Logo, company name

Description automatically generated

**Introducción a la Ciencia de Datos**

Parcial 12 de Octubre 2022

1. **En clase hemos visto que el modelo de regresión logística es un modelo altamente interpretable. El mismo queda expresado por medio de cualquiera de las siguientes fórmulas (logit es la función inversa de la función logística):**

**Suponga que usted entrenó el modelo con 2 variables explicativas y obtuvo los coeficientes .**

1. **Interprete el modelo conociendo los coeficientes.**
2. **Interprete el modelo a partir de las posibilidades (odds):**

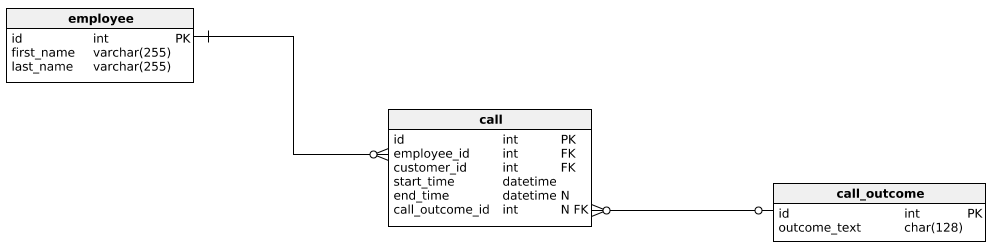
Esto se puede interpretar de la siguiente manera:

* es la intersección con el eje y. Son las probabilidades logarítmicas (logits) del evento en que **y=1**, cuando los predictores . Al exponenciar, podemos ver que cuando las probabilidades del evento en que **y=1** son 1 en 1000, 10^-3. De manera similar, la probabilidad del evento de que **y=1** cuando puede calcularse como =
* significa que aumentar en 1 unidad aumenta las probabilidades logarítmicas (logits) en 1. Entonces, si aumenta en 1 unidad, las probabilidades de que **y=1** aumentan en un factor de 10^1 = 10. Tenga en cuenta que la probabilidad de **y=1** también ha aumentado, pero no ha aumentado tanto como lo han hecho las posibilidades (odds).
* significa que aumentar en 1 unidad aumenta las probabilidades logarítmicas (logits) en 2. Entonces, si aumenta en 1 unidad, las posibilidades/odds de que **y=1** aumentan en un factor de 10^2 = 100. Observe cómo el efecto de en las probabilidades logarítmicas es dos veces mayor que el efecto de **,** pero el efecto sobre las posibilidades/odds es 10 veces mayor. Pero el efecto sobre la probabilidad (probabilidad NO es lo mismo que posibilidad/odds) de **y=1** no es 10 veces mayor, solo el efecto sobre las posibilidades/odds es 10 veces mayor.

1. **Suponga que usted tiene un dataset con 10 columnas y quiere encontrar clusters de datos. De las 10 columnas, 9 son continuas y 1 es categórica. Las columnas continuas se encuentran en rangos de valores muy distintos y la columna categórica posee una altísima cardinalidad, aunque la frecuencia de los valores es alta solo para 3 categorías. Explique el pre procesamiento que llevaría a cabo.**

El preprocesamiento se divide en dos partes. La primera es escalar/normalizar las columnas numéricas continuas a un rango común para que las columnas que poseen rangos de valores mayores no opaquen el efecto de las otras columnas. La segunda parte es cómo lidiar con la columna categórica. Debemos hacer un One Hot Encoding. Sin embargo, como posee una alta cardinalidad resultaría en la creación de numerosas columnas y tendríamos el problema de la maldición de la dimensionalidad. Como la letra dice que hay unas pocas categorías con valores predominantes (alta frecuencia), conviene hacer un encoding de las categorías con baja frecuencia en una sola categoría y luego aplicar el OHE.

