|  |  |  |
| --- | --- | --- |
| **Nombre**: Ruben Alejandro Deambrossi | | **Matrícula**: 3012348 |
| **Nombre del curso:**  Aprendizaje Automático | **Nombre del profesor**:  Rolando Evelio Pérez Versón | |
| **Módulo**: 1 | **Actividad**: 4 | |
| **Fecha**: 18-Jun-2022 | | |
| **Bibliografía**:  Roldan, P. (s.f.). Modelo de regresion. *economipedia*. Recuperado de <https://economipedia.com/definiciones/modelo-de-regresion.html>  *Comprende Principal Component Analysis* (2018). Aprende Machine Learning. Recuperado de <https://www.aprendemachinelearning.com/comprende-principal-component-analysis/>  Dubey, A. (2018). The Mathematics Behind Principal Component Analysis. *Towards Datascience*. Recuperado de <https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643>  Brownlee, J. (2019). A Gentle Introduction to Probability Density Estimation. *Machine Learning Mastery*. Recuperado de <https://machinelearningmastery.com/probability-density-estimation/>  (s.d.). Histogramas de Areas – Curvas de densidad. *UBA - Exactas*. Recuperado de <http://www.dm.uba.ar/materias/analisis_de_datos/2006/1/teoricas/Teor2b.pdf>  *Principales Algoritmos usados en Machine Learning* (2017). Aprende Machine Learning. Recuperado de <https://www.aprendemachinelearning.com/principales-algoritmos-usados-en-machine-learning/>  (s.d.). El algoritmo k-means aplicado a clasificación y procesamiento de imágenes. *Universidad de Oviedo*. Recuperado de <https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html> | | |

**Título**: Optimización y selección de modelos de aprendizaje automático.

**Introducción**:

El siguiente reporte corresponde a la actividad 4 de Aprendizaje Automático.

**Desarrollo**:

1. Mínimo valor estimado para el cual la función converge.

Notebook Link:

1. Optimización.

Notebook Link:

1. Modelos matemáticos.

* Regresión: Es un modelo matemático que determina la relación entre una variable dependiente (x) y otras variables llamadas independiente (x).

La forma más simple de un modelo de regresión es:

**Y = A + BX + u**

En donde:

Y es la variable dependiente/endógena.

X es la variable independiente/explicativa.

A, B son parámetros fijos y desconocidos.

u = es un término de error, su función es englobar todos los demás parámetros que afectan Y pero no fueron incluidos en el modelo.

El objetivo de un modelo de regresión es estimar los valores de A y B. (Roldan, P., s.f.).

* Reducción de dimensionalidad: Es un método utilizado para extraer características más importantes de un set de datos. Es muy utilizado para reducir las dimensiones tanto de un data set como de imágenes.

Según Dubey, A. (2018), existen varios modelos para reducir la dimensionalidad, siendo uno de los más utilizados el PCA (Principal Component Analysis).

Para aplicar PCA se deben seguir los siguientes pasos:

* Se calcula la media de cada dimensión de la matriz del dataset.
* Se calcula la matriz de covarianza de todo el dataset. El resultado es otra matriz

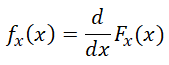


* Calcular los autovectores y autovalores de la matriz de covarianza
* Ordenar los autovectores en base a los autovalores de manera descendente y escoger los autovectores asociados con los autovalores más grandes. Aquellos autovectores con los autovalores con menor valor son aquellos que corresponden a dar menor información sobre la distribución de los datos, por lo tanto, son aquellos que se pueden descartar.
* Con los autovalores ordenados en orden descendente, se crea una nueva matriz donde se combinan los autovectores y autovalores.
* Con la matriz calculada en el paso anterior se aplica una transformación a la matriz con los datos que se quiere reducir su dimensión en base a la siguiente ecuación y = W’ \* X, donde W’ es la transpuesta de la matriz con los autovectores y autovalores.
* La estimación de la densidad: Es la relación entre la salida de una variable random y su probabilidad.

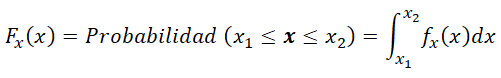
De acuerdo con Brownlee, J. (2019), Si la variable random es continua, entonces su probabilidad puede ser calculada por la probabilidad de densidad (PDF).

La función de densidad de probabilidad de la variable aleatoria se podría estimar contando el número de veces a lo largo del intervalo de observación en que la señal toma cada valor de la amplitud.

Matemáticamente PDF:



Donde Fx(x) es la probabilidad acumulativa de que una variable aleatoria x no supere un valor particular de la misma x.



La probabilidad de que la variable aleatoria caiga en una región especifica del espacio de posibilidades está dada por la integral de la variable entre 1 y otro limite.

* Clasificación: un algoritmo de clasificación clasifica automáticamente objetos o datos con el objetivo de poder hacer predicciones precisas a partir de la observación con datos previos.

Uno de los algoritmos de clasificación no supervisada más utilizado es el k-means, el cual agrupa objetos en k grupos basándose en sus características. Este agrupamiento se consigue minimizando la suma de la distancia entre cada objeto y el centroide de su grupo o cluster.

Este algoritmo consta de tres pasos:

* Inicialización: después de escoger el número de grupos, k, se establecen k centroides en el espacio de manera aleatoria.
* Asignación objetos a los centroides: cada objeto es asignado a su centroide más cercano.
* Actualización: se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos de dicho grupo.

Los pasos 2 y 3 se repiten hasta que los centroides no se mueven (El algoritmo k-means aplicado a clasificación y procesamiento de imágenes, s.f.)

1. Articulo.

No pude darle seguimiento al articulo con auto AI, ya que tuve problemas con mi cuenta de IBM.

Como conclusión puedo decir que el articulo se me hizo muy completo y que abarca diferentes elementos para tener en cuenta al momento de hacer un análisis de este tipo. Hay diferentes terminologías y técnicas que aun desconozco pero parecen ser muy útiles y fundamentales al momento de preprocesar los datos.

Se me hizo muy interesante la manera de comparar y evaluar los 4 modelos utilizados, sobre todo la parte de evaluación de los resultados, ya que es un elemento fundamental para elegir cual algoritmo es el más adecuado.

Me sorprendió como el modelo de regresión logística fue el modelo que mejor se adapta a la solución y como también es necesario evaluar el contexto, ya que si bien este modelo presentaba un alto numero de falsos positivos, de todas maneras es mejor bloquear transacciones “sospechosas” que dejar pasar transacciones fraudulentas.