Cuando termine el curso, sería importante que sepan tener respuestas a las siguientes preguntas sobre los distintos temas abordados en el curso. Estas preguntas tienen dos objetivos.

1- Ayudar a redondear conceptos de cada uno de los temas.

2- Preparación de la parte de preguntas teóricas el día del examen.

Estas preguntas no forman parte del material que deben entregar. Lo que deben entregar son lo que les indico en el otro topic.

Iré actualizando esta lista en la medida que recorramos los próximos temas.

**ÁLGEBRA LINEAL**

1.     **¿Qué es una forma cuadrática?**

**2.     ¿Qué propiedades tienen los valores propios de una matriz simétrica?** Los valores propios de una matriz simétrica son todos reales, y los vectores propios correspondientes son ortogonales entre sí.

**3.     ¿Qué significa que una matriz simétrica sea definida positiva o semidefinida positiva? Indicar dos definiciones equivalentes entre sí.**

**4.     ¿Qué relación existe entre la traza de una matriz y sus valores propios? ¿Y entre el determinante de la matriz y sus valores propios?** La traza de una matriz es igual a la suma de sus valores propios. El determinante de una matriz es el producto de sus valores propios.

**5.     ¿Qué dice el teorema espectral? Dar dos resultados equivalentes entre sí.** El teorema espectral establece que cualquier matriz simétrica real puede diagonalizarse mediante una matriz ortogonal. Es decir, A=QΛQ^T , donde Q techo es una matriz ortogonal cuyas columnas son los vectores propios y Λ techo es una matriz diagonal con los valores propios.

**6.     ¿Qué dice el teorema de la esfera unidad?**

**7.     ¿Qué propiedades tienen las matrices de covarianzas?** Las matrices de covarianzas son simétricas y semidefinidas positivas, lo que significa que todos sus valores propios son no negativos. Además, la diagonal contiene las varianzas de las variables.

**ANÁLISIS EN COMPONENTES PRINCIPALES (PCA)**

1**.     ¿Cuál es el objetivo principal que aborda el análisis en componentes principales?** El objetivo es reducir la dimensionalidad de los datos, manteniendo la mayor cantidad posible de la variabilidad total. Esto se logra transformando las variables originales en un conjunto más pequeño de variables no correlacionadas llamadas **componentes principales**.

2.     **¿Cómo se hallan las componentes principales? ¿Qué problema resuelven?** Las componentes principales se obtienen calculando los valores y vectores propios de la matriz de covarianzas o la matriz de correlaciones de los datos. Resuelven el problema de maximizar la varianza explicada por las nuevas variables (componentes) con respecto a las originales.

**3.     ¿Qué problema trae aparejada la existencia de datos atípicos en el PCA?** Los datos atípicos pueden distorsionar los resultados del PCA, ya que las componentes principales están basadas en la varianza, y los atípicos pueden inflar la varianza de manera desproporcionada.

**4.     Si las variables tienen distintas unidades de medida o distintas magnitudes ¿qué problema puede aparecer en el PCA y cómo se puede hacer para resolverlo?** Si las variables tienen diferentes escalas, las que tienen mayores magnitudes dominarán las componentes principales. Para resolver este problema, se estandarizan las variables, es decir, se les da la misma escala convirtiendo sus medias en 0 y sus desviaciones estándar en 1.

**5.     ¿Qué desventaja puede tener estandarizar los datos antes de hacer un PCA?** Al estandarizar, se pierde la interpretación directa de la importancia de las variables originales, ya que todas las variables contribuyen por igual a la varianza total, independientemente de su escala original.

**6.     ¿Qué es un biplot y para qué sirve?** Un biplot es una representación gráfica que muestra simultáneamente tanto las observaciones como las variables en el espacio definido por las componentes principales. Permite interpretar las relaciones entre las observaciones y cómo las variables influyen en dichas observaciones.

**7.     ¿Cómo se interpretan en el biplot el lugar que ocupan las observaciones? ¿Y el ángulo entre los vectores? ¿Y             la longitud de los mismos?**

* El lugar de las observaciones en el biplot refleja sus relaciones con las componentes principales.
* El ángulo entre los vectores de las variables indica la correlación entre ellas: ángulos pequeños significan alta correlación positiva, ángulos rectos indican no correlación y ángulos grandes indican correlación negativa.
* La longitud de los vectores refleja la importancia de las variables en la varianza explicada por las componentes.