1. **Suponga que usted trabaja con una base de datos de un call center. Usted sospecha que el empleado Victor Lustig ha mentido sobre su performance y quiere verificarlo usando la base de datos. Necesitamos todas las llamadas con su hora de inicio y hora de finalización. Para cada llamada, queremos mostrar cuál fue el resultado, así como el nombre y apellido del empleado que realizó esa llamada. Realice la query.**

****

La query require hacer un triple INNER JOIN hasta llegar a los datos del resultado de la llamada y filtrar por el empleado en cuestión:

SELECT employee.first\_name, employee.last\_name, call.start\_time, call.end\_time, call\_outcome.outcome\_text

FROM employee

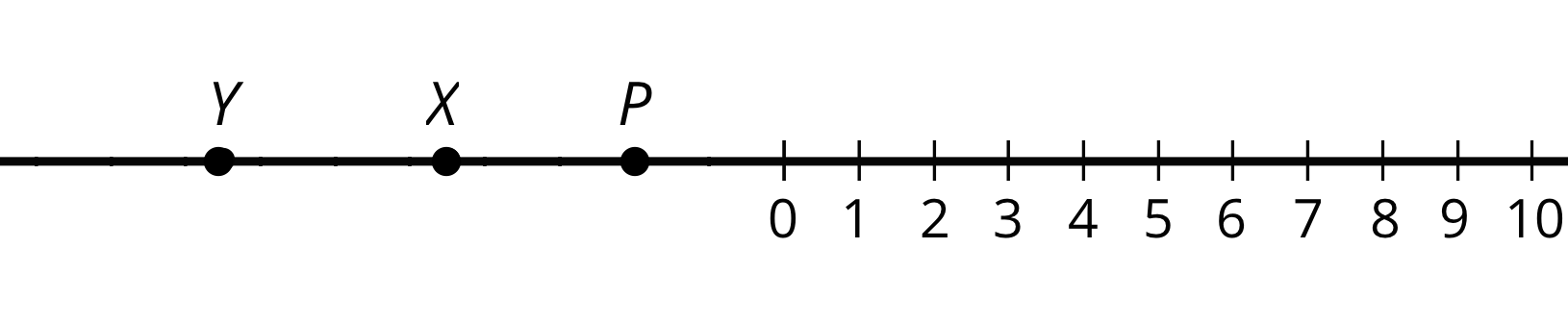
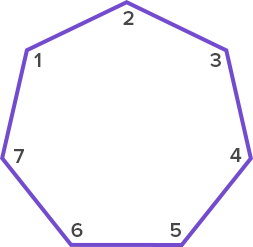
WHERE employee.first\_ name = ‘Victor’ AND employee.last\_ name = ‘Lustig’

INNER JOIN call ON call.employee\_id = employee.id

INNER JOIN call\_outcome ON call.call\_outcome\_id = call\_outcome.id

ORDER BY call.start\_time ASC;

