



**UNIVERSIDAD SIMÓN BOLÍVAR
DECANATO DE ESTUDIOS PROFESIONALES
COORDINACIÓN DE MATEMÁTICAS**

**CONTRIBUCIÓN DE DISTINTOS TIPOS DE MATERIALES DE APRENDIZAJE EN EL
CONOCIMIENTO ADQUIRIDO EN EL CURSO CO6612**

Por:
Leonardo Jose Martin Amendolia

PROYECTO DE GRADO
Presentado ante la Ilustre Universidad Simón Bolívar
como requisito parcial para optar al título de
Licenciado en Matemáticas, opción Estadística y Matemáticas Computacionales

Sartenejas, abril del 2022



**UNIVERSIDAD SIMÓN BOLÍVAR
DECANATO DE ESTUDIOS PROFESIONALES
COORDINACIÓN DE MATEMÁTICAS**

**CONTRIBUCIÓN DE DISTINTOS TIPOS DE MATERIALES DE APRENDIZAJE EN EL
CONOCIMIENTO ADQUIRIDO EN EL CURSO CO6612**

Por:

Leonardo Jose Martin Amendolia

Realizado con la asesoría de:

Prof. Minaya Villasana

PROYECTO DE GRADO

Presentado ante la Ilustre Universidad Simón Bolívar
como requisito parcial para optar al título de
Licenciado en Matemáticas, opción Estadística y Matemáticas Computacionales

Sartenejas, abril del 2022

UNIVERSIDAD SIMÓN BOLÍVAR
DECANATO DE ESTUDIOS PROFESIONALES
COORDINACIÓN DE MATEMÁTICAS

**CONTRIBUCIÓN DE DISTINTOS TIPOS DE MATERIALES DE APRENDIZAJE EN EL
CONOCIMIENTO ADQUIRIDO EN EL CURSO CO6612**

PROYECTO DE GRADO

Por: Leonardo Jose Martin Amendolia

Realizado con la asesoría de: Prof. Minaya Villasana

RESUMEN

El siguiente trabajo propone aplicar un modelo matemático que prediga el nivel de conocimiento de los estudiantes, los agrupe según sus habilidades de aprendizaje y analice su comportamiento, a lo largo del tiempo, cuando interactúan con diversos tipos de materiales educativos: evaluados y no evaluados. Dicho modelo es aplicado por medio de dos casos distintos y resuelve un problema de optimización en el que minimiza una función de costo o de pérdida por medio del algoritmo del descenso del gradiente estocástico, a través de estrategias de entrenamiento basadas en el aprendizaje automatizado. El modelo es aplicado a datos del curso CO6612 de la Universidad Simón Bolívar. Se concluye que los estudiantes del curso fueron clasificados en dos grupos de aprendices obteniendo una mayor adquisición de conocimiento por parte de los recursos Quiz Evaluativo PM y Quiz Evaluativo Proyecto Final seguidos por cada una de las tareas asignadas. Además, según sus habilidades de aprendizaje, fueron caracterizados en 2 grupos de aprendices; el primer grupo, consiguió adaptarse y asimilar los conceptos de manera más lenta, que los estudiantes del segundo grupo en el cual su adquisición de conocimiento fue más rápida.

Palabras clave: modelar, función de costo, descenso del gradiente.

Para todas las personas que no conocen esta hermosa carrera.

AGRADECIMIENTOS

Le agradezco al Universo este hermoso camino lleno de sabiduría y aprendizaje.

Mil gracias a mi familia (Mariela, Jose, Fabricio y Claudia) por estar conmigo, no solo en esta oportunidad, sino en cada paso de mi vida personal, laboral y estudiantil, por ser quienes son y lo que son.

Le doy las gracias a cada uno de los profesores que entregó más de lo que debía, que me ha dado las herramientas y, sobre todo, la motivación y ejemplo para llegar hasta el final. Gracias a mis compañeros de estudio. Y gracias a la increíble Universidad Simón Bolívar por ser el centro del conocimiento.

ÍNDICE GENERAL

ÍNDICE DE TABLAS	ix
ÍNDICE DE FIGURAS	x
1 PLANTEAMIENTO DEL PROBLEMA	1
1.1 Planteamiento del problema	1
1.2 Justificación	2
1.3 Objetivos	3
1.3.1 Objetivo general	3
1.3.2 Objetivos específicos	3
2 MARCO REFENCIAL	4
2.1 Aprendizaje automatizado	4
2.1.1 Funciones de costo o de pérdida	5
2.1.2 Sobre-estimación y regularización	6
2.1.3 Descenso del gradiente	6
2.1.3.1 Limitaciones del descenso del gradiente	9
2.2 Estrategia de entrenamiento	9
2.3 Medidas de precisión	11
2.4 Explicación del modelo MVKM	11
2.4.1 Definición de las variables y modelo	14
2.4.2 Construcción de la función de costo	19
2.4.3 Descenso del gradiente	21
2.4.3.1 Derivadas parciales de \mathcal{L}_1	21
3 METODOLOGÍA	25
3.1 Datos del curso CO6612	25
3.2 Transformación de los datos	27
3.3 Elección de conjuntos para entrenar y probar	28
3.4 Proceso de entrenamiento	29
4 EXPERIMENTACIÓN Y DISCUSIÓN	31
4.1 Calibración de hiperparámetros	31
4.2 Aporte de cada recurso al nivel de conocimiento de los estudiantes	34
4.3 Correspondencia entre los recursos y los conceptos	35
4.4 Análisis de los estudiantes	37
4.5 Construcción de los grupos según las habilidades de aprendizajes de cada estudiante	39

4.6	Análisis de cada grupo de estudiantes en la adquisición de conocimiento	43
4.7	Análisis del conocimiento adquirido por los estudiantes para cada concepto	45
4.8	Análisis del conocimiento de los estudiantes para cada concepto	46
CONCLUSIONES		48
REFERENCIAS		49

ÍNDICE DE TABLAS

3.2	Datos transformados del curso CO6612.	28
3.1	Número de identificación de cada recurso.	29
4.1	Posibles valores de los hiperparámetros para la búsqueda en cuadrícula.	32
4.2	Valores óptimos de los hiperparámetros para cada caso.	33
4.3	RMSE y MAE para cada caso.	33
4.4	Agrupaciones de los estudiantes.	43

ÍNDICE DE FIGURAS

2.1	Pasos de actualización del descenso del gradiente sobre la función f .	7
2.2	Elección de distintos valores para α .	8
2.3	Conjuntos para entrenar y probar	10
2.4	Factorización matriz tridimensional $X^{[r]}$.	13
2.5	Matriz Tridimensional $X^{[r]}$	16
2.6	Matriz $X_a^{[r]}$	16
2.7	Vector $x_{s,a}^{[r]}$	17
2.8	Escalar $x_{s,a,p}^{[r]}$	17
2.9	Gráfica función σ	18
4.1	Aporte al conocimiento por parte de cada recurso, caso 1.	34
4.2	Aporte al conocimiento por parte de cada recurso, caso 2.	35
4.3	Peso relativo de cada concepto en cada recurso, caso 1.	36
4.4	Peso relativo de cada concepto en cada recurso, caso 2.	37
4.5	Adquisición de conocimiento de los estudiantes, caso 1.	38
4.6	Adquisición de conocimiento de los estudiantes, caso 2.	38
4.7	Visualización de los datos de la matriz S , caso 1.	40
4.8	Visualización de los datos de la matriz S , caso 2.	40
4.9	Método del codo para la asignación óptima de los grupos, caso 1.	41
4.10	Método del codo para la asignación óptima de los grupos, caso 2.	41
4.11	Visualización de lo grupos con $k = 2$ según el algoritmo k -medias, caso 1.	42
4.12	Visualización de lo grupos con $k = 2$ según el algoritmo k -medias, caso 2.	42
4.13	Conocimiento adquirido por los estudiantes en cada intento, caso 1.	44
4.14	Conocimiento adquirido por los estudiantes en cada intento, caso 2.	45
4.15	Conocimiento para cada concepto en cada intento, caso 1.	46
4.16	Conocimiento para cada concepto en cada intento, caso 2.	46
4.17	Conocimiento para cada concepto en cada intento por cada grupo, caso 2.	47

CAPÍTULO 1

PLANTEAMIENTO DEL PROBLEMA

El auténtico pensar, como el vivir, es tarea
que se aprende. Quien aprende, debe
dialogar con aquellos que han pensado
previamente: recorrer sus itinerarios,
seguir sus rutas y estelas, consultar sus
bitácoras

Ernesto Mayz Vallenilla
Rector fundador de la USB

En este capítulo, se describe el problema objeto de esta investigación, justificación y sus objetivos.

