es evaluando si este alcanzó el máximo número de convergencias (iteraciones)

- VARIADAS

➢ Usar todas las muestras de la clase minoritaria en la fase de entrenamiento y dejar únicamente muestras de la clase mayoritaria para la fase de validación.
➢ El umbral EER(Igual Tasa Error) se estima sobre la función de densidad de los scores de salida del modelo, mientras que el umbral MCP(Mínimo Coste)se estima sobre la función de distribución acumulada.
➢ Usar el umbral de mínimo coste para tomar la decisión final
➢ SMOTE es un método de submuestreo inteligente
➢ Para más de dos clases y una codificación one-hot encoding de la clase a predecir, es recomendable usar una función de activación sigmoidal.
➢ Para más de dos clases y una codificación one-hot encoding de la clase a predecir, es recomendable usar una función de activación sigmoidal

Preguntas

¿Qué consideraciones debe tener un método de submuestreo “inteligente”?

Respuesta Debe identificarse la presencia de datos atípicos o de datos redundantes para descartarlos

¿Que problema presenta el entrenamiento de una red neuronal artificial usando el algoritmo tipo batch?

Respuesta La trayectoria del gradiente batch tiende a estancarse en puntos silla de la función de costo.

¿A qué tipo corresponde un modelo predictivo, donde los datos de entrenamiento son requeridos explícitamente para realizar el test de una nueva muestra?

Respuesta Modelo no paramétrico, es necesario almacenar todas las muestras de entrenamiento. Un ejemplo es el método de K-vecinos más cercanos en el que se

mide la distancia entre la nueva muestra y todas las muestras almacenadas en el entrenamiento para encontrar los vecinos más cercanos.

Cuál de las siguientes definiciones corresponde a un ensamble de clasificadores tipo voting:

Una combinación de diferentes modelos entrenados sobre el conjunto de entrenamiento original.

Explique una forma simple de resolver un problema de aprendizaje que corresponde a un paradigma de aprendizaje de múltiples instancias ¿como se puede usar un modelo supervisado convencional y como se toma una decisión final?

Respuesta Lo primero es convertir esas múltiples instancias en una sola característica sacando el promedio de estas.

¿Que diferencia existe entre árboles de decisión para resolver problemas de clasificación y para problemas de regresión?

Respuesta:

¿Que tipo de función de activación se debe usar en la capa de salida de una red neuronal entrenada para resolver un problema de clasificación de 4 clases, las cuales se codifican usando one-hot encoding?

¿Para que se usa la entropía en el entrenamiento de árboles de decisión?

Respuesta:

Como métrica de la impureza en cada nodo del árbol y como parte de la medición de la ganancia de información.

¿Que función de activación debe usar en la capa de salida de una red neuronal artificial si el problema que estoy resolviendo es de múltiples etiquetas?

Respuesta

La función de activación Sigmoide

¿por que el algoritmo de gradiente que se usa en el aprendizaje on-line se conoce como gradiente estocástico?

Respuesta:

Porque en lugar de consistir en una reducción progresiva de la función de error a partir de la regla de la cadena , consiste en una reducción de esa función de error. En base a una caminata aleatoria.

¿Cual de los modelos vistos en clase es equivalente a un solo perceptron?

Respuesta:
Regresión logística debido a que su salida se basa en la función sigmoide.

¿Cómo se llama el algoritmo de entrenamiento de una red neuronal recurrente?

Backpropagation en el tiempo (BTT ó backpropagation through time)

¿Cuales son los tres pasos del algoritmo de entrenamiento de un mapa auto-organizable (SOM)? Si no recuerda los nombres describa qué se hace en cada paso.
Respuesta
Competición
Cooperación
Adaptación

¿Por qué las capas internas de una red neuronal artificial se llaman capas ocultas?
Respuesta:
Los valores y resultados parciales entre cada una de las capas ocultas no son visibles durante el proceso de operación de la red neuronal

¿cuales de los siguientes modelos son discriminativos y cuales generativos, cuales paramétricos y cuales no paramétricos ?
Respuesta:
ventana de parzen-no paramétrico-generativo
RNA-paramétrico-discriminativo
SVM-no paramétrico-discriminativo
GMM-paramétrico-generativo

¿Qué es la matriz de confusión?
Respuesta: Matriz que presenta la cantidad o porcentaje de muestras de cada clase, clasificadas correctamente y las clasificadas en otra clase (mal clasificadas.)

