Logotipo, nombre de la empresa

Descripción generada automáticamente

Universidad de Montevideo – Introducción a la Ciencia de Datos

Examen 2021 - S1

5 de Julio, 2021

1. **Explique el método de validación k-fold cross-validation haciendo un diagrama e indicando el propósito de cada dataset.**

Paso 1. De manera similar al holdout method, dividimos el conjunto de datos en dos partes, un conjunto de entrenamiento y un conjunto de prueba independiente; guardamos el conjunto de prueba para el paso de evaluación del modelo final al final.

Paso 2. En este segundo paso, ahora podemos experimentar con varios ajustes de hiperparámetros; podríamos utilizar la optimización bayesiana, la búsqueda aleatoria o la búsqueda en cuadrícula, por ejemplo. Para cada configuración de hiperparámetro, aplicamos el método de validación cruzada de k veces en el conjunto de entrenamiento, lo que da como resultado múltiples modelos y estimaciones de rendimiento.

Paso 3. Tomando los ajustes de hiperparámetros que produjeron los mejores resultados en el procedimiento de validación cruzada de k veces. Podemos usar el conjunto de entrenamiento completo para el ajuste del modelo con estos ajustes.

Paso 4. Ahora, usamos el conjunto de prueba independiente que retuvimos anteriormente (Paso 1); utilizamos este conjunto de pruebas para evaluar el modelo que obtuvimos del Paso 3.

Paso 5. Finalmente, después de que completamos la etapa de evaluación, opcionalmente podemos ajustar un modelo a todos los datos (conjuntos de datos de entrenamiento y prueba combinados), que podría ser el modelo para producción (deployment).

Hacer diagrama visto en clase.

1. **Usted entrena un modelo predictivo con un conjunto de training y evalúa la performance utilizando una métrica. Luego, evalúa la performance en un conjunto de testing y nota una gran diferencia de performance (mucho menor performance en testing en relación a training). Qué problema es este? Sugiera al menos una manera de resolverlo.**

El problema es overfitting ya que existe una gran diferencia de performance entre el conjunto de entrenamiento y el conjunto de prueba. Existen numerosas maneras de resolver dicho problema:

* Probar un proporción de split distinto o una seed distinta (poco probable).
* Cambiar de estrategia de validación.
* Entrenar menos epochs/iteraciones.
* Ajustar hiperparámetros más robustamente.
* Agregar regularización.
* Probar un algoritmo menos complejo o uno que ajuste mejor las hipótesis.
* Remover outliers.
* Seleccionar bien las features.
* Data augmentation/obtener más datos.
* Combinar varios modelos.

1. **Explique al menos 4 pre-procesamientos utilizados en NLP.**

Explicar 4 de los siguientes pre-procesamientos: tokenization, short/stop word removal, filtering, stemming, lemmatization, lower case, bag of words, etc.

1. **Explique al menos 3 métodos de feature engineering para datos tabulares.**

Ejemplos podrían ser One Hot Encoding, Ordinal encoding, binning, missing data imputation, remoción de outliers, normalización, estandarización, etc. Explicar, dar ejemplos.

1. **Corredores de distintos países corren 10 carreras y obtienen los tiempos que se muestran en la tabla del anexo. La desviación estándar para los corredores del país 1 es de 103 segundos mientras que para los corredores del país 2 es de 101 segundos. Estableciendo un nivel de significancia alpha = 0.05, existe una diferencia significativa en la media de tiempos de ambos corredores?**

**La tabla del anexo posee valores negativos del z-score porque es de la cola izquierda, pero es igual para los valores positivos del z-score si fuera la cola derecha.**

Acá en este ejemplo se hace un two sample z-test para medias. Hay que plantear la hipótesis nula y la hipótesis alternativa, usar la fórmula vista en clase y obtener el z-score. El valor del z-score da 0.2499 (se aproxima a 0.25), se usa la tabla y el p-value asociado es 0.4013. No se rechaza la hipótesis nula ya que para el nivel de significancia 0.05 el p-value es mayor.

1. **Comente los errores que existen en las gráficas presentes en el anexo.**

En la primer gráfica hay demasiadas categorías, lo que hace que sea imposible determinar el valor para cada estado. Asimismo, los colores son muy similares. Conviene utilizar un gráfico distinto. Podría también cambiarse a una escala distinta donde las diferencias se visualicen mejor. Hay dos estados cuyos valores exceden significativamente al resto. Podría utilizarse un gráfico de barras o incluso mejor un mapa geográfico donde se demuestre para cada estado el valor. Como es temporal puede utilizarse un filtro de tiempo en el gráfico, esto sería un gráfico dinámico.

El problema con el segundo gráfico es que las gráficas están superpuestas, es decir, los valores para el sexo femenino arrancan donde termina el sexo masculino. Habría que hacer una diferencia para obtener los valores para el sexo femenino. Un camino es hacer un gráfico superpuesto con transparencia o un histograma lado a lado. Otra opción sería hacer dos histogramas por separado.