1.1. Planteamiento del problema

Según los investigadores Zhao, S., Wang, C., y Sahebi, S., [1]

Las investigaciones actuales se han centrado en modelar un solo tipo de recurso de aprendizaje a la vez ignorando la heterogeneidad de los recursos de aprendizajes disponibles para los estudiantes. Los sistemas modernos de aprendizaje en línea ofrecen con frecuencia a los estudiantes aprender y evaluar sus conocimientos utilizando varios tipos de recursos de aprendizaje, como lecturas, conferencias, tareas, cuestionarios y debates.

En la última década, se ha registrado un crecimiento de inscripciones en cursos en línea. En 2012, solo en Estados Unidos, el 32.5 % de los estudiantes de educación superior estaban inscritos en algún curso dictado a través de tecnologías digitales y, gracias a la llegada de la pandemia por el COVID 19, se aceleró y se generó la necesidad de que los estudiantes e institutos educativos se trasladarán al uso de las plataformas digitales de modo que durante los años 2020 y 2021 el 49% de la población estudiantil en todo el mundo afirmó haberse inscrito en un curso en línea.

Así pues, es claro que los cursos dictados hoy día, de manera presencial, y sobre todo de forma virtual, van acompañados de distintos tipos de materiales de aprendizaje que pueden ser agrupados en dos grandes categorías: materiales evaluativos –tareas, proyectos, asignaciones, exámenes, entre otros- y materiales no evaluativos –videos, lecturas, entre otros-.

Cuando se busca evaluar el conocimiento del estudiante se apelan a los recursos evaluativos para medir el conocimiento adquirido. La razón es simple, provee un mecanismo para cuantificar el aprendizaje. No obstante, se sabe que este aprendizaje es producto de la interacción con todos

los tipos de materiales; pero que su efecto o cuantificación en el proceso del estudiante permanece oculto hasta llegar a una evaluación.

Es por lo anterior que nace la curiosidad de estudiar algún modelo matemático que posibilite predecir el conocimiento de los estudiantes que interactúan con distintos tipos de materiales de aprendizaje y cuantificar el aporte de esos materiales no evaluados al proceso de aprendizaje del estudiante. Esto permitiría responder a preguntas comunes en el área de la educación como, por ejemplo, ¿existe algún modelo que nos permita predecir el conocimiento de un estudiante en algún momento dado aunque este no haya sido evaluado? ¿qué nos puede decir ese modelo sobre los tipos de estudiantes? ¿cuántos estilos de aprendices hay en un curso? ¿es posible agrupar y caracterizar a los estudiantes del curso? ¿cómo afectan o apoyan los recursos educativos en los aprendices? entre otras preguntas.

1.2. Justificación

Actualmente, no se conocen investigaciones, y por ende implementaciones de modelos matemáticos, que permitan predecir el conocimiento de los estudiantes que interactúan con diversos tipos de materiales educativos y que propicien un entendimiento entre las habilidades de los estudiantes y los conceptos dados en el curso.

Por lo anterior, se desea aplicar un modelo que pueda hacer lo mencionado anteriormente; ya que las clases dictadas a través de las plataformas digitales, y usando distintos tipos de materiales de aprendizaje, han mostrado un incremento año tras año en los estudiantes que en ellos participan.

La intención de este proyecto es aplicar el modelo denominado *Multi-View Knowledge Modeling* (MVKM), publicado en el artículo *Modeling Knowledge Acquisition from Multiple Learning Resource Types*, en 2020, y elaborado por Zhao, S., Wang, C., y Sahebi, S., de la Universidad de Albany, Nueva York, Estados Unidos.

Se experimentó con el modelo MVKM con datos reales, específicamente, con los datos del curso Redes Neuronales, código CO6612, de la Universidad Simón Bolívar de los trimestres enero-marzo y abril-julio del 2021, dictado por la profesora Minaya Villasana.

El enfoque del presente trabajo es práctico; lo cual resuelve una dificultad actual en la modelización del conocimiento de los estudiantes, a través de un problema de optimización que emplea técnicas de aprendizaje automatizado para su resolución. Este trabajo permite asentar las bases a futuras investigaciones a nivel educativo, en las que se desee entender la relación entre los cursos y sus estudiantes. La implementación de esta herramienta en combinación con diversas investigaciones, estudios y/o planteamientos pueden ser poderosas estrategias para el pleno entendimiento del

rendimiento de los estudiantes en los cursos educativos, para una comprensión en cuanto a la presentación y asimilación de los conceptos por parte de los estudiantes, al aporte de cada uno de los recursos educativos, entre otras cosas; todo ello, con la finalidad y el objetivo de tomar las mejores decisiones, que afecten de forma positiva al avance educativo de los estudiantes.

1.3. Objetivos

1.3.1. Objetivo general

Aplicar el modelo MVKM al curso de CO6612 que contiene dos grandes categorías de recursos, para predecir el conocimiento adquirido por los estudiantes durante el mismo más allá de la calificación final.

1.3.2. Objetivos específicos

- Modelar el conocimiento de los estudiantes en diversos puntos de tiempo
- Agrupar a los estudiantes según sus habilidades de aprendizaje
- Caracterizar a los grupos de estudiantes
- Analizar el comportamiento de los conceptos a lo largo del tiempo
- Observar el efecto de los conceptos sobre cada grupo de estudiantes
- Cuantificar el aporte al conocimiento de cada uno de los recursos del tipo evaluativo
- Comparar la aplicación del modelo mediante dos métodos distintos

CAPÍTULO 2

MARCO REFENCIAL

Simplicidad es un prerequisite de
confiabilidad

Edsger W. Dijkstra

Este capítulo tiene la finalidad de establecer las bases teóricas detrás de la aplicación e implementación del modelo MVKM. En la sección 1.1 se presentan las definiciones del aprendizaje automatizado y los conceptos relacionados a ese tema como las funciones de costo, el descenso del gradiente, la regularización, las técnicas de entrenamiento de un modelo y las medidas de precisión. En la sección 1.2 se definen las ecuaciones del modelo MVKM, los hiperparámetros, la función de costo del modelo y la aplicación del descenso de gradiente.

2.1. Aprendizaje automatizado

Desde que las computadoras fueron inventadas el ser humano siempre se ha preguntado si ellas serían capaces de aprender por sí mismas, tal como pensaba Mitchell, T., [2] “imaginemos computadoras aprendiendo de registros médicos para saber cuál tratamiento es el más efectivo, casas aprendiendo de la experiencia para optimizar costos energéticos basados en el uso de los ocupantes”.