Es una herramienta en la que se compara el desempeño de un algoritmo respecto a la predicción de salidas con las salidas reales que se observaron en la toma de datos original. Se crea una matriz en la que en las filas representan las salidas observadas en la realidad y las columnas representan las predicciones hechas por el algoritmo. Cada cruce fila-columna indica cuántos elementos en la predicción fueron de una clase siendo en realidad de la misma u otra clase.

¿Qué son o cómo se definen la sensibilidad, la especificidad y la precisión?

°**Sensibilidad:** Mide el porcentaje de muestras positivas clasificadas correctamente:
TP/(TP+FN)

°Especificidad: Mide el porcentaje de muestras negativas clasificadas correctamente:
TN/(TN+FP)

°Precisión: Mide el porcentaje de muestras clasificadas como positivas, que en realidad debían ser positivas:
TP/(TP+FP)

Sensibilidad o tasa de verdaderos positivos: Mide para un modelo clasificador cuántos verdaderos positivos logra en la clasificación:

Sensibilidad = # verdaderos positivos / (# verdaderos positivos + # falsos negativos)

Especificidad o tasa de verdaderos negativos: Mide para un modelo clasificador cuántos verdaderos negativos logra en la clasificación:

Especificidad = # verdaderos negativos / (# verdaderos negativos + # falsos positivos)

Precisión: Define la tasa de predicción de verdaderos positivos:

Precisión = # verdaderos positivos / (# verdaderos positivos + # falsos positivos)

Durante la validación de un modelo, ¿por qué es importante tener en cuenta el intervalo de confianza de las medidas de desempeño?

> **Respuesta:**Es importante ya que nos expresa un margen dentro del cual el resultado puede moverse, con el cual se puede dar una buena predicción.

¿En qué consiste un paradigma de aprendizaje de múltiples instancias?

> **Respuesta:** Consiste en tener muestras que estén representadas por más de una instancia, se puede ver como varias ventanas de datos que hacen referencia a una sola muestra.

¿En qué consiste el problema del desbalance y qué efectos causa?

Respuesta: Muestras de una clase mucho mayor a la cantidad de otra clase. El efecto es que al seleccionar muestras de entrenamiento y validación, el subconjunto de entrenamiento puede quedar con muestras de una sola clase o muy pocas de la otra, esto causaría que el sistema aprendiera a clasificar solo una clase.

¿Qué función cumple la ventana o kernel en el modelo ventana de Parzen?

Respuesta: Es la función encargada de asignar un peso a cada muestra con respecto a la distancia entre esta y la muestra a clasificar.

¿Qué diferencia existe entre la técnica de BAGGING estándar a partir de árboles de decisión y el método conocido como Random Forest?

Respuesta: Bagging construye M número de bases de datos(BD) tomando subconjuntos aleatorios de la BD original y construye en este caso un árbol para cada BD, por su parte, Random Forest construye M árboles utilizando todas las muestras de entrenamiento pero aleatoriza las variables a medir en cada nodo de decisión.

- **Explique las diferencias entre las metodologías: Validación Cruzada y Bootstrapping**

Respuesta La validación cruzada define un número K de subconjuntos, utiliza K-1 para entrenar y 1 para validar, se repite usando un conjunto diferente para validar. En cambio, Bootstrapping define un porcentaje de muestras para entrenar y uno para validar, pero en cada iteración se vuelven a particionar las muestras de manera aleatoria.

¿Cuáles son las tres estrategias básicas de clasificación multiclase basadas en clasificadores bi-clase?

- Uno contra uno (one vs one)
- Uno contra todos (one vs rest)
- Jerárquica

En un problema de aprendizaje con M número de salidas (M variables a predecir), ¿cuándo es útil utilizar un modelo de aprendizaje que utilice información de todas las salidas en conjunto en lugar de dividir el problema en M problemas individuales?

Respuesta Se considera necesario cuando las variables de salida están relacionadas.

El considerar si se debe utilizar en conjunto información de todas las salidas depende de si se sospecha o se sabe de la existencia de correlación o de independencia entre variables de salida. Mientras se pueda considerar que haya independencia entre las salidas, se puede dividir el problema en P problemas individuales, caso contrario a que se pueda considerar que haya relación entre las variables de salida.

¿Por qué razón los árboles de decisión hacen parte de los métodos NO métricos?

Respuesta Porque de entrada no se realizan suposiciones sobre la forma de los datos, de esta manera la información para construir el árbol es extraída directamente de los datos.

¿Explique la diferencia entre un paradigma de aprendizaje de múltiples clases VS uno de múltiples etiquetas?