El tercer gráfico tiene las fechas en desorden. Además el orden de las categorías varía fecha a fecha. Hay inconsistencias. El gráfico puede ser engañoso, ya que si no se mira con detenimiento el eje x uno puede pensar que los casos disminuyen en el tiempo. Se debe ordenar correctamente el gráfico por fecha y mantener un orden consistente entre categorías. Además se puede también agregar más ticks en el eje y ya que solo hay 3.

1. **En clase vimos grandes categorías de tareas relacionadas a Data Science y Machine Learning. Estas eran: supervised learning, unsupervised learning y reinforcement learning. Explique con sus palabras estas categorías de tareas.**

Definir las distintas tareas, explicar como es el aprendizaje en cada caso, como son los datos (hay ground truth?), que problemas se intentan resolver, etc.

1. **Explique 3 de los siguientes conceptos relacionados a una serie de tiempo: tendencia, estacionalidad, cambio cíclico y cambio irregular. Explique además como se reflejan dichos conceptos en la gráfica del anexo.**

La tendencia es un patrón en los datos que muestra el movimiento de una serie a valores relativamente más altos o más bajos durante un largo período de tiempo. En otras palabras, se observa una tendencia cuando hay una pendiente creciente o decreciente en la serie de tiempo. La tendencia suele pasar durante algún tiempo y luego desaparece, no se repite. Puede verse por medio de una regresión lineal en intervalos prolongados de tiempo. En la serie de tiempo del ejemplo se ve una tendencia creciente hasta 1980 aprox., luego estable.

La estacionalidad se manifiesta como variaciones repetitivas y periódicas en una serie de tiempo. En el ejemplo pueden verse en los incrementos y reducciones anuales.

Los cambios cíclicos son movimientos observados cada pocas unidades de tiempo, pero ocurren con menos frecuencia que las fluctuaciones estacionales. En el ejemplo, se pueden ver recesiones, una en 1975, otra en 1983 y otra en 1991.

El cambio irregular o ruido es una componente de error irreductible que es aleatorio y no exhibe dependencia sistemática del índice de tiempo. En el ejemplo, son cambios pequeños que hacen la gráfica más irregular haciendo que los patrones repetitivos no sean tan iguales sino que presenten pequeñas fluctuaciones a intervalos fijos de tiempo.

1. **Explique un problema que podría ser resuelto utilizando unsupervised learning, citando los datos que utilizaría y qué algoritmo.**

Como hemos visto en clase los problemas de unsupervised learning por lo general intentan encontrar patrones o grupos/cluster de datos con características similares entre los integrantes del grupo y lo más disímiles posibles a los integrantes de otros grupos. Otro problema a resolver, aunque menos común, es modelar paramétricamente (es decir, mediante una distribución estadística) un conjunto de datos.

Un ejemplo podría ser encontrar grupos/clusters de pacientes similares, usuarios con gustos parecidos, clientes con mismos patrones de compra, agrupación jerárquica de categorías, etc.

Los datos no poseen una verdad absoluta (y\_true/ground truth) sino que solo datos en X. Los datos usualmente deben ser reducidos a un número acotado de categorías para evitar el problema de la maldición de la dimensionalidad, requeriendo normalmente algún tipo de normalización o estandarización. En especial, si el algoritmo utiliza el concepto de distancia internamente (como kNN).

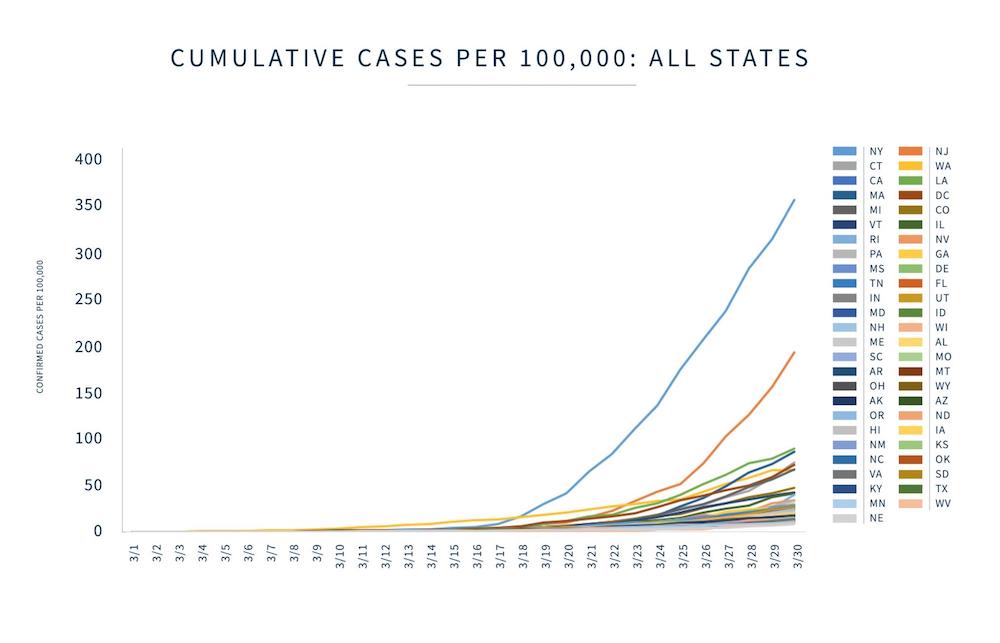
Debe describir algún tipo de algoritmo, como herarchical clustering, k-Means, DBSCAN, OPTICS, árboles de clasificación, etc.