Hoy día, gracias a los avances en distintas disciplinas, como la estadística y las probabilidades, la computación, la teoría de la información, la psicología, la filosofía, entre otras, somos testigos de estos hechos. Las computadoras están aprendiendo a realizar tareas sin necesidad de ser programadas explícitamente por un humano; por lo que, se coincide con la definición de Michie, D. y Spiegelhalter, D. J y Taylor, C. C.,[3], y se define el aprendizaje automatizado o *Machine Learning* como “procedimientos informáticos automáticos basados en operaciones lógicas o binarias que aprenden una tarea a partir de una serie de ejemplos”.

Ahora, al imaginarse el uso del aprendizaje automático en el área de la educación: ¿sería posible, utilizando datos de estudiantes en cursos académicos, encontrar modelos que permitan clasificar a los estudiantes según sus habilidades e inferir sus niveles de conocimiento adquirido?

¿Se podría usar la información de esa categorización para relacionar los conceptos o conocer la importancia de cada uno de los recursos en el aprendizaje del estudiante?, ¿es posible modelar el fenómeno de olvido de conceptos por parte de los estudiantes, entre otros fenómenos?

Hoy día, el aprendizaje automatizado, junto con todas las disciplinas involucradas, se ha convertido en una herramienta útil para realizar estudios, investigaciones y tomar las mejores decisiones en áreas tan diversas como la educación.

2.1.1. Funciones de costo o de pérdida

Al intentar ajustar los parámetros de un modelo a un conjunto de datos, es importante contar con una función que cuantifique el error cometido. Así, para cada conjunto de parámetros del modelo, se puede probar su idoneidad. Esta función de error naturalmente depende de los datos usados, del modelo a ajustar y de los parámetros de ese modelo.

Los parámetros que mejor ajusten el modelo a los datos dados son aquellos que producen el mínimo error posible entre todas las combinaciones de parámetros existentes. Dicha función es lo que se denomina función de costo. De modo que, la función de costo se utiliza como medida para comprender la relación entre los datos observados y las predicciones; en otras palabras, calcula la distancia existente entre la predicción y la observación.

Existen diversas funciones de pérdida usadas en distintos algoritmos. A continuación, se enlistan dos de las más usadas:

- **Función de pérdida del error cuadrático:** Esta función mide la diferencia cuadrática entre los valores reales y los predichos.

Sea x el valor real, \hat{x} el valor predicho y $n \in \mathbb{N}$, el número total de datos, entonces, el error cuadrático medio (EC) se define como sigue:

$$EC = \sum_{i=1}^n (x_i - \hat{x}_i)^2 \quad (2.1)$$

- **Función de pérdida del error absoluto (EA):** Calcula la suma de las diferencias absolutas entre las variables reales y predichas.

$$EA = \sum_{i=1}^n |x_i - \hat{x}_i| \quad (2.2)$$

2.1.2. Sobre-estimación y regularización

En la búsqueda del menor error posible, es factible que el modelo sea muy complejo o con muchos parámetros libres que haga que se ajuste en demasía a los datos, que pudiesen contener errores, y perder de vista la naturaleza real de la función (modelo) que se busca, incurriendo en el problema de la sobre-estimación.

Brownlee, J., [4], indica que “El sobre-ajuste ocurre cuando un modelo aprende los detalles y el ruido en los datos de entrenamiento en la medida en que afecta negativamente el rendimiento del modelo en datos nuevos”. Entonces, una de las maneras en que ocurre la sobre-estimación es cuando el modelo aprende muy bien la data de entrenamiento y se ajusta de manera casi perfecta a ella. El problema es que los datos de entrenamiento contienen errores (de medición por ejemplo), entonces ajustará también esos errores, perdiendo la relación fundamental entrada-salida. Lo esperado es tener un error mínimo en la data de entrenamiento y de validación. Idealmente se prueba la capacidad de generalizar del modelo encontrado, observando el error cometido en datos que no fueron usados para el proceso de entrenamiento, a este conjunto se le conoce como datos de prueba.

Otra manera en que podemos caer en modelos sobre-estimados es mediante la construcción de modelos muy complejos. A la función de costo se le agrega un término de complejidad o regularización escalada por una constante, llamada constante de regularización. Gareth, J., Daniela, W., Trevor, H., y Robert, T., [5], mencionan que “El término de regularización tiene el efecto de reducir las estimaciones de los parámetros hacia cero”; dicho de otra manera, disminuye la complejidad del modelo.

Por su parte, la constante de regularización será un hiperparámetro que determina cuan severa será la penalización que imprime un cierto nivel de simplicidad al modelo. Su función es compensar entre un buen ajuste del modelo y su simplicidad, evitando así la sobreestimación.

2.1.3. Descenso del gradiente

Una vez que se ha planteado la función de costo o de pérdida, se debe utilizar algún algoritmo que estime numéricamente dónde dicha función genera sus valores más bajos. Esta estimación se logra actualizando los valores de los parámetros del modelo, de tal modo que se minimice el error de la función de pérdida. Existen diversos algoritmos de optimización como el método de Cauchy, el descenso del gradiente, el gradiente conjugado, el método de Newton, entre otros. Este trabajo se enfocará en el descenso del gradiente.

Según, Brownlee, J., [4], “El descenso de gradiente es un algoritmo de optimización utilizado para encontrar los valores de los parámetros (coeficientes) de una función (f) que minimiza una

función de costo.” En virtud de ello, el descenso de gradiente es un método iterativo usado para minimizar alguna función de pérdida que produce una secuencia que va convergiendo a los parámetros óptimos del modelo, tomando el camino opuesto a la dirección del gradiente (opuesto a la dirección de ascenso más pronunciada) de la función de error; es decir, en dirección negativa del gradiente.

En otras palabras, sea $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ una función que define un paraboloide elíptico y sea $\alpha \in \mathbb{R}$ constante. Sea x_0 un punto de la función y suponga que se desplaza una distancia positiva α en la dirección de descenso entonces el siguiente punto x_1 será calculado por:

$$x_1 = x_0 - \alpha \cdot \nabla f(x_0) \quad (2.3)$$

De forma general,

$$x_{n+1} = x_n - \alpha \cdot \nabla f(x_n) \quad (2.4)$$

Por consiguiente, a partir de un primer punto, se va iterando, utilizando la ecuación anterior, hasta encontrar un mínimo local. En la figura 2.1 se visualiza el paso de actualización del descenso del gradiente sobre un mapa de contorno de la función f , arrancando en un punto x_0 hasta llegar al mínimo.

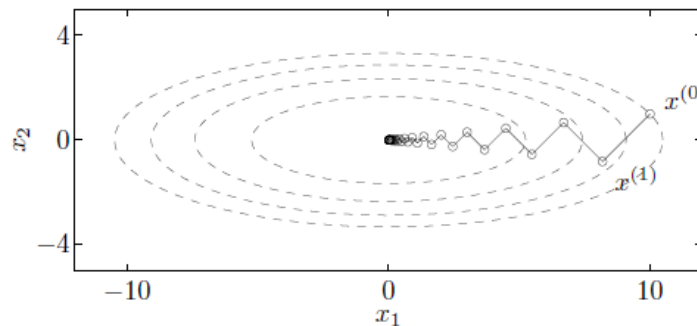


Figura 2.1: Pasos de actualización del descenso del gradiente sobre la función f .

Imágen tomada del libro *Optimización Convexa*, [6].

La constante α de la ecuación 2.4 se conoce como tasa de aprendizaje. La tasa de aprendizaje

es el tamaño del paso para moverse en la dirección de descenso. Al fijar este valor se evita la búsqueda lineal en la dirección del descenso. Es un parámetro sensible. Seleccionar un valor muy alto para la tasa de aprendizaje puede causar que el algoritmo no converja; ya que podría estar llegando a espacios donde la función no se minimiza; mientras que una tasa de aprendizaje muy pequeña consumirá bastantes recursos computacionales y más tiempo de lo esperado. La figura 2.2 muestra, sobre una función cualquiera de costo, cómo los distintos valores de la tasa de aprendizaje influyen en la elección del siguiente punto.