Respuesta En múltiples clases se tienen muestras de diferentes clases, pero cada una de ellas pertenece a una sola clase, por otro lado, múltiples etiquetas manejan muestras que pertenecen a más de una clase (una muestra está etiquetada en más de una clase)

En el aprendizaje de múltiples clases se encontrarán situaciones en las que las clases se excluyen mutuamente, en el aprendizaje de múltiples etiquetas se pueden encontrar situaciones en las que a un objeto clasificado se le puede asignar varias etiquetas simultáneas

¿Que es smote?

respuesta: Es un método de submuestreo inteligente en donde se crean muestras sintéticas . 2 Muestras y tomando un punto de la recta que las une, esto con el fin de favorecer la clase minoritaria.

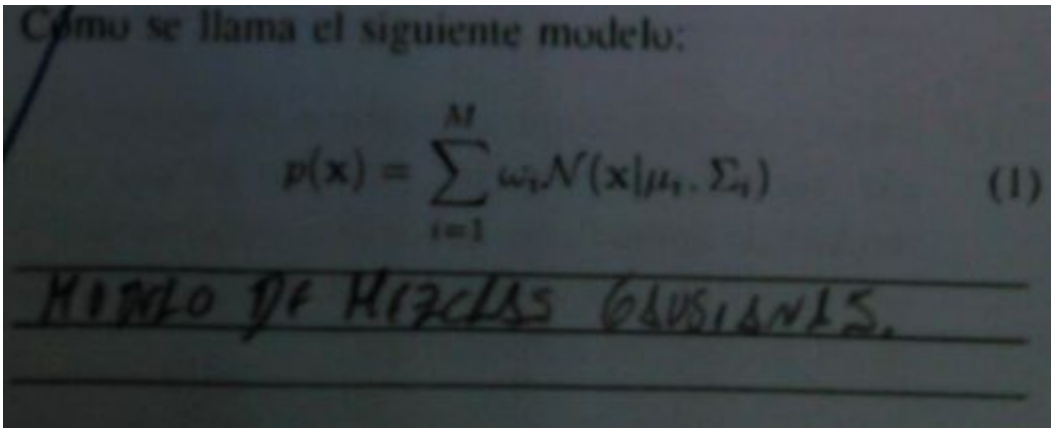
Diga una ventaja y una desventaja de los modelos paramétricos vs los no paramétricos

respuesta:

Ventaja: Los paramétricos son menos costosos computacionalmente que los no paramétricos cuando se tiene una base de datos de muestras muy extensas.

Desventaja: el modelo paramétrico hace suposiciones sobre la función que le presenta los datos, por tanto habrán algunos que no se puedan predecir correctamente.

¿Como se llama el siguiente modelo ?



Durante la validación de un modelo ¿por que es importante tener en cuenta el intervalo de confianza de las medidas de desempeño?
Respuesta:

¿En qué consiste un paradigma de aprendizaje de múltiples instancias?
respuesta:
En que cada muestra está descrita por varios vectores de características y no por uno sólo.

¿Que diferencia existe entre el método random forest y el extremely randomized tree?
Respuesta:

¿Como se llama el criterio de entrenamiento y el algoritmo de entrenamiento del modelo GMM vistos en clase?

Respuesta:

Criterio de máxima verosimilitud
algoritmo esperanza y maximización

Suponga que le solicitan modificar un sistema de información médico en el que se encuentran las historias clínicas de millones de pacientes, con el objetivo de adicionar una nueva funcionalidad al sistema para que a partir de las historias clínicas agrupe los pacientes de acuerdo con un concepto conocido como Grupo Relacionado de Diagnóstico es decir, pacientes con condiciones clínicas similares. El objetivo es que un médico pueda identificar pacientes similares a alguno que esté tratando y de esa manera pueda determinar si el tratamiento que piensa prescribir ha sido o no efectivo y optimizar de esa manera el tiempo y los recursos. El problema de establecer Grupos Relacionados de Diagnóstico podría resolverse a través de técnicas de aprendizaje supervisadas o no supervisadas. Explique su respuesta

Respuesta:

No supervisada, se parte de un conjunto de características y lo que se busca son patrones de comportamiento interesantes, para este caso se asocian condiciones clínicas similares y se puede determinar el tratamiento a prescribir.

En una máquina de vectores de soporte (SVM) el parámetro C también llamado restricción caja, se puede considerar como un parámetro de regularización ¿Si el valor de C es grande eso significa mayor o menor regularización?