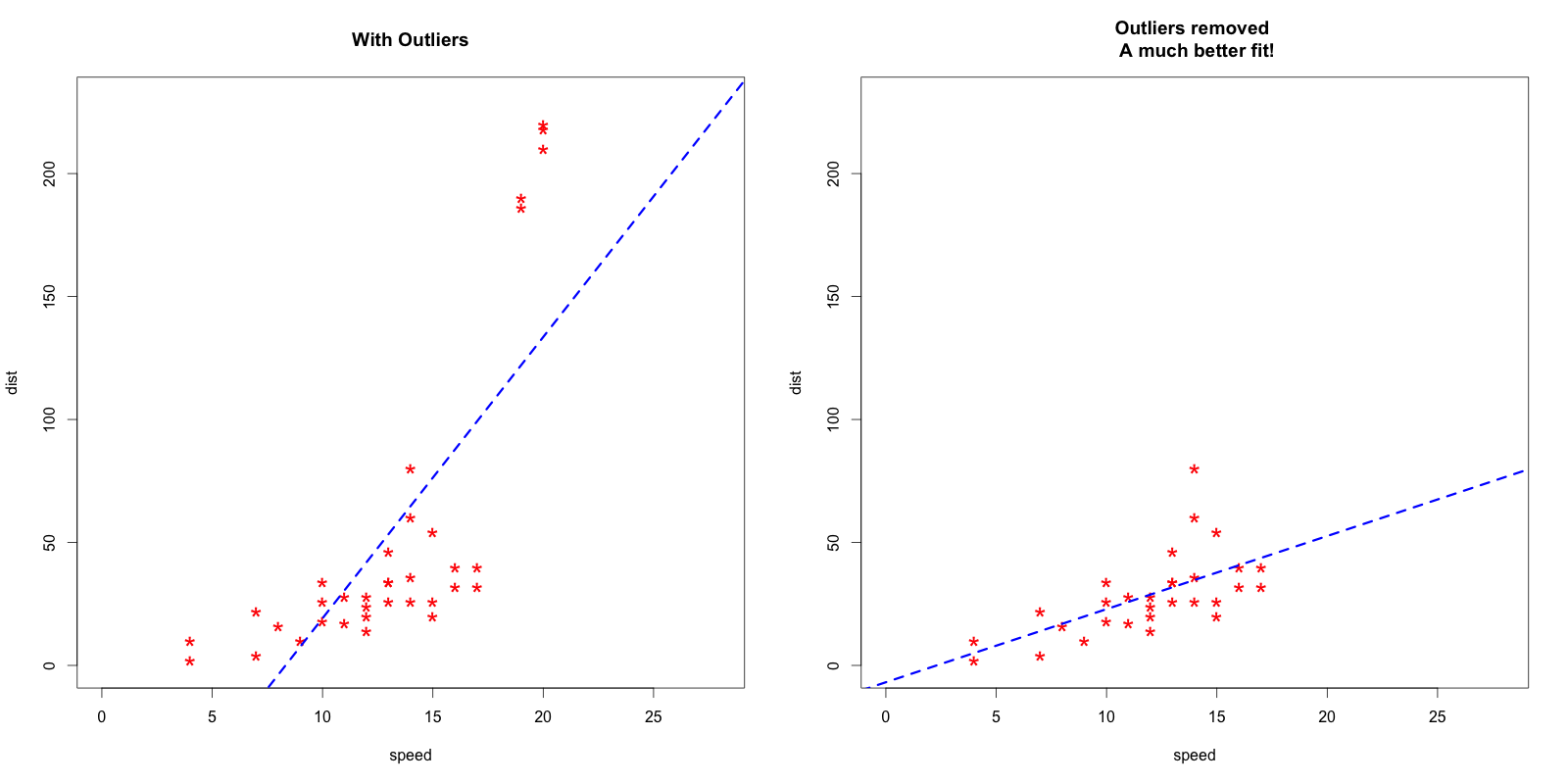
1. **Un feature engineering muy común para datos temporales es expresarlos mediante una función seno y coseno, es decir, “graficar” los datos en un círculo. ¿Por qué cree que esto es una buena idea en lugar de expresarlos numéricamente?**

** **

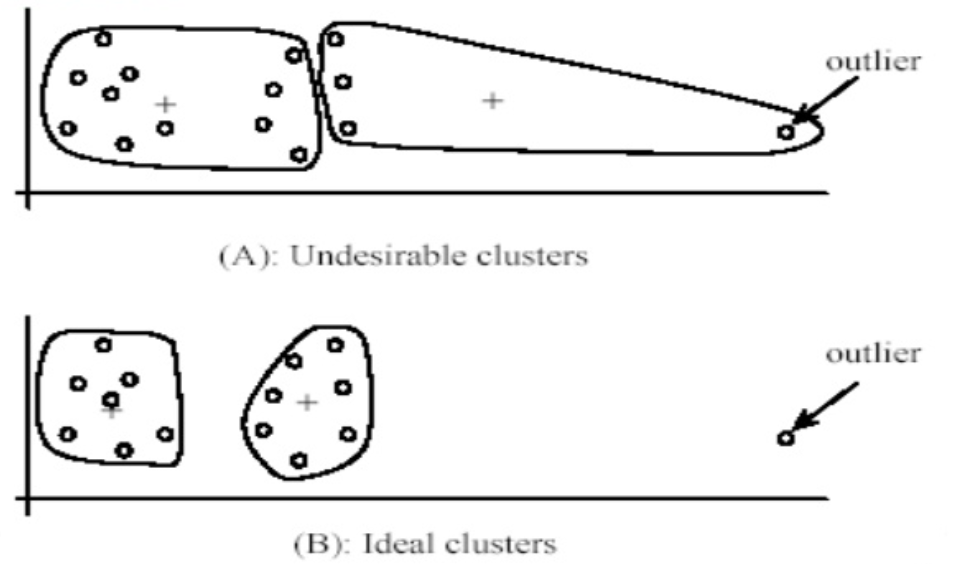
Los datos temporales son inherentemente cíclicos. Por eso conviene hacer este tipo de feature engineering. Por ejemplo, si no lo hiciésemos, el domingo y el lunes estarían separados por 6 días cuando en realidad la separación es de 1. En esencia, representar las variables temporales en un círculo por medio de funciones seno y coseno (funciones cíclicas), captura bien el concepto de distancia entre los valores y la periodicidad. Se utilizan seno y coseno (las dos en conjunto y no solo una) porque si se utiliza solo seno o solo coseno existen valores distintos en tiempo que reibirían el mismo valor trigonométrico.