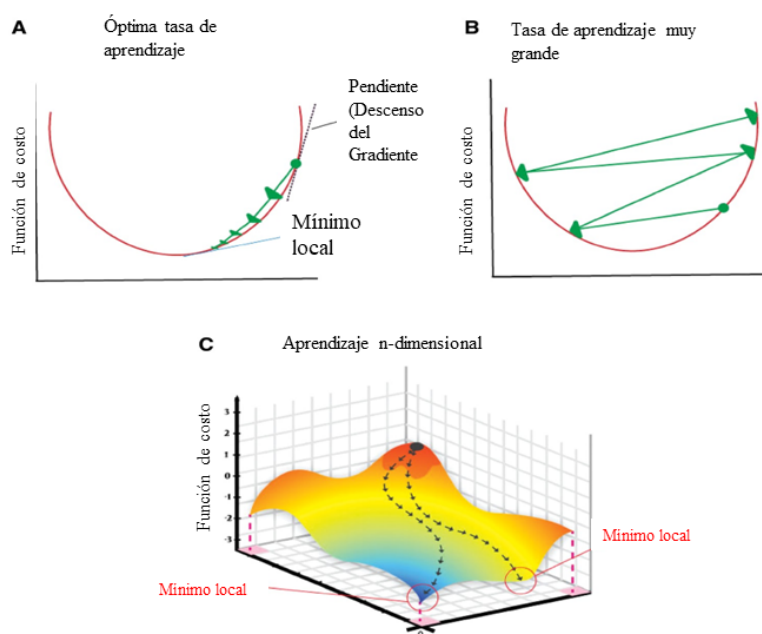


Figura 2.2: Elección de distintos valores para α .

Imagen tomada del artículo *Deep learning for cardiovascular medicine: a practical primer*, [7].

Un aspecto a considerar del descenso del gradiente es que el paso de actualización se realiza cuando se presentan todos los datos al modelo, esto se conoce como presentación por lotes.

Según Haykin, S., [8], algunas ventajas del descenso del gradiente son que realiza una estimación precisa garantizando, bajo condiciones simples, la convergencia del método, además de ser económico computacionalmente ya que paraleliza el proceso de aprendizaje. Sin embargo, como el descenso del gradiente ejecuta una presentación por lotes de los datos al modelo esto ocasiona que exista una mayor probabilidad de quedar atrapado en un mínimo local y requiere de gran capacidad

de almacenamiento por parte de la máquina.

Para combatir dichos problemas, Haykin, S., [8], recomienda realizar el paso de actualización del descenso del gradiente cada vez que se le presenta un dato al modelo. Esta técnica, denominada descenso del gradiente estocástico, reduce la posibilidad de quedar atrapado en un mínimo local y requiere de menor capacidad de almacenamiento.

2.1.3.1. Limitaciones del descenso del gradiente

Algunas limitaciones del descenso de gradiente son las siguientes:

- El descenso del gradiente solo encuentra mínimos locales; por ende, si una función tiene más de un mínimo, no es garantizado que sea el mínimo global.
- La elección de la tasa de aprendizaje es de real importancia; ya que de su elección depende una rápida o lenta convergencia al valor mínimo o que nunca se logre converger.

2.2. Estrategia de entrenamiento

Después de que se ha definido la función de costo y el algoritmo de optimización, es momento de encontrar los mejores parámetros que se ajusten a los datos disponibles; y, por consiguiente, tener la mayor precisión posible. Para ello, se requiere de una estrategia en la presentación de los datos al modelo.

Una manera es dividir los datos en al menos dos conjuntos, un conjunto para entrenar y otro conjunto para probar los ajustes encontrados; dado que al construir un modelo, se debe comprobar sobre el conjunto de datos que no ha sido usado previamente, ya que el interés está en el error cometido por estos datos de prueba. Por ejemplo, si se construyó un modelo para predecir si un paciente sufrirá o no de diabetes, basados en registros médicos de otros pacientes (datos de entrenamiento), el objetivo es predecir si un paciente nuevo, es decir, un dato que no ha sido visto por el modelo, sufrirá o no de diabetes, dado el conocimiento adquirido.

La estrategia en la que se presentan los datos, explicada en el inciso anterior, es conceptualmente la más sencilla, y consiste en tomar de forma aleatoria dos conjuntos de datos; uno para entrenar y otro para probar los parámetros encontrados en el entrenamiento. En primer lugar, se ajustan los parámetros del modelo con los datos de entrenamiento y, luego, se usan para realizar predicciones en el conjunto de validación y calcular los errores cometidos por la función de costo. Pero, este enfoque trae consigo algunas problemáticas, tal como lo indica Gareth, J., Daniela, W., Trevor, H., y Robert, T., [5]:

El enfoque del conjunto de validación es conceptualmente simple y fácil de implementar, pero tiene dos inconvenientes potenciales. El primero, es que la estimación de la tasa de error para el conjunto de validación puede ser muy variable, dependiendo precisamente de qué observaciones se incluyen en el conjunto de entrenamiento y qué observaciones se incluyen en el conjunto de validación. El segundo es que, en el enfoque de validación, solo un subconjunto de las observaciones, las que se incluyen en el conjunto de entrenamiento, en lugar del conjunto de validación, se utilizan para ajustar el modelo. Dado que los métodos estadísticos tienden a funcionar peor cuando se entrenan con menos observaciones, esto sugiere que la tasa de error del conjunto de validación puede tender a sobrestimar la tasa de error de prueba para el ajuste del modelo en todo el conjunto de datos.

Debido a las problemáticas anteriores, es posible implementar otra estrategia que consiste en escoger dos conjuntos de datos: un conjunto para entrenar- el cual se particiona en dos subconjuntos de datos: datos de entrenamiento y datos de validación- y el segundo conjunto para probar el modelo. La figura 2.3 explica dichas particiones.

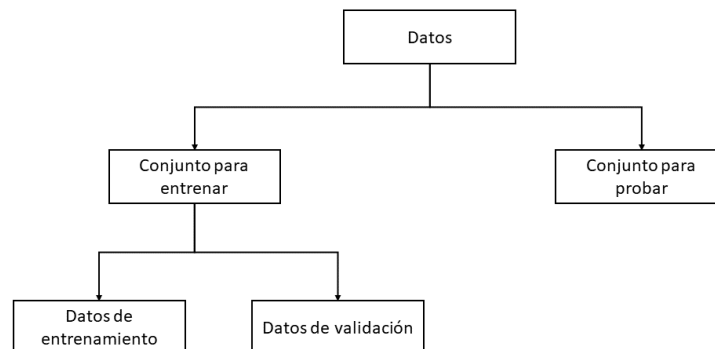


Figura 2.3: Conjuntos para entrenar y probar

La división en conjunto para entrenar y de probar se realiza antes de comenzar las iteraciones del entrenamiento; usualmente se destinan 20% de los datos para probar. Después, en cada iteración del entrenamiento, se escogen de forma aleatoria los datos de entrenamiento y de validación. Luego, se entrena el modelo con los datos de entrenamiento y se verifica el error cometido con los datos de validación. Después, se vuelven a unir ambos subconjuntos y se toma nuevamente, de forma aleatoria, el subconjunto de entrenamiento y el de validación. Este proceso se repite de forma iterativa hasta completar un ciclo compuesto por n iteraciones y así se evita entrenar con un

mismo conjunto de datos y se hace la comprobación con nuevos datos.