**8.     ¿Cómo se sabe cuándo un biplot es mucho o poco  informativo del conjunto de datos que tenemos?** Un biplot es más informativo cuando las primeras dos componentes principales explican un alto porcentaje de la varianza total de los datos. Si explican poco, el biplot será menos representativo del conjunto de datos completo.

**9.     Cuando vemos la salida de la  función prcom() en R ¿qué propiedades cumplen los valores que nos arroja la matriz de rotación**? Los valores en la matriz de rotación son los coeficientes que definen cómo las variables originales se combinan para formar las componentes principales. Esta matriz es ortogonal, lo que significa que sus columnas son ortonormales (vectores propios).

En Python, el equivalente a la función prcomp() de R, que se utiliza para realizar Análisis en Componentes Principales (PCA), es la clase PCA de la biblioteca scikit-learn. Esta clase proporciona una funcionalidad similar para realizar PCA y generar la matriz de rotación.

*# Obtener la matriz de componentes principales (rotación) rotacion = pca.components\_*

***¿Qué es la matriz de rotación (o components\_ en Python)?***

*La* ***matriz de rotación*** *(que en scikit-learn se llama components\_) es una matriz cuyas filas son los* ***vectores propios*** *de la matriz de covarianzas o correlaciones de los datos originales. Estos vectores propios definen las direcciones de las nuevas componentes principales.*

* *En el contexto de Python,* ***pca.components\_*** *es equivalente a la* ***matriz de rotación*** *que devuelve prcomp() en R.*
* *Las columnas de esta matriz representan cómo las* ***variables originales*** *se combinan linealmente para formar las nuevas componentes principales.*
* *Al igual que en R, esta matriz es* ***ortonormal****, es decir, sus columnas son vectores propios y son ortogonales entre sí.*

***Interpretación de la salida***

* ***pca.components\_****: Cada fila de esta matriz corresponde a un vector propio asociado a una de las componentes principales. Los coeficientes en estas filas nos indican las ponderaciones (o combinaciones lineales) de las variables originales que forman cada componente principal.*
* ***Ortonormalidad****: Igual que en prcomp(), las filas de la matriz en Python son ortogonales, lo que significa que las componentes principales son no correlacionadas entre sí. Además, las longitudes de estos vectores son 1 (norma 1), garantizando la ortonormalidad.*

***Propiedades importantes:***

1. ***Ortogonalidad****: Las componentes son ortogonales entre sí, lo que asegura que cada nueva componente captura una parte independiente de la variabilidad de los datos.*
2. ***Varianza explicada****: Se puede acceder a la varianza explicada por cada componente principal usando pca.explained\_variance\_ratio\_. (Esta salida indica qué porcentaje de la varianza total es capturada por cada componente principal, lo que también es una parte clave del análisis.)*

**CLUSTERING**

1. **¿Cuál es el objetivo general que persigue la técnica llamada clustering?** El objetivo del clustering es agrupar observaciones o datos en subconjuntos (clusters) de manera que las observaciones dentro de un mismo cluster sean similares entre sí y diferentes de las de otros clusters. Es una técnica de aprendizaje no supervisado utilizada para descubrir estructuras ocultas en los datos.
2. **¿En qué consiste la técnica de k-means? ¿Qué problema computacional tiene?** 
   * Definición: K-means es un método iterativo que particiona los datos en kkk clusters definidos por sus centroides. Busca minimizar la variación intracluster (distancia entre los puntos y el centroide del cluster).
   * Problema computacional: Puede quedar atrapado en mínimos locales debido a la dependencia del punto inicial. Además, el costo computacional crece con el número de puntos y clusters.
3. **¿Cómo funciona el algoritmo de separación en grupos de k-means? Describirlo.**

El algoritmo sigue estos pasos:

1. Inicializar K centroides aleatoriamente.
2. Asignar cada punto al cluster con el centroide más cercano (según alguna métrica, como la distancia euclidiana).
3. Recalcular los centroides como el promedio de los puntos asignados a cada cluster.
4. Repetir los pasos 2 y 3 hasta que las asignaciones no cambien o se alcance un número máximo de iteraciones.
5. **Si corro varias veces el programa que calcula las k-means ¿pueden dar grupos distintos? ¿Por qué?** Sí, pueden dar grupos distintos porque los centroides iniciales se seleccionan aleatoriamente, lo que puede llevar a diferentes soluciones debido a la convergencia hacia diferentes mínimos locales.
6. **Responder las preguntas 3 y 4 para la metodología PAM. ¿Qué ventajas y desventajas tiene respecto a kmeans?**

* PAM (Partitioning Around Medoids):
* Encuentra un medoid (punto real en el conjunto de datos) por cluster.
* Minimiza la suma de distancias absolutas en lugar de cuadráticas.
* Ventajas respecto a k-means: Más robusto ante outliers porque usa medoids en lugar de centroides.
* Desventajas: Más costoso computacionalmente, ya que implica evaluar todas las posibles asignaciones de medoids.
* Variabilidad: Si se ejecuta varias veces con inicializaciones diferentes, puede generar resultados distintos, aunque menos sensibles que k-means debido a su enfoque basado en medoids.

1. **¿Qué son los clusters jerárquicos?**

Son métodos de agrupamiento que crean una estructura jerárquica de clusters, representada mediante un dendrograma. Pueden ser:

* Aglomerativos: Comienzan con cada punto como un cluster individual y los combinan.
* Divisivos: Comienzan con un único cluster que se divide progresivamente.

1. **¿Cómo se realiza el método de agrupamiento (aglomerativo) que termina con el gráfico llamado dendrograma? Describirlo.**

* Cada observación comienza como un cluster individual.
* Calcular las distancias entre todos los clusters.
* Combinar los dos clusters más cercanos (según una métrica como la distancia mínima).
* Repetir hasta que todos los datos estén en un único cluster.
* Representar el proceso en un dendrograma.

1. **En un dendrograma ¿qué van en las abscisas y en las ordenadas? ¿En las abscisas hay unidad de medida?**

* Abscisas: Representan las observaciones o clusters. No tienen unidad de medida.
* Ordenadas: Representan la distancia o disimilitud a la que se fusionan los clusters. Sí tienen unidad de medida, que depende de la métrica utilizada.

1. **Definición de cada una de las posibles distancias que se pueden utilizar en el agrupamiento por dendrogramas (single, complete, average, centroid, Ward).**

* Single linkage: Distancia mínima entre puntos de dos clusters.
* Complete linkage: Distancia máxima entre puntos de dos clusters.
* Average linkage: Promedio de las distancias entre puntos de dos clusters.
* Centroid linkage: Distancia entre los centroides de dos clusters.
* Ward’s method: Minimiza la varianza intracluster al fusionar clusters.

1. **¿Cómo se estudia la correlación entre un dendrograma y las distancias (o disimilaridades) que tienen los datos originales?**

Se usa el cophenetic correlation coefficient, que mide la correlación entre las distancias cophenéticas (extraídas del dendrograma) y las distancias originales. Valores cercanos a 1 indican que el dendrograma representa bien las distancias originales.

1. **Si en un dendrograma hay dos individuos que están cercanos con distancias cercanas ¿eso implica que son parecidos?**

Sí, indica que tienen una baja disimilitud según la métrica utilizada, pero esta similitud depende del criterio de distancia seleccionado.

1. **Definición e interpretación de los índices de Silhouette y Dunn.**

* Índice de Silhouette: Mide qué tan bien se ajusta cada punto a su cluster. Valores cercanos a 1 indican buen ajuste; valores negativos sugieren mala asignación.
* Índice de Dunn: Relación entre la menor distancia intercluster y el mayor diámetro intracluster. Valores altos indican clusters bien separados y compactos.

1. **¿Qué son los heatmaps? Saber explicar cómo se definen, cómo se interpretan y para qué sirven.**

Son representaciones visuales de matrices, donde los valores se codifican por colores. En clustering, se usan para mostrar la similitud entre observaciones (e.g., matriz de distancias) junto con un dendrograma. Ayudan a identificar patrones visualmente.