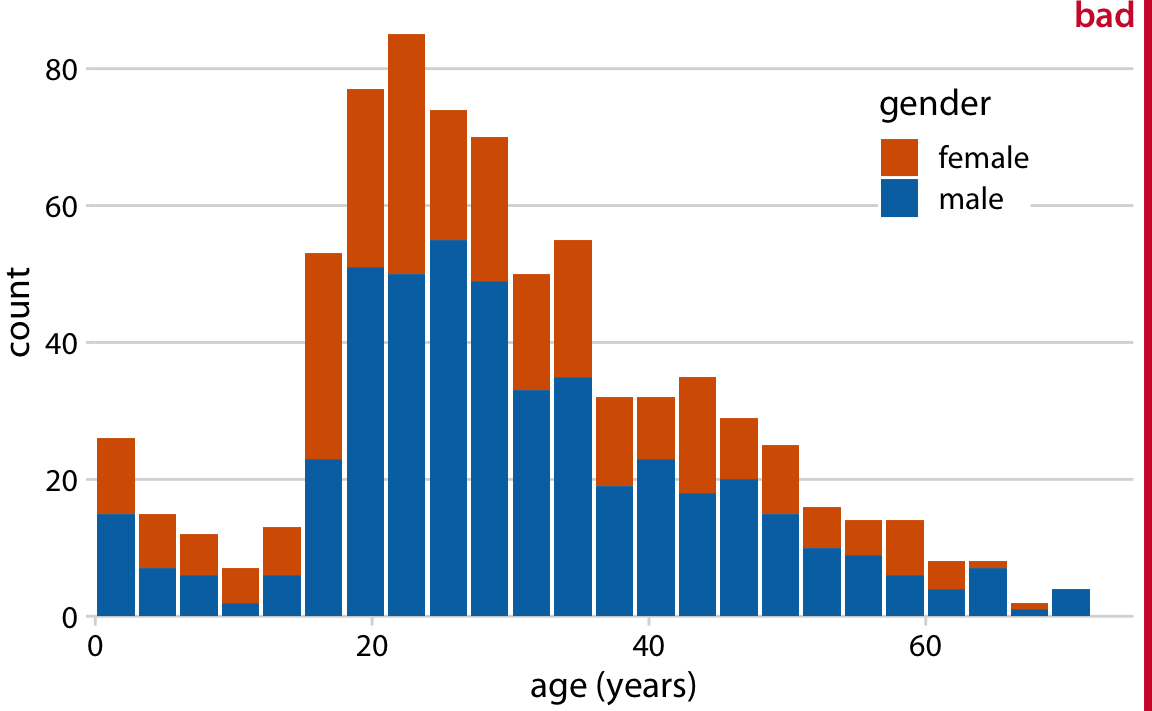
1. **Usted quiere usar una regresión lineal para generar predicciones de ventas. Qué hipótesis asume el modelo lineal? Cómo se interpreta el modelo lineal? Explique los conceptos de pendiente, intercepción y R2.**

Las hipótesis son:

* Linealidad: la relación entre X y la media de Y es lineal.
* Homoscedasticidad: la varianza del residual es la misma para cualquier valor de X.
* Independencia: las observaciones son independientes entre sí.
* Normalidad: para cualquier valor fijo de X, Y se distribuye normalmente.

La pendiente explica el incremento de la variable Y cuando se incrementa en una unidad la variable X. El intercept nos explica cual es el valor de Y cuando la variable X es cero. El coeficiente de determinación, denotado R2 o r2 y pronunciado "R cuadrado", es la proporción de la varianza en la variable dependiente que es predecible a partir de la variable independiente. Cuanto más cercano a 1 mejor se ajusta el modelo a los datos. Explicar la fórmula de R2, es decir numerador (error cuadrático respecto a la línea de mejor ajuste) y denominador (error cuadrático respecto a la línea de y\_promedio).

Problema 6:  
  




A screenshot of a computer

Description automatically generated with medium confidence

Problema 5:

Table

Description automatically generated

Table

Description automatically generated

Problema 8:

Chart, line chart

Description automatically generated