Respuesta:

Si C es alto se obtiene una menor regularización

Enumere los 4 pasos del algoritmo k-means

1. **1**
2. **1**
3. **1**
4. **1**

¿Cuántos modelos de clasificación deben ser entrenados si se usa una estrategia uno-contra-uno (en inglés one vs one)?

Si se tienen k clases, se deben entrenar [k (k - 1)] / 2 modelos

Por qué los sistemas de reconocimiento de patrones que usan técnicas de extracción de características, durante la etapa de producción incluyen una fase adicional de procesamiento de información, mientras que los que usan selección de características no la usan?

Respuesta://

Debido a que la extracción necesita realizar el procesamiento de las características para poder obtener las nuevas características que aportarán mayor información. Mientras que la selección solo busca eliminar las que no aportan información y se puede realizar antes de la fase de producción para evitar trabajar con características que no aportan información.

La regularización es una estrategia en la que se le adiciona un término a la función objetivo, para evitar el sobreajuste. El error de entrenamiento en un modelo muy regularizado tiende a ser mayor o menor que en un modelo sin regularización? por qué?

Respuesta://

La regularización consiste en agregar una cantidad adicional al resultado del cálculo de la función objetivo que, en muchos casos es la función de error cuadrático medio; entonces a mayor regularización, más alta es la adición hecha a dicha función. Se concluye que el error de entrenamiento es mayor mientras más regularizado esté el modelo.

Describa dos ventajas de la selección de características en comparación con la extracción de características

Respuesta://

- En la selección de características no se transforman las características originales en características nuevas sin relación con el contexto original del problema
- Los criterios de selección de características (correlaciones mutuas, covarianzas respecto a la variable de respuesta) son relativamente
- La selección de características, sea usando criterios tipo filtro o criterios tipo wrapper, gozan de capacidad de generalización
- En la selección de características se tienen tiempos de ejecución más reducidos

El análisis de componentes es una técnica supervisada o no supervisada? por qué?
Es una técnica no supervisada porque en ella se considera distintas combinaciones lineales de las variables predictoras (las características) para definir un nuevo espacio de características de dimensión reducida, independientemente del conjunto de variables respuesta.

Por qué el modelo SVM es un modelo disperso?
Porque para definir la frontera del hiperplano de clasificación/regresión se usa cantidades menores del conjunto de datos.

En una máquina de vectores de soporte (SVM) ¿Un mayor nivel de regularización implica un mayor o menor número de vectores de soporte?
A mayor regularización, mayor número de vectores de soporte

Por qué en el análisis de PCA, las direcciones principales que se usan para transformación de datos, corresponden a los vectores propios asociados a los mayores valores propios y no a los menores?
Respuesta://

> El objetivo es que la varianza sea muy grande, entonces eso implica que el vector propio que debemos seleccionar es aquel que esté asociado al mayor valor propio.
>
> Así mismo, se busca que haya independencia lineal en entre cada una de las nuevas variables que conformen el nuevo espacio de características, esto exige que haya pocas o ninguna variable expresable como combinación lineal de las demás variables.

En el contexto de métodos de selección secuencial de características ¿por qué los criterios tipo filtro tienen mayor generalidad pero menor capacidad de generalización que los criterios tipo Wrapper?

Respuestas://

> El filtro tiene como criterio una función genérica evaluando las características por su contenido de información, la distancia entre clases o correlación y con esto obtiene una mayor generalidad. Mientras que el wrapper tiene como criterio el error obtenido en la validación lo cual lo hace que aumente la capacidad de generalización.

En una empresa fabricante de vehículos se quiere desarrollar un sistema de visión artificial que realice un reconocimiento automático e identifique los tipos de objetos capturados en una imágen tomada por una cámara que va instalada abordo de un vehículo. En la imágen pueden aparecer simultáneamente diferentes tipos de objetos, árboles, señales de tránsito, personas, perros, etc. El objetivo es diseñar un sistema basado en técnicas de Machine Learning que pueda solucionar el requerimiento planteado por la empresa. ¿A qué tipo de paradigma de aprendizaje corresponde el problema planteado en el párrafo anterior?

Respuesta:// Es supervisado porque se conocen las clases y además de múltiples etiquetas.

¿Una red neuronal es un modelo paramétrico o no paramétrico? Justifique su respuesta.

Respuesta://

Son paramétricos ya que tienen una gran cantidad de parámetros, uno para cada peso que se ajusta durante el entrenamiento.

Como el número de pesos generalmente se mantiene constante, técnicamente tienen grados fijos de libertad.