1. **¿A qué se le llama estudio de la tendencia en clustering? Explicar en qué consiste.**

Es el análisis preliminar para determinar si los datos tienen una estructura intrínseca que permita su separación en clusters. Implica evaluar si existe una agrupación natural en los datos.

1. **¿Qué herramientas existen para saber si es razonable separar un conjunto de observaciones en clusters? En clase hablamos de 5 (en particular describir la idea que está detrás del estadístico de Hopkins).**

* Índice de Hopkins: Mide la aleatoriedad de los datos. Valores cercanos a 1 sugieren que los datos son agrupables.
  + Otras herramientas:
    - PCA
    - Silhouette
    - Heatmaps
    - VAT – Virtual Assessment of Cluster Tendency: herramienta visual. VAT reordena la matriz de distancia de datos de modo que los individuos cercanos quedan cerca , y luego pinta la matriz como un heatmap para visualizarlo de forma mas fácil.
    - Hopkins

1. **Definir los índices de Rand y de Jaccard para comparar el grados de similitud entre dos distintas particiones y explicar qué miden y cómo se interpretan.**

* Índice de Rand: Proporción de pares de puntos correctamente clasificados (ya sea en el mismo cluster o en diferentes clusters). Rango: 0 a 1.
* Índice de Jaccard: Proporción de pares de puntos correctamente clasificados en el mismo cluster respecto al total de pares clasificados juntos. Rango: 0 a 1.
* Ambos índices miden la similitud entre particiones, siendo el de Jaccard más estricto al ignorar puntos en diferentes clusters.

**CLASIFICACIÓN I (5 MÉTODOS CLÁSICOS)**

1. **Describir el problema de la clasificación binaria o no binaria.**

Clasificación binaria: Se trata de asignar una observación a una de dos clases posibles, por ejemplo, positivo/negativo o 0/1.

Clasificación no binaria (multiclase): Consiste en asignar una observación a una de varias clases posibles (más de dos), como por ejemplo, clasificar tipos de frutas (manzana, plátano, naranja).

El objetivo en ambos casos es encontrar un modelo que prediga correctamente la clase de nuevas observaciones basándose en un conjunto de datos de entrenamiento.

1. **¿Qué es la regla de decisión Bayes y qué propiedad tiene? ¿Qué problema tiene a la hora de intentar llevarla a la práctica?**

Regla de decisión Bayes: Clasifica una observación xxx en la clase CkC\_kCk​ que tiene la mayor probabilidad posterior, es decir:

A white background with black text

Description automatically generated

Propiedad: Es el clasificador óptimo que minimiza el error de clasificación teórico (Bayes error rate).

Problema práctico: No siempre es posible conocer las distribuciones P(x∣Ck) y P(Ck​), por lo que se necesitan estimaciones aproximadas.

1. **Describir el método de k vecinos más cercanos (knn) para clasificar. ¿Qué papel juega la elección del valor de k? Entender el significado de las figuras 2.15 y 2.16 del libro de James-Witten-Hastie-Tibshirani.**

Definición: Clasifica una observación x0x\_0x0​ según las clases de los kkk puntos más cercanos en el espacio de características (medido por una métrica como la distancia euclidiana).

Elección de kkk:

* Valores bajos (k=1,2k = 1, 2k=1,2): Más sensible al ruido.
* Valores altos (kkk grande): Tiende a generalizar más, lo que puede resultar en pérdida de detalle.

Figuras del libro (2.15 y 2.16): Muestran cómo kkk afecta el sesgo y la varianza:

* Fig. 2.15: Ejemplo de una frontera de decisión con diferentes valores de kkk.
* Fig. 2.16: Impacto en el error de clasificación según kkk.

1. **Describir el método de clasificación basado en la logística.**

La regresión logística modela la probabilidad de que una observación pertenezca a una clase como: A black and orange math equation

Description automatically generated with medium confidence

Luego, clasifica a la observación en la clase con la mayor probabilidad.