2.3. Medidas de precisión

Para evaluar qué tan cerca están los valores pronosticados de los reales, se utiliza la raíz del error cuadrático medio (RMSE), véase ecuación (2.5), y el error absoluto medio (MAE), ecuación (2.6). El RMSE tiene el efecto de ser sensible a los errores, ya que los eleva al cuadrado; esto causa un desplazamiento en la media de los errores y por ende una puntuación de error mayor. El MAE calcula el promedio de los valores de error de pronóstico y se mantiene en las unidades originales de los valores predichos y, a diferencia del RSME, no es sensible a valores atípicos.

$$RMSE = \sqrt{\frac{\sum_{i=0}^n (x_i - \hat{x}_i)^2}{n}} \quad (2.5)$$

$$MAE = \frac{1}{n} \sum_{i=0}^n |x_i - \hat{x}_i| \quad (2.6)$$

2.4. Explicación del modelo MVKM

Los cursos actuales ofrecen a sus estudiantes dos tipos de materiales, los cuales son evaluativos y no evaluativos; compuestos por diversos recursos de aprendizaje que les permiten aprender y poner en práctica sus conocimientos.

Tal y como establece Zhao, S., Wang, C., y Sahebi, S. [1], “Ignorar los distintos tipos de materiales educativos en la modelización del conocimiento de los estudiantes limita nuestro entendimiento en saber cómo aprenden”. El hecho de saber cómo los estudiantes aprenden es importante; ya que permite modelar el conocimiento de un estudiante de la manera más acertada posible.

Existen diversos modelos que plantean distintas alternativas, como por ejemplo *Rank-Based Tensor Factorization* (RBTF), elaborado por Sahebi, S., y Doan, T., [9]; o *Automatic Knowledge Tracing* (AKT), realizado por González-Brenes, J., [10]; los cuales resuelven el problema de la predicción del conocimiento del estudiante.

Los modelos RBTF y AKT buscan modelar el conocimiento adquirido por los estudiantes. En el caso del RBTF, el conocimiento es modelado mediante la representación en el tiempo de la adquisición de los conceptos c inherentes a los recursos p . Este conocimiento resulta de la multiplicación matricial entre el vector de conceptos q_p y el vector de conocimientos, $t_{a,s}$, en esos conceptos del estudiante s en el tiempo a . El RBTF resuelve un problema de optimización por medio del descenso del gradiente; pero, no emplea ningún parámetro que haga referencia a las habilidades de aprendizaje de los estudiantes ni a los distintos tipos de materiales de aprendizaje. Por su parte, el AKT se basa en un vector ordenado de rendimiento de respuestas y en las habilidades de aprendizaje del estudiante para el cálculo del conocimiento del estudiante y tampoco emplea ningún parámetro para la modelización de los diversos tipos de materiales de aprendizaje.

Como alternativa, el modelo MVKM plantea una nueva forma de modelar el conocimiento de cada estudiante para cada recurso en cada instante de tiempo cuando interactúa con materiales evaluativos y no evaluativos. Por otra parte, a la vez que este modelo permite modelar el conocimiento del estudiante, como consecuencia de su estructura, descubre las similitudes entre los conceptos subyacentes en los materiales educativos.

Por lo tanto, este modelo intenta resolver de forma simultánea los problemas de modelización del conocimiento y conocimiento del área, tomando en cuenta la heterogeneidad de los materiales, y a su vez mitiga la sobreestimación del conocimiento de los estudiantes, mediante la introducción de factores de olvido de conceptos.

En el modelo MVKM, la predicción del conocimiento de cada estudiante para cada recurso en un punto de tiempo se realiza aplicando operaciones matriciales sobre S , T y Q . La matriz S representa las habilidades de aprendizaje de cada estudiante; la matriz T da una interpretación del conocimiento adquirido entre cada habilidad de aprendizaje y cada concepto del curso; y, por último, la matriz Q muestra la correspondencia entre los recursos y los conceptos.

La figura 2.4 da una representación gráfica del producto de las matrices S , T y $Q^{[r]}$ con $r = 0$ ó $r = 1$, cuando se tienen dos tipos de materiales educativos.

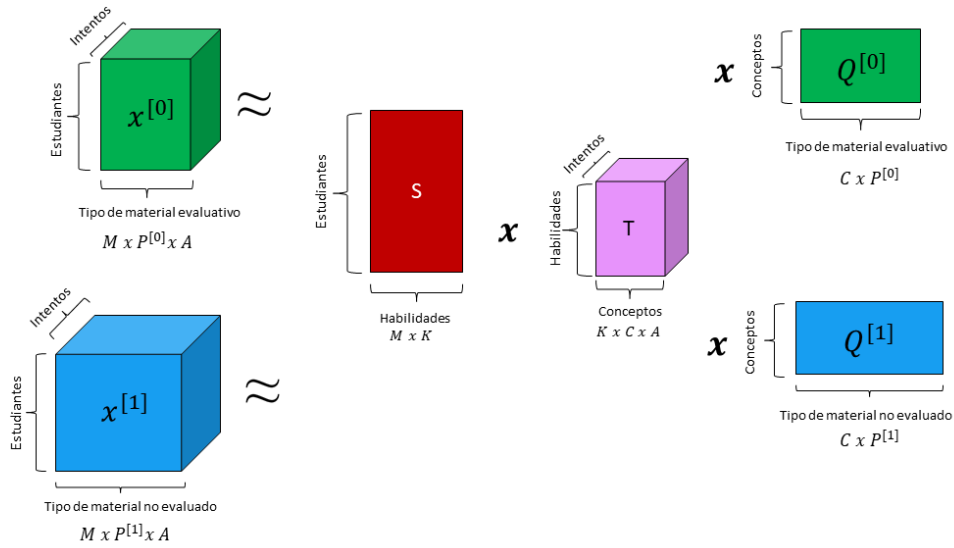


Figura 2.4: Factorización matriz tridimensional $X^{[r]}$.

Adaptación de imagen tomada del artículo *Modeling Knowledge Acquisition from Multiple Learning Resource Types*, [1]

El modelo cumple con las siguientes suposiciones: (a) Cada material de aprendizaje cubre algunos conceptos que se presentan en un curso; el conjunto de todos los conceptos del curso se comparte entre los materiales de aprendizaje. (b) Diferentes materiales de aprendizaje tienen distintos niveles de dificultad o utilidad para los estudiantes. (c) El curso puede seguir una tendencia en la presentación del material de aprendizaje: pasar de conceptos más fáciles a más difíciles o alternar entre conceptos fáciles y difíciles; a pesar de eso, los estudiantes pueden interactuar libremente con los materiales de aprendizaje y no están sujetos a una secuencia específica. (d) A medida que los estudiantes interactúan con estos materiales, aprenden los conceptos que se presentan en ellos; es decir, que aumenta su conocimiento en estos conceptos. (e) Dado que los estudiantes pueden olvidar algunos conceptos del curso, este aumento de conocimiento no es estricto. (f) Distintos estudiantes vienen con diversas habilidades de aprendizaje y valores de conocimiento inicial. (g) El cambio gradual de conocimiento varía entre estudiantes. Pero, los estudiantes se pueden agrupar de acuerdo a cómo cambia su conocimiento. (h) Eventualmente, el desempeño de un estudiante en un recurso de aprendizaje del tipo material evaluativo depende de los conceptos cubiertos, del conocimiento del estudiante en esos conceptos, la dificultad/utilidad del recurso y la habilidad general del estudiante. (i) El conocimiento adquirido por un estudiante es transferible para ser usado en otros recursos.