1. **Qué problemas causan los outliers en el modelo de regresión lineal (supervised) y en k-Means (unsupervised)?**

En el caso de la regresión lineal los outliers distorsionan fuertemente la línea de mejor ajuste en la dirección del outlier (a menos que se diese la casualidad que existen dos outliers perfectamente simétricos respecto a la línea de mejor ajuste lo cual es muy improbable).



Respecto a k-Means, los valores atípicos pueden influir significativamente en la configuración final de los clústers y deben eliminarse para obtener soluciones de calidad. La posición de los centroides quedaría distorsionada en la dirección del outlier resultando en clusters no representativos o no homogéneos y de mala calidad. En el ejemplo de la imagen se puede apreciar como la posición de los centroides (símbolo +) queda distorsionada por la presencia del outlier.



**6) Una empresa internacional de jugos naturales sin azúcar quiere comenzar a comercializar sus productos en la cantina de la Universidad de Montevideo. Su estrategia de marketing se basa en información recolectada de diferentes fuentes que apunta a que los estudiantes universitarios consumen, en promedio, 3 litros de jugo natural a la semana.**

**Usted es recientemente contratado por esta empresa y duda si dicha premisa es correcta. Confeccione un experimento para comprobar dicha premisa definiendo**

**- Universo de estudio y muestra a realizar**

**- Proceso por el cual realizará el experimento**

**- Como comprobará estadísticamente la significancia de sus resultados**

La pregunta es abierta por lo que se evaluará cada propuesta estudiando su viabilidad y concordancia con el problema planteado.

Posible solución:

El univeso del estudio son todas aquellas personas concurrentes a la cantina de UM donde la empresa quiere comercializar sus productos.

Para corroborar la hipótesis se plantea siguente experimento: realizar una encuesta a los concurrentes a la cantina y preguntar a los encuestados la cantidad de litros de jugos consumidos por semana.

Para reducir los costos se realiza un muestreo, se seleccionan aleatoriamente 6 horas en la semana en la cual el encuestador se parará en la puerta de la cantina y realizará la encuesta a 1 de cada 5 personas que ingresen por la puerta. La muestra estará entonces comprendida por todos estos encuestados.

Con las encuestas obtendremos el valor promedio del consumo de jugo para la muestra y pondremos a prueba el promedio de la población (u) con un test estadístico.

Para esto aplicaremos un test estadístico Z con dos colas donde nuestra hipótesis nula es H0 u = 3 y la hipótesis alternativa es H1 u =! 3 con un nivel de significancia del 95%

**7) A continuación, se mencionan diferentes variables relativas a una persona. Determine la naturaleza de las siguientes variables en Nominales, Ordinales, Intervalo y Ratio**

**- Su nombre**

**- Su fecha de nacimiento en formato ISO 8601 (Ej: 2000 – 01– 15)**

**- Su edad**

**- Su género**

**- Su altura**

**- Su cédula de identidad**

**- La hora en más probable en la que se levanta los lunes**

**- La distancia que viaja hasta su trabajo**

**- La puntuación, del 1 al 5, que dio a la última película que vio.**

**- Su comida favorita**

* Nombre - Nominal
* Fecha de nacimiento - Ordinal
* Edad - Ratio
* Género - Nominal
* Altura - Ratio
* Cédula de identidad - Nominal
* Hora más probable que se levanta los lunes - Ordinal
* Distancia que viaja a su trabajo - Ratio
* La puntuación, del 1 al 5, que dio a la última película que vio - Ordinal
* Su comida favorita - Nominal