1. **¿Cómo se estiman los parámetros en la logística?** Se utilizan máxima verosimilitud, buscando los valores de β0,β1,… que maximizan la probabilidad de observar los datos dados los parámetros.
2. **Describir el método de análisis discriminante (LDA) para clasificar y el análisis discriminante cuadrático (QDA).**

LDA: Asume que las clases tienen distribuciones normales con la misma matriz de covarianza. Clasifica según la probabilidad posterior calculada con esta suposición.

QDA: Similar, pero permite que cada clase tenga su propia matriz de covarianza.

Diferencia: LDA tiene fronteras de decisión lineales; QDA, cuadráticas.

1. **¿Qué limitaciones tienen los métodos LDA y QDA? ¿En qué casos uno es mejor que el otro?**

Limitaciones:

* LDA: No captura relaciones no lineales debido a la suposición de covarianza común.
* QDA: Necesita más datos para estimar matrices de covarianza específicas por clase.

Mejor elección:

* LDA: Cuando las clases tienen fronteras aproximadamente lineales.
* QDA: Cuando las fronteras son no lineales y se cuenta con suficientes datos.

1. **Describir el método naive Bayes para clasificar.**

Supone que las características son independientes dentro de cada clase. Calcula:

**A black text on a white background

Description automatically generated**

Es simple y eficiente, aunque la suposición de independencia puede no ser realista.

1. **¿Cuáles de los métodos de clasificación vistos es paramétrico y cuáles son no paramétricos?**

Paramétricos: Suponen una forma funcional para el modelo, como logística, LDA y QDA.

No paramétricos: No hacen suposiciones sobre la distribución de los datos, como kNN.

1. **Saber entender e interpretar gráficos como las figuras 4.6, 4.9 del libro de James-Witten-Hastie-Tibshirani.**

Fig. 4.6: Muestra la relación entre el tamaño del árbol y el error de clasificación.

Fig. 4.9: Compara fronteras de decisión entre métodos como LDA y kNN. Ayuda a entender la capacidad de ajuste y generalización de cada método.

1. **Indicar al menos una ventaja y al menos una desventaja de cada uno de los métodos de clasificación.**

Logística:

* Ventaja: Fácil interpretación.
* Desventaja: No captura relaciones no lineales.

kNN:

* Ventaja: No hace suposiciones sobre los datos.
* Desventaja: Sensible a la escala y elección de

LDA/QDA:

* Ventaja: Rápidos y efectivos si las suposiciones se cumplen.
* Desventaja: Mal desempeño si las suposiciones no se cumplen.

Naive Bayes:

* Ventaja: Escalable.
* Desventaja: Suposición de independencia poco realista.

1. **Definir el training error test y el test error rate. ¿Cuál es el más importante y por qué? ¿Cómo se calculan?**

Definiciones:

1. Training Error Rate:
   * Es la proporción de errores (instancias clasificadas incorrectamente) en el conjunto de datos de entrenamiento.
   * Mide qué tan bien el modelo se ajusta a los datos sobre los que fue entrenado.
   * Fórmula: Training Error Rate=Numero de clasificaciones incorrectas en el entrenamiento/Total de observaciones en el entrenamiento ​
2. Test Error Rate:
   * Es la proporción de errores en un conjunto de datos nuevo (de prueba) que no se utilizó para entrenar el modelo.
   * Mide la capacidad de generalización del modelo a datos no vistos.
   * Fórmula: Test Error Rate=Numero de clasificaciones incorrectas en el conjunto de prueba/Total de observaciones en el conjunto de prueba

¿Cuál es más importante y por qué?

* El Test Error Rate es más importante porque:
  + Indica cómo se comportará el modelo en datos futuros, es decir, su capacidad de generalización.
  + Un modelo con un bajo Training Error Rate pero un alto Test Error Rate sufre de sobreajuste (overfitting), lo que significa que se ajusta demasiado a los datos de entrenamiento y falla en generalizar.
  + Un Test Error Rate bajo garantiza que el modelo sea útil en el mundo real.

¿Cómo se calculan?