Adicionalmente, se permite el olvido ocasional de los conceptos que se están aprendiendo por parte de los estudiantes; esto quiere decir, que no necesariamente el conocimiento adquirido por el estudiante entre cada habilidad de aprendizaje y cada concepto en el intento $a + 1$ será mayor al conocimiento entre dichas habilidades y conceptos en el intento a , sin embargo, los olvidos ocasionales son penalizados a través del uso de una función indicatriz que señala el concepto para el cual se ha producido el olvido. Esta penalización evita variaciones abruptas en el conocimiento de los estudiantes consiguiendo un crecimiento gradual de conocimiento.

2.4.1. Definición de las variables y modelo

Sean:

- M el total de estudiantes inscritos en el curso.
- K el número de habilidades de aprendizaje por cada estudiante.
- C la cantidad de conceptos del curso.
- R el conjunto de los distintos tipos de materiales, materiales evaluativos o no evaluativos.
- $r \in R$ el índice de cada tipo de material
- $P^{[r]} \in r$ el conjunto de recursos pertenecientes al tipo de material r .
- $p \in P^{[r]}$ un recurso de $P^{[r]}$.
- A el número de intentos consecutivos en el tiempo.
- $a \in A$ un punto en el tiempo.

Y sean las matrices S , T y Q definidas como siguen:

- La matriz S de dimensión $M \times K$ que representa las habilidades de cada estudiante.
- La matriz tridimensional T de dimensión $K \times C \times A$ que captura la relación entre las habilidades de los estudiantes y los conceptos en sus distintos intentos.
- La matriz Q , que a partir de ahora se escribirá como $Q^{[r]}$, sugiriendo que es la matriz que pertenece a los recursos del tipo r de dimensión $C \times P^{[r]}$, muestra la relación entre los recursos educativos de un tipo de material de aprendizaje r y los conceptos.

Dado que el objetivo es cuantificar la interacción de cada estudiante s para todo el tiempo A para cada recurso educativo p del tipo de material r , esto se conseguirá al aplicar la multiplicación tensorial entre las matrices S , T y $Q^{[r]}$, tal que:

$$X^{[r]} = S \cdot T \cdot Q^{[r]} \quad (2.7)$$

Pero, en la práctica, no se realiza la multiplicación tensorial de las matrices S , T y $Q^{[r]}$; en cambio, de la matriz S se extrae el vector de habilidades del estudiante s , representado por s_s , que corresponde a la fila s -ésima de dimensión $1 \times K$, que será multiplicado con la matriz T_a de dimensión $K \times C$ -obtenida al hacer un corte transversal en la a -ésima cara de la matriz T - que captura y cuantifica el crecimiento de conocimiento de los estudiantes con cada una de las habilidades de aprendizaje para cada uno de los conceptos en el intento a . Dicho producto punto produce un vector de dimensión $1 \times C$ que indica la adquisición de conocimiento del estudiante s en cada uno de los conceptos en el intento a .

Seguido, de la matriz $Q^{[r]}$ se selecciona el vector del recurso p , definido como q_p , que señala la relación entre el recurso p y los conceptos; este vector tendrá dimensión $C \times 1$. Finalmente, el producto punto entre el vector de dimensión $1 \times C$, del producto parcial de S con T , y el vector q_p será igual a un escalar, $\hat{x}_{s,a,p}$, que da el valor del conocimiento del estudiante s para todo el tiempo a en el recurso p .

Si para el estudiante s se repite el proceso anterior en todos sus intentos y para todos los recursos se consigue el valor de conocimiento del estudiante s para cada intento y cada recurso.

Suponiendo que se realiza para los M estudiantes del curso, entonces el conocimiento de los estudiantes para cada intento y recurso puede ser representado por la matriz tridimensional $X^{[r]}$. Figura 2.5.

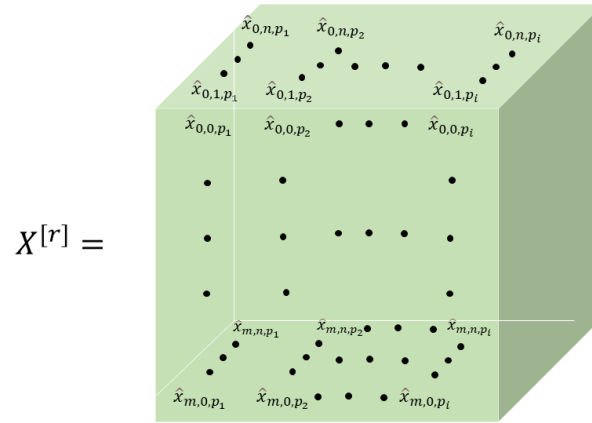


Figura 2.5: Matriz Tridimensional $X^{[r]}$

La a -ésima cara de la matriz tridimensional $X^{[r]}$, denotada por $X_a^{[r]}$, es una matriz que representa la interacción que tuvieron todos los estudiantes con los recursos $P^{[r]}$ del tipo de material r en un instante de tiempo a . Figura 2.6.

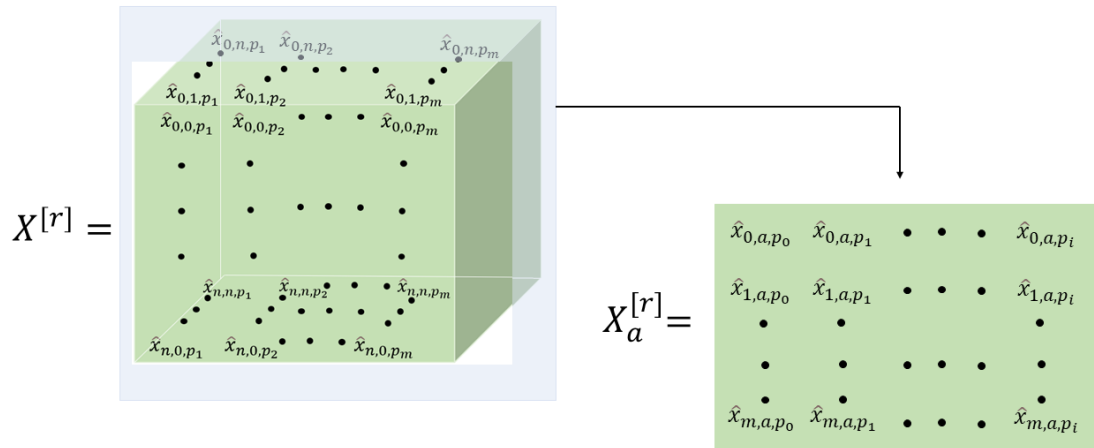


Figura 2.6: Matriz $X_a^{[r]}$

La s -ésima fila de esta matriz, $X_a^{[r]}$, será un vector definido como $x_{s,a}^{[r]}$ que muestra información sobre la interacción del estudiante s con todos los recursos $P^{[r]}$ del tipo de material r en el intento a . Figura 2.7.

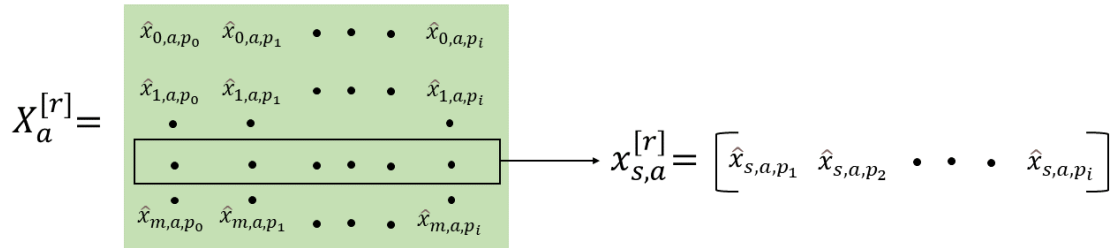


Figura 2.7: Vector $x_{s,a}^{[r]}$

Y por último, el escalar $x_{s,a,p}^{[r]}$ determina el valor del conocimiento del estudiante s en el recurso p del tipo de material r en el intento a . Figura 2.8.