¿Por Qué el algoritmo de gradiente descendente que se usa en el aprendizaje *on-line* se conoce como gradiente descendente estocástico ?

Al usarse una muestra a la vez durante el proceso de entrenamiento, la trayectoria que sigue el algoritmo para encontrar el mínimo de la función de costo se asemeje a una caminata aleatoria, causa por la cuál se tiene un gradiente descendente estocástico

¿Cuál es la principal diferencia entre la función de activación tangente hiperbólica y la función de activación sigmoidal?

La diferencia es que la función de activación de la tangente hiperbólica varia de -1 y 1 y el gradiente es más fuerte que para la sigmoidal.

En la sigmoidal su variación es entre 0 y 1.

¿Cuántos modelos de clasificación deben ser entrenados si se usa una estrategia Uno contra Uno (en inglés One vs One)?

Respuesta://

Se deberían entrenar (NumClasses * (NumClasses - 1)) / 2 clases

Dé un ejemplo de aplicación donde la entrada corresponda a una secuencia de observaciones y la salida corresponda a una sola clase para toda la secuencia.

¿En qué consiste la estrategia de comité de máquinas conocida como Boosting ?

El objetivo de Boosting es entrenar múltiples modelos simples y combinar sus respuestas para producir un mejor resultado mediante la combinación de una manera secuencial.

La idea es que los modelos subsiguientes se enfoquen en las muestras que no han podido ser modeladas correctamente por los modelos anteriores.

Es una combinación en serie de máquinas en las que la máquina siguiente en el proceso parte de los resultados obtenidos por la máquina anterior, siempre evaluando y comparando la función de costo y buscando la reducción de errores resultantes del trabajo de la máquina anterior.

¿Cuál es la diferencia entre un ensamble de clasificadores tipo voting y uno tipo stacking?

La diferencia fundamental entre voting y stacking es cómo se realiza la agregación final. En voting, los pesos especificados por el usuario se utilizan para combinar los clasificadores, mientras que en el stacking realiza esta agregación utilizando una mezcla / meta clasificador.

La diferencia ente los dos es cómo se toma la decisión final a partir de la combinacion de varios modelos.

Dé un ejemplo de aplicación donde la entrada corresponda a una secuencia de observaciones y la salida corresponda a una sola clase para toda la secuencia.
Clasifique los siguientes modelos como generativos o discriminatorios, como paramétricos o no paramétricos

¿Qué diferencia existe entre un paradigma de múltiples etiquetas y uno de múltiples etiquetadores?

Dé un ejemplo de aplicación donde la entrada (el objeto sobre el cual se desea hacer una predicción) corresponda a una única observación (puede ser un vector, una matriz, etc.) y la salida corresponda a una secuencia observaciones (secuencia de valores que tienen un orden, puede ser temporal, espacial, etc.).

¿Qué diferencia existe en un modelo Gradient Boosting Tree y un Stochastic Gradient Boosting Tree?

Modelo	generativo	discriminativo	paramétrico	No paramétrico
regresión logística		x	x	
k-vecinos	x			x
Naive Bayes	x		x	
Funciones discriminantes gaussianas	x		x	
Modelo de mezcla gaussianas	x		x	
Ventana de Parzen	x			x
Redes Neuronales Artificiales		x	x	
Máquinas de vectores de soporte		x	x	

Tipos de funciones de activación

- ➢ https://missinglink.ai/guides/neural-network-concepts/7-types-neural-network-activation-functions-right/

Árboles de decisión

- ➢ https://bookdown.org/content/2031/arboles-de-decision-parte-i.html

Gradient Boosting y Stochastic

- ➢ http://docs.salford-systems.com/StochasticBoostingSS.pdf
- ➢ https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab

PREGUNTAS DEL 2 EXAMEN DE SIMULACIÓN

Me piden construir una aplicación para un dispositivo móvil que reconozca el texto escrito a mano usando un lápiz especial sobre la pantalla. La información con la que cuento es la posición en x y y del lápiz durante el tiempo en que la persona está escribiendo. ¿Cuál de los siguientes modelos es el más apropiado para resolver el problema?

Seleccione una:

- a. Una red tipo SOM (Mapa Autorganizable)
- b. Un Gradient Boosting Tree
- c. Una red neuronal MLP
- d. Una red neuronal recurrente
- e. Un modelo de mezclas de funciones Gaussianas

Respuesta: mapa autoorganizable es opción, recurrente modela sistemas en los cuales las salidas futuras dependen de salidas anteriores.

¿Cuál de las siguientes afirmaciones es falsa?