1. Training Error Rate:
   * Se calcula utilizando el conjunto de datos de entrenamiento:
     + Entrenar el modelo con los datos de entrenamiento.
     + Evaluar cuántas observaciones del conjunto de entrenamiento se clasifican incorrectamente.
   * Porcentaje de errores en el conjunto de entrenamiento.
2. Test Error Rate:
   * Se calcula utilizando el conjunto de prueba (test set), que es independiente del conjunto de entrenamiento:
     + Aplicar el modelo ya entrenado a los datos de prueba.
     + Comparar las predicciones con las etiquetas reales del conjunto de prueba.
   * Porcentaje de errores en el conjunto de prueba.
3. **Definir la curva ROC ¿Cómo puede ser utilizada para comparar métodos de clasificación?** Representa la sensibilidad vs. 1-especificidad a diferentes umbrales. Permite comparar métodos midiendo el AUC (Área Bajo la Curva): valores más altos indican mejor rendimiento
4. **Describir el método de cross validation para obtener el valor de k en KNN.**

Divide los datos en K subconjuntos y evalúa el rendimiento del modelo para diferentes valores de K. Elige el K que minimiza el error promedio.

1. **¿En qué consiste el criterio de información de Akaike para la selección de modelos? ¿Cómo se puede utilizar en la logística? En particular vale la pena ver y entender la figura 5.7 del libro de James-Witten-Hastie-Tibshirani.**

Es un método para seleccionar modelos basado en el equilibrio entre ajuste y complejidad. **Definición del AIC (Criterio de Información de Akaike):**

El AIC es una métrica para comparar diferentes modelos estadísticos. Evalúa el equilibrio entre:

* **Ajuste al conjunto de datos:** Qué tan bien el modelo explica los datos.
* **Complejidad del modelo:** Penaliza modelos más complejos para evitar el sobreajuste.

Su formula es:

con p = nr parámetros estimados en el modelo

**Interpretación del AIC:**

* Valores más bajos de AIC indican un mejor modelo (mejor equilibrio entre ajuste y simplicidad).
* No mide directamente la calidad del modelo, sino qué modelo es **relativamente mejor** entre un conjunto de modelos candidatos.

**¿Cómo se usa en la selección de modelos de regresión logística?**

1. **Ajustar varios modelos:** Construir diferentes modelos logísticos con diferentes conjuntos de predictores.
2. **Calcular el AIC para cada modelo:** Para cada modelo ajustado, calcular el valor de AIC.
3. **Comparar los modelos:** Elegir el modelo con el AIC más bajo, ya que representa el mejor equilibrio entre ajuste y simplicidad.

**Ventajas del AIC:**

* Ayuda a evitar el **sobreajuste**, al penalizar la complejidad del modelo.
* Es aplicable a cualquier modelo que se pueda ajustar mediante verosimilitud, incluida la regresión logística.

**Limitaciones:**

* Solo compara modelos dentro del mismo conjunto de datos. No dice nada sobre la calidad absoluta del mejor modelo.
* Puede preferir modelos más simples si el tamaño de los datos es pequeño.

**Relación con la figura 5.7 del libro de James-Witten-Hastie-Tibshirani:**

* La figura ilustra cómo el AIC se usa para seleccionar predictores en un modelo logístico:
  + El eje X representa el número de predictores incluidos.
  + El eje Y muestra el valor de AIC.
  + El modelo óptimo es el que minimiza el AIC, mostrando cuántos predictores deberían incluirse para un buen equilibrio.

Por ejemplo, si el AIC comienza a subir al agregar más predictores, eso indica que los nuevos predictores no aportan suficiente información y solo están agregando complejidad innecesaria.

**CLASIFICACIÓN II (SVM, CART)**

1. **Describir en qué consiste el método de support vector machines para clasificar datos.**

SVM es un método de clasificación que busca encontrar el hiperplano óptimo que separa las clases de datos en el espacio de características con el mayor margen posible. El margen se define como la distancia más corta entre el hiperplano y las observaciones más cercanas de cada clase, conocidas como vectores soporte.