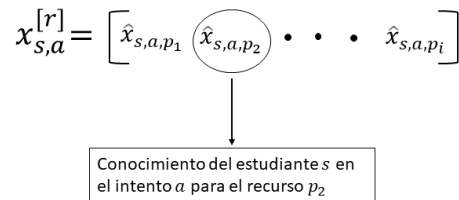


Figura 2.8: Escalar $x_{s,a,p}^{[r]}$

Por lo tanto, al operar como se describió antes, se calcula la predicción del conocimiento del estudiante s en un instante de tiempo a para el recurso p , $\hat{x}_{s,a,p}$, como:

$$\hat{x}_{s,a,p} = s_s \cdot T_a \cdot q_p \quad (2.8)$$

El caso particular en que el tipo el recurso p sea parte de los materiales no evaluativos se emplea la función sigmoidea; esta función se simboliza por convención con la letra griega σ .

Según Shai, S.S. y Shai, B.D.,[11], la función sigmoidea se define como $\sigma(\cdot) : \mathbb{R} \rightarrow [0, 1]$ tal que:

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \quad (2.9)$$

La figura 2.9 muestra la gráfica de la función sigmoidea.

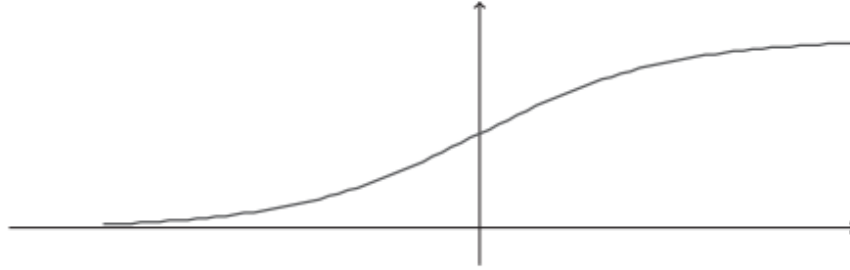


Figura 2.9: Gráfica función σ

Imagen tomada del libro *Understanding Machine Learning from Theory to Algorithms*, [11].

El valor observado del conocimiento, por definición, de dichos recursos será igual a 1; por lo que para la predicción del conocimiento será acotada superiormente por 1 mediante la función sigmoidea. De esta forma $\hat{x}_{s,a,p}$, estará definido por la siguiente ecuación:

$$\hat{x}_{s,a,p} = \sigma(s_s \cdot T_a \cdot q_p) \quad (2.10)$$

2.4.2. Construcción de la función de costo

En principio, la función de costo, \mathcal{L}_1 , se define como la diferencia cuadrática entre el valor real y el predicho.

$$\mathcal{L}_1 = \sum_{r,s,a,p} (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]})^2 \quad (2.11)$$

Con la finalidad de mantener la simplicidad del modelo y evitar su sobre-estimación se agregan los términos de regularización, $\|s_s\|^2$ y $\|T_a\|^2$, dados por la norma 2 o norma de Frobenius, junto a las constantes de regularización, λ_s y λ_t . Además, se añade el último hiperparámetro, γ , que ajusta el peso relativo del error en la regularización del modelo para cada tipo de material.

De esta manera, la función de costo queda como sigue:

$$\mathcal{L}_1 = \sum_{r,s,a,p} \gamma^{[r]} (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]})^2 + \lambda_t \|T_a\|^2 + \lambda_s \|s_s\|^2 \quad (2.12)$$

Por otra parte, como se permite el olvido ocasional de conocimiento de los estudiantes, entonces, el conocimiento adquirido para el par valor habilidad-concepto en el instante $a+1$ puede no ser mayor al conocimiento adquirido en el instante a . Sin embargo, lo deseado es que los estudiantes obtengan una adquisición gradual del conocimiento, por lo tanto, se espera que la matriz T_{a+1} sea mayor a la matriz T_a componente a componente. Por ende:

$$T_{a+1} - T_a \geq 0 \quad (2.13)$$

De esta manera, la ecuación (2.13) pudiese considerarse una restricción al problema de optimización que se incorpora al problema. Una forma de manejar dicha restricción es mediante la incorporación de la misma como una penalización en la función de costo en aquellos espacios donde se infringe. Esta penalización tiene el efecto de aumentar la habilidad de aprendizaje necesaria para generar un aumento en la adquisición de conocimiento en dicho concepto. Esa penalización viene dada por:

$$\mathcal{L}_2 = F(a) \quad (2.14)$$

Donde $F(a)$ se define como:

$$F(a) = \begin{cases} T_{a+1} - T_a & \text{si } T_{a+1}^{(i,j)} - T_a^{(i,j)} < 0 \\ T_a - T_{a-1} & \text{si } T_a^{(i,j)} - T_{a-1}^{(i,j)} < 0 \\ 0 & \text{caso contrario} \end{cases} \quad (2.15)$$

De este modo la derivada de \mathcal{L}_2 respecto a T_a es como sigue:

$$\frac{d\mathcal{L}_2}{dT_a} = \begin{cases} -1 & \text{si } T_{a+1}^{(i,j)} - T_a^{(i,j)} < 0 \\ 1 & \text{si } T_a^{(i,j)} - T_{a-1}^{(i,j)} < 0 \\ 0 & \text{caso contrario} \end{cases} \quad (2.16)$$

La derivada de \mathcal{L}_2 respecto a T_a resulta en una matriz de dimensión $K \times C$ de entradas iguales a 1, -1 ó 0 que indica los espacios en donde se incumple la restricción o no.

Finalmente, se desea minimizar \mathcal{L}_1 por medio del descenso del gradiente estocástico a la vez que se penalizan los olvidos en los pares valores habilidad-concepto tales que la diferencia entre el tiempo $a+1$ y el tiempo a es menor a 0 para cada una de las entradas de la matriz $T_{a+1} - T_a$. Por lo que lo que la función final \mathcal{L} , ecuación (2.17), incluye la función de la minimización de la norma de los errores al cuadrado y penalización dada en \mathcal{L}_2 siendo ω un parámetro que permita escalar entre ambos e indica el grado de penalización deseado.

$$\mathcal{L} = \mathcal{L}_1 - \omega \mathcal{L}_2 \quad (2.17)$$

Cabe señalar que, si $\omega = 0$ entonces \mathcal{L} queda compuesta solo por \mathcal{L}_1 , es decir, $\mathcal{L} = \mathcal{L}_1$.

2.4.3. Descenso del gradiente

Una vez que se ha definido la función de costo se debe encontrar su dirección de descenso. Para esto, es necesario calcular el gradiente de la función \mathcal{L} ; en otras palabras, se deben calcular las derivadas parciales de la función \mathcal{L} respecto a s_s , T_a y q_p ; y, aplicar el paso de actualización a cada uno de los parámetros.

El paso de actualización de cada parámetro queda como:

$$s_{s_{n+1}} = s_{s_n} - \alpha \left[\frac{\partial \mathcal{L}_1}{\partial s_s} \right] \quad (2.18)$$

$$T_{a_{n+1}} = T_{a_n} - \alpha \left[\frac{\partial \mathcal{L}_1}{\partial T_a} - \omega \frac{\partial \mathcal{L}_2}{\partial T_a} \right] \quad (2.19)$$

$$q_{p_{n+1}} = q_{p_n} - \alpha \left[\frac{\partial \mathcal{L}_1}{\partial q_p} \right] \quad (2.20)$$

Según lo que se observa en el paso de actualización de cada parámetro se tiene que $s_{s_{n+1}}$ y s_{s_n} son vectores de dimensión $1 \times K$ por lo que $\frac{\partial \mathcal{L}_1}{\partial s_s}$ debe resultar en un vector de la misma dimensión que $s_{s_{n+1}}$ y s_{s_n} . De igual forma, para $\frac{\partial \mathcal{L}_1}{\partial T_a}$ su resultado debe llevar a una matriz de dimensión $K \times C$ y para $\frac{\partial \mathcal{L}_1}{\partial q_p}$ se debe obtener un vector de dimensión $1 \times C$.