Seleccione una:

- a. El método de agrupamiento K-means se usa para resolver problemas no supervisados.
- b. La forma de determinar la convergencia del algoritmo k-means es evaluando si éste alcanzó el máximo número de iteraciones.
- c. El criterio de entrenamiento del método K-means se puede interpretar como la minimización de la dispersión intra-cluster.
- d. El método de agrupamiento K-means se puede considerarse una simplificación del algoritmo EM.
- e. El dendograma es un tipo particular de gráfico que permite visualizar el resultado de la aplicación de una técnica de clustering jerárquico.

Respuesta: Descartadas:E,A,D,B… respuesta la C

¿Cuál de las siguientes afirmaciones es falsa?

Seleccione una:

- a. La función softmax es una generalización de la función sigmoidal para más de dos clases.
- b. Las RNA tipo Perceptrón Multi-Capa (MLP) requieren el término independiente (peso w_0) en todas sus capas.
- c. La diferencia entre las RNA para resolver problemas de regresión y clasificación, está en la función de activación usada en la capa de salida.
- d. Para más de dos clases y una codificación one-hot encoding de la clase a predecir, es recomendable usar una función de activación sigmoidal.
- e. Los Perceptrones Multi-Capa (MLPs) son un tipo de RNA Feed Forward o de propagación hacia adelante.

¿Cuántos modelos de clasificación deben ser entrenados si se usa una estrategia Uno contra Uno (en inglés One vs One)?

¿Cuál de las siguientes afirmaciones es falsa?

Seleccione una:

- a. En un modelo de mezcla de funciones Gaussianas que usa matrices de covarianza diagonal, las componentes pueden formar elipsoides cuyo eje principal es paralelo a alguno de los ejes coordenados.
- b. Un modelo de clasificación basado en mezcla de funciones Gaussianas con matrices de covarianza completa tiene mayor capacidad de modelamiento que un modelo similar pero con matrices de covarianza esféricas.
- c. El algoritmo de Esperanza y Maximización corresponde a una implementación del criterio de Máxima Verosimilitud.
- d. En un problema de clasificación biclase en el que cada clase se modela a partir de una función de densidad Gaussiana con media dada por el conjunto de muestras de cada clase pero con varianza constante para todas las clases, la frontera de decisión entre las clases tiene la forma de una recta, es decir, es lineal.
- e. Un modelo de mezcla de funciones Gaussianas es un modelo no paramétrico que tiene la capacidad de modelar diferentes funciones arbitrarias de densidad de probabilidad.

Cuál de las siguientes definiciones corresponde a un ensamble de clasificadores tipo *voting*:

Seleccione una:

- a. Una combinación secuencial de modelos en la que cada nuevo modelos se enfoca en los errores del anterior.
- b. Una estrategia de combinación de modelos del mismo tipo, en la cada modelo se entrena sobre una muestra boostrap del conjunto de entrenamiento.
- c. Una combinación de diferentes árboles de decisión que seleccionan aleatoriamente un subconjnto de variables a evaluar en cada nodo.
- d. Una combinación de diferentes modelos entrenados sobre el conjunto de entrenamiento original.
- e. Una combinación de diferentes árboles de decisión que seleccionan aleatoriamente un subconjnto de variables a evaluar en cada nodo, y el mejor umbral de un subconjunto aleatorio de umbrales candidatos.

¿Cuál de las siguientes afirmaciones es falsa?

Seleccione una:

- a. Para un mismo número de épocas M, el esquema de entrenamiento *on-line* realiza un mayor número de actualizaciones de los pesos de una Red Neuronal Artificial que el esquema tipo *batch*.
- b. El algoritmo Backpropagation consiste en propagar el error medido en la capa de salida, a las capas ocultas de la red neuronal usando la regla de la cadena de la derivada, para ajustar los pesos de toda la red.
- c. El aprendizaje *on-line* difiere del aprendizaje tipo *batch* en la función de costo o criterio usado para el ajuste de los pesos de una Red Neuronal Artificial.
- d. El aprendizaje *on-line* puede usarse para realizar entrenamiento de una Red Neuronal Artificial que no ha sido entrenada previamente, y también para ajustar una Red que ya fue entrenada pero se requiere ajustar con muestras nuevas.
- e. El esquema de entrenamiento *on-line* es más susceptible a la tasa de aprendizaje que el esquema tipo *batch*.

¿Cuál de las siguientes afirmaciones es falsa?