1. **Plantear matemáticamente el problema el problema de optimización que se resuelve para encontrar el hiperplano separador por svm, en ambos casos (cuando los datos están linealmente separados y cuando no lo están).**
2. **Explicar de dónde salen los vectores soporte que intervienen en la construcción del hiperplano de separación.**

Los vectores soporte son los puntos de datos más cercanos al hiperplano en cada clase. Determinan el margen máximo porque el hiperplano se define exclusivamente por estos puntos. Matemáticamente, son los puntos para los cuales:

A black and white math symbols

Description automatically generated

1. **¿Cómo se adapta el método de svm para clasificar cuando los grupos que conforman los datos no están linealmente separados?**

Cuando los datos no son linealmente separables:

1. Se proyectan los datos a un espacio de mayor dimensionalidad utilizando una función de transformación ϕ(x), de modo que sean linealmente separables en ese espacio.
2. Se usa un kernel (función de similitud) para calcular eficientemente ϕ(x), evitando trabajar explícitamente en el espacio transformado.

Ejemplos de kernels:

* Lineal: xi⋅xj ​
* Polinómico: (xi⋅xj+c)^d
* RBF (Radial Basis Function): exp⁡(−γ∥xi−xj∥^2)

1. **Enunciar y explicar el potencial que tiene en svm el teorema de Mercer.**

El teorema de Mercer garantiza que una función de kernel válida equivale a un producto escalar en algún espacio de características de alta dimensión. Esto permite:

* Trabajar implícitamente en ese espacio sin calcular explícitamente las transformaciones.
* Resolver problemas no lineales de manera eficiente y con buena generalización.

1. **Dar al menos una ventaja y al menos una desventaja de aplicar el método de svm**

Ventaja: Es robusto y efectivo para datos de alta dimensionalidad y problemas con bordes de decisión complejos.

Desventaja: Es computacionalmente costoso para conjuntos de datos grandes debido a la dependencia cuadrática o cúbica en el número de muestras.

1. **Describir el método de árboles de regresión y clasificación. Dar ventajas y desventajas del método.**

Definición: Construye un árbol binario donde cada nodo realiza una partición del espacio de características basándose en una regla de decisión que minimiza la impureza (para clasificación) o el error cuadrático (para regresión).

Ventajas:

* Fácil de interpretar y visualizar.
* Puede manejar datos categóricos y numéricos.

Desventajas:

* Sensible a los datos de entrenamiento (puede sobreajustarse).
* No generaliza bien en comparación con métodos más sofisticados.

1. **¿Cómo se obtienen las variables que aparecen en el encabezado de las distintas ramas y los valores que aparecen en las mismas? ¿Qué son los valores que aparecen en los nodos del árbol?**

Variables en las ramas: Se eligen mediante un criterio de selección, como:

* Impureza de Gini: G=1 − ∑ k . pk ^2
* Entropía: H = − ∑ k pk log (pk)
* MSE (regresión): Promedio del error cuadrático.

Valores en los nodos: Representan la clase mayoritaria (para clasificación) o el promedio de las etiquetas (para regresión).

1. **¿¿Qué significan los métodos de agregación en el tema predicción o regresión?**

La agregación consiste en combinar múltiples predicciones para mejorar la precisión y reducir la varianza. Ejemplos:

* Bagging (Bootstrap Aggregating): Entrena múltiples modelos en muestras bootstrap y promedia sus predicciones.
* Boosting: Combina modelos secuenciales, dando mayor peso a las observaciones mal clasificadas.
* Random Forest: Combina árboles de decisión generados de manera aleatoria.

1. **Explicar en qué consiste el método de Random Forest para clasificar ¿cuál es la ventaja respecto a CART? ¿Y la desventaja?**

Random Forest:

Es un método basado en la construcción de múltiples árboles de decisión (CART) en subconjuntos aleatorios del conjunto de datos y promediando sus predicciones (clasificación por mayoría de votos).

Ventajas sobre CART:

1. Menor varianza: Reduce el sobreajuste al combinar múltiples árboles.
2. Generalización: Mejora el rendimiento en datos no vistos.

Desventajas respecto a CART:

1. Menos interpretable: La combinación de muchos árboles hace que el modelo sea más difícil de entender.
2. Mayor costo computacional: Requiere entrenar múltiples árboles, lo que consume más tiempo y recursos.