**8) Se accede a información de consumos de la cantina de facultad. Usted sospecha que existen dos perfiles distinguibles de clientes en la cantina por lo que aplica un algoritmo de aprendizaje no supervizado k-means inicializándolo con 2 clusters. El algoritmo inicializa con ubicando 2 centrodies: A y B en las siguientes posiciones: A (2,100) y B (14,800). Determine a que cluster se asigna cada cliente en la primer iteración del algoritmo.**

|  |  |  |
| --- | --- | --- |
| **Frecuencia mensual compras en cantina** | **Ticket Promedio** | **Cluster** |
| **15** | **500** |  |
| **6** | **200** |  |
| **4** | **240** |  |
| **8** | **0** |  |
| **12** | **1000** |  |

* (15, 500): Cluster B
* (6,200): Cluster A
* (4,240): ClusterA
* (8,0): Cluster B
* (12,1000): ClusterB

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| x | y | (x\_P-x\_A)^2 | (y\_P-y\_A)^2 | D(P,A) | (x\_P-x\_B)^2 | (y\_P-y\_B)^2 | D(P,B) | Cluster |
| 15 | 500 | 169 | 160000 | 400 | 1 | 90000 | 300 | B |
| 6 | 200 | 16 | 10000 | 100 | 64 | 360000 | 600 | A |
| 4 | 200 | 4 | 10000 | 100 | 100 | 360000 | 600 | A |
| 8 | 0 | 36 | 10000 | 100 | 36 | 640000 | 800 | A |
| 12 | 1000 | 100 | 810000 | 900 | 4 | 40000 | 200 | B |

**9) Que diferencias existen entre k-Means y DBSCAN? ¿En qué situaciones es preferible utilizar uno u otro?**

La primera gran diferencia está en el procedimiento. K-means considera inicialmente a todos los puntos equivalentes y luego, utilizando distancia euclediana, va asignando pertenenciasa centroides provisorios hasta que estos convergen. Por otro lado, DBSCAN categoriza los puntos según su densidad en Core, Border y Noise.

Esta diferencia de procedimiento genera los siguientes efectos:

- El algoritmo de DBSCAN asegura convergencia mientras que K-means no.

-  El algoritmo de DBSCAN es mucho más rápido que K-means, esto tiene principal relevancia cuando se aplican a grandes cantidades de datos

-  El algoritmo de DBSCAN es parcial dado que algunos puntos no son asignados a clusters sino definidos como ruido o outliers, mientras que el K-means asigna completamente todos los puntos a un cluster

-  K-means requiere conocimientos previos de la naturaleza de los clusters, inicializar el algoritmo con un número errado de clusters puede llevar a clusters de un único punto o clusters a los que no se les asigna ningún punto.

- K-means inicializa de forma estocástica (la posición de los centroides en la primera iteración es al azar) por lo que pueden azarosamente darse los problemas del punto anterior. Los resultados de DBSCAN no son sensibles al azar.

-  K-means es fuertemente sensible a la presencia de outliers, pudiendo estos generar no convergencia o resultados extraños, por lo que es necesario filtrar los datos previo a su utilización. DBSCAN asigna a los outliers como ruido y no afecta la posición de los clusters resultantes.

**10) Cuales son los componentes necesarios para una buena visualización? Justifique.**

Una buena visualización tiene 4 componentes fundamentales.

·        Tiene que mostrar datos de una forma confiable y honesta. Los datos pueden ser reales o sintéticos si es que estos son debidamente aclarados y cumplen con un fin común al presentador y a la audiencia.

·        Debe tener un diseño interesante, con el uso de colores y fuentes que, además de contribuir a la correcta interpretación de la visualización, agraden desde el punto de vista estético.

·        Tiene que estar enmarcada dentro de una historia y servir al desarrollo de la misma. Las visualizaciones sin contexto o que no permitan empujar la historia hacia una conclusión carecen de sentido.

·        Tiene que poder ser compartida por la audiencia con otros actores de interés por lo que debe de ser entendible en ausencia del presentador. Esto se logra mediante la correcta inclusión del contexto en la visualización ya sea por el correcto uso de títulos, etiquetas, ejes, etc.