Seleccione una:

- a. Un modelo de mezcla de funciones Gausianas es un modelo parámetrico que tiene la capacidad de modelar diferentes funciones arbitrarias de densidad de probabilidad.
- b. En un modelo de mezcla de funciones Gausianas que usa matrices de covarianza diagonal, las componentes pueden formar elipsoides cuyo eje principal es paralelo a alguno de los ejes coordenados.
- c. Un modelo de clasificación basado en mezcla de funciones Gausianas con matrices de covarianza completa tiene menor capacidad de modelamiento que un modelo similar pero con matrices de covarianza esféricas.
- d. En un problema de clasificación biclase en el que cada clase se modela a partir de una función de densidad Gausiana con media dada por el conjunto de muestras de cada clase pero con varianza constante para todas las clases, la frontera de decisión entre las clases tiene la forma de una recta, es decir, es lineal.
- e. El algoritmo de Esperanza y Maximización corresponde a una implementación del criterio de Máxima Verosimilitud.

¿Cuál de las siguientes afirmaciones es falsa?

Seleccione una:

- a. El aprendizaje *on-line* difiere del aprendizaje tipo *batch* en la función de costo o criterio usado para el ajuste de los pesos de una Red Neuronal Artificial.
- b. El aprendizaje *on-line* puede usarse para realizar entrenamiento de una Red Neuronal Artificial que no ha sido entrenada previamente, y también para ajustar una Red que ya fue entrenada pero se requiere ajustar con muestras nuevas.
- c. Para un mismo número de épocas M, el esquema de entrenamiento *on-line* realiza un mayor número de actualizaciones de los pesos de una Red Neuronal Artificial que el esquema tipo *batch*.
- d. El algoritmo Backpropagation consiste en propagar el error medido en la capa de salida, a las capas ocultas de la red neuronal usando la regla de la cadena de la derivada, para ajustar los pesos de toda la red.
- e. El esquema de entrenamiento *on-line* es más susceptible a la tasa de aprendizaje que el esquema tipo *batch*.

¿Cuál de las siguientes afirmaciones es falsa?

Seleccione una:

- a. En el método conocido como Random Forest el conjunto de árboles generados para clasificación no se podan.
- b. Un árbol de regresión siempre requiere de un límite de crecimiento o de la aplicación de un método de poda.
- c. La ganancia de información es una medida de la bondad de una partición del espacio de características usada en modelos de árboles de decisión para la solución de problemas de regresión.
- d. En un árbol de regresión, cada nodo terminal tiene asociados un solo valor de predicción.
- e. El índice gini se puede usar como medida de impureza de un espacio de características durante el entrenamiento de árboles de decisión.

¿Cuál de las siguientes afirmaciones es falsa?

Seleccione una:

- a. Para más de dos clases y una codificación one-hot encoding de la clase a predecir, es recomendable usar una función de activación sigmoidal.
- b. La diferencia entre las RNA para resolver problemas de regresión y clasificación, está en la función de activación usada en la capa de salida.
- c. La función softmax es una generalización de la función sigmoidal para más de dos clases.
- d. Las RNA tipo Perceptrón Multi-Capa (MLP) requieren el término independiente (peso w_0) en todas sus capas.
- e. Los Perceptrones Multi-Capa (MLPs) son un tipo de RNA Feed Forward o de propagación hacia adelante.

¿Cuál de las siguientes afirmaciones es falsa?

Seleccione una:

- a. En un modelo de mezcla de funciones Gausianas que usa matrices de covarianza diagonal, las componentes pueden formar elipsoides cuyo eje principal es paralelo a alguno de los ejes coordenados.
- b. El algoritmo de Esperanza y Maximización corresponde a una implementación del criterio de Máxima Verosimilitud.
- c. En un problema de clasificación biclase en el que cada clase se modela a partir de una función de densidad Gausiana con media dada por el conjunto de muestras de cada clase pero con varianza constante para todas las clases, la frontera de decisión entre las clases tiene la forma de parábola, es decir, es lineal.
- d. Un modelo de clasificación basado en mezcla de funciones Gausianas con matrices de covarianza completa tiene mayor capacidad de modelamiento que un modelo similar pero con matrices de covarianza esféricas.
- e. Un modelo de mezcla de funciones Gausianas es un modelo parámetrico que tiene la capacidad de modelar diferentes funciones arbitrarias de densidad de probabilidad.

¿Cuál de las siguientes afirmaciones es falsa?

Seleccione una:

- a. El índice gini se puede usar como medida de impureza de un espacio de características durante el entrenamiento de árboles de decisión.
- b. En el método conocido como Random Forest el conjunto de árboles generados para clasificación no se podan.
- c. Un árbol de regresión siempre requiere de un límite de crecimiento o de la aplicación de un método de poda.
- d. En un árbol de regresión, cada nodo terminal tiene asociados un solo valor de predicción.
- e. La ganancia de información es una medida de la bondad de una partición del espacio de características usada en modelos de árboles de decisión para la solución de problemas de regresión.

El algoritmo k-means se puede entender con un caso particular del algoritmo EM cuando se asume que la covarianza de todas las componentes es la misma y que todas las componentes tienen el mismo peso en el modelo.

Seleccione una:

- Verdadero
- Falso

Dé un ejemplo de aplicación donde la entrada corresponda a una secuencia de observaciones y la salida corresponda a una clase para cada una de las observaciones en la secuencia.

Por ejemplo con seguridad biométrica o seguridad por voz, cada muestra está representada por una secuencia de observaciones pero la salida corresponde a una clase como por ejemplo permitir o no permitir

¿Una red neuronal es un modelo paramétrico o no paramétrico? Justifique su respuesta.

Una RNA es un modelo parametrico ya que se hacen suposiciones sobre la forma del conjunto de datos.

¿Cuáles son los tres pasos del algoritmo de entrenamiento de un mapa auto-organizable (SOM)? Si no recuerda los nombres, describa qué se hace en cada paso.

Competición: Dada una muestra $\mathbf{x}$, en esta etapa se debe determinar la neurona ganadora.
Cooperación: Teniendo en cuenta la neurona ganadora, se define la vecindad de las neuronas excitadas, lo que permite generar el mapa topográfico.
Adaptación sináptica: Es el paso de adaptación, en el que se reajustan los pesos de la red, para favorecer las neuronas activadas.

Lo de mapas autoorganizable está en las notas que pase del 2 parcial, página 4

¿Cuál de las siguientes afirmaciones es falsa?

Seleccione una:

- a. El criterio de entrenamiento del método K-means se puede interpretar como la minimización de la dispersión intra-cluster.
- b. El método de agrupamiento K-means se puede considerarse una simplificación del algoritmo EM.
- c. El dendograma es un tipo particular de gráfico que permite visualizar el resultado de la aplicación de una técnica de clustering jerárquico.
- d. El método de agrupamiento K-means se usa para resolver problemas no supervisados.
- e. La forma de determinar la convergencia del algoritmo k-means es evaluando si éste alcanzó el máximo número de iteraciones.

¿Qué desventaja puede presentar la estrategia de clasificación Uno contra el resto (en inglés One vs All ó One vs the rest)

En one vs all se induce al desbalanceo

¿En qué consiste la estrategia de comité de máquinas conocida como Boosting ?

Boosting entrena modelos de manera secuencial de modo que los siguientes modelos se enfoquen en las muestras que no se han modelado correctamente, por lo tanto cada modelo tiene un peso diferente en la decisión final. Boosting busca reducir el sesgo en el error pero puede presentar problemas de sobreajuste.

El modelo GMM se puede utilizar para resolver problemas supervisados de clasificación y también para resolver problemas no supervisados de agrupamiento.

Seleccione una:

- Verdadero
- Falso

¿Cuál es el propósito de usar la estrategia de entrenamiento por Mini-batch en las Redes Neuronales Artificiales?

Mini-batch es un punto intermedio entre trabajo por lotes (Batch) y on-line, en donde se parten las muestras de entrenamiento en mini-batchs y se actualizan los pesos a partir del error que comete ese mini-batch de muestras

En un hospital desean desarrollar una aplicación que le permita a un médico, con base en la información disponible de un paciente (demográfica, clínica, etc.), encontrar un conjunto de pacientes similares dentro de la base de datos, para determinar el tratamiento más efectivo que le debe formular al paciente en evaluación, con base en el historial de los demás pacientes encontrados. Si la solución al problema se aborda usando técnicas de Machine Learning, ¿A qué tipo de problema de aprendizaje corresponde el problema planteado?

Corresponde a un problema de tipo no supervisado ya que se parte de un conjunto de características y se buscan patrones de comportamiento interesante, en este caso especifico se asocian las condiciones clínicas similares y con esto se puede determinar el tratamiento a prescribir

¿Qué diferencia existe entre un modelo Gradient Boosting Tree y un Stochastic Gradient Boosting Tree?

La diferencia es que con Stochastic Gradient Boosting Tree en cada iteración se hace un submuestreo sin reemplazo de los datos de entrenamiento de forma aleatoria