2.4.3.1. Derivadas parciales de \mathcal{L}_1

Para el calculo de las derivadas parciales de \mathcal{L}_1 existen dos casos a considerar. Caso 1: si p forma parte de los recursos del tipo evaluativo entonces el valor predicho $\hat{x}_{s,a,p}$, está dado por la ecuación (2.8). Caso 2: si p pertenece al conjunto de los materiales no evaluativos entonces $\hat{x}_{s,a,p}$ estará dado por la ecuación (2.10).

Dicho lo anterior, $\frac{\partial \mathcal{L}_1}{\partial s_s}$ cuando p es un recurso del tipo evaluado, se calcula como:

$$\begin{aligned}
\frac{\partial \mathcal{L}_1}{\partial s_s} &= \frac{\partial \left[\sum_{r,s,a,p} \gamma^{[r]} (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]})^2 + \lambda_t \|T_a\|^2 + \lambda_s \|s_s\|^2 \right]}{\partial s_s} \\
&= \sum_{r,s,a,p} \frac{\partial \left[\gamma^{[r]} (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]})^2 + \lambda_t \|T_a\|^2 + \lambda_s \|s_s\|^2 \right]}{\partial s_s} \\
&= \sum_{r,s,a,p} \left[\frac{\partial \left[\gamma^{[r]} (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]})^2 \right]}{\partial s_s} + \frac{\partial \left[\lambda_t \|T_a\|^2 + \lambda_s \|s_s\|^2 \right]}{\partial s_s} \right] \\
&= \sum_{r,s,a,p} \left[2\gamma^{[r]} \cdot (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]}) \frac{\partial \left[(x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]}) \right]}{\partial s_s} + \frac{\partial \left[\lambda_t \|T_a\|^2 + \lambda_s \|s_s\|^2 \right]}{\partial s_s} \right] \quad (2.21) \\
&= \sum_{r,s,a,p} \left[2\gamma^{[r]} \cdot (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]}) \frac{\partial \left[(x_{s,a,p}^{[r]} - s_s \cdot T_a \cdot q_p) \right]}{\partial s_s} + \frac{\partial \left[\lambda_t \|T_a\|^2 + \lambda_s \|s_s\|^2 \right]}{\partial s_s} \right] \\
&= \sum_{r,s,a,p} 2\gamma^{[r]} \cdot (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]}) \cdot (-)(T_a \cdot q_p) + 2\lambda_s \cdot s_s \\
&= \sum_{r,s,a,p} -2\gamma^{[r]} \cdot (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]}) \cdot T_a \cdot q_p + 2\lambda_s \cdot s_s
\end{aligned}$$

Podemos notar que $\frac{\partial \mathcal{L}_1}{\partial s_s}$ resulta en un vector de dimensión $1 \times K$ según el producto punto indicado por T_a y q_p . Por lo tanto, se preserva la dimensión entre los vectores $s_{s_{n+1}}$, s_{s_n} y el resultado dado por $\frac{\partial \mathcal{L}_1}{\partial s_s}$.

Análogamente, se repiten los pasos de la ecuación (2.21) y se hallan las derivadas parciales $\frac{\partial \mathcal{L}_1}{\partial T_a}$ y $\frac{\partial \mathcal{L}_1}{\partial q_p}$.

De tal forma que,

$$\frac{\partial \mathcal{L}_1}{\partial T_a} = \sum_{r,s,a,p} -2\gamma^{[r]} \cdot (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]}) \cdot s_s \cdot q_p + 2\lambda_t \cdot T_a \quad (2.22)$$

y,

$$\frac{\partial \mathcal{L}_1}{\partial q_p} = \sum_{r,s,a,p} -2\gamma^{[r]} \cdot (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]}) \cdot s_s \cdot T_a \quad (2.23)$$

De igual forma que para el caso anterior, tanto $\frac{\partial \mathcal{L}_1}{\partial T_a}$ y $\frac{\partial \mathcal{L}_1}{\partial q_p}$ resultan en una matriz y un vector que son equivalentes en dimensión a la matriz $T_{a_{n+1}}$ y al vector $q_{p_{n+1}}$, respectivamente.

Para el caso 2, se usará la derivada de la función sigmoidea que es la siguiente:

$$\sigma'(\cdot) = \sigma(\cdot)(1 - \sigma(\cdot)) \quad (2.24)$$

De modo tal:

$$\begin{aligned} \frac{\partial \mathcal{L}_1}{\partial s_s} &= \frac{\partial \left[\sum_{r,s,a,p} \gamma^{[r]} (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]})^2 + \lambda_t \|T_a\|^2 + \lambda_s \|s_s\|^2 \right]}{\partial s_s} \\ &= \sum_{r,s,a,p} \frac{\partial \left[\gamma^{[r]} (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]})^2 + \lambda_t \|T_a\|^2 + \lambda_s \|s_s\|^2 \right]}{\partial s_s} \\ &= \sum_{r,s,a,p} \left[\frac{\partial \left[\gamma^{[r]} (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]})^2 \right]}{\partial s_s} + \frac{\partial \left[\lambda_t \|T_a\|^2 + \lambda_s \|s_s\|^2 \right]}{\partial s_s} \right] \\ &= \sum_{r,s,a,p} \left[2\gamma^{[r]} \cdot (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]}) \frac{\partial \left[(x_{s,a,p}^{[r]} - \sigma(\hat{x}_{s,a,p}^{[r]})) \right]}{\partial s_s} + \frac{\partial \left[\lambda_t \|T_a\|^2 + \lambda_s \|s_s\|^2 \right]}{\partial s_s} \right] \quad (2.25) \\ &= \sum_{r,s,a,p} \left[2\gamma^{[r]} \cdot (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]}) \frac{\partial \left[(x_{s,a,p}^{[r]} - \sigma(s_s \cdot T_a \cdot q_p)) \right]}{\partial s_s} + \frac{\partial \left[\lambda_t \|T_a\|^2 + \lambda_s \|s_s\|^2 \right]}{\partial s_s} \right] \\ &= \sum_{r,s,a,p} 2\gamma^{[r]} \cdot (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]}) \cdot (-)(s_s \cdot T_a \cdot q_p)(1 - (s_s \cdot T_a \cdot q_p))(T_a \cdot q_p) + 2\lambda_s \cdot s_s \\ &= \sum_{r,s,a,p} -2\gamma^{[r]} \cdot (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]}) \cdot \hat{x}_{s,a,p}^{[r]} \cdot (1 - \hat{x}_{s,a,p}^{[r]}) \cdot T_a \cdot q_p + 2\lambda_s \cdot s_s \end{aligned}$$

Análogamente, se procede de la misma manera para las derivadas parciales $\frac{\partial \mathcal{L}_1}{\partial T_a}$ y $\frac{\partial \mathcal{L}_1}{\partial q_p}$.

$$\frac{\partial \mathcal{L}_1}{\partial T_a} = \sum_{r,s,a,p} -2\gamma^{[r]} \cdot (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]}) \cdot \hat{x}_{s,a,p}^{[r]} \cdot (1 - \hat{x}_{s,a,p}^{[r]}) \cdot s_s \cdot q_p + 2\lambda_t \cdot T_a \quad (2.26)$$

$$\frac{\partial \mathcal{L}_1}{\partial q_p} = \sum_{r,s,a,p} -2\gamma^{[r]} \cdot (x_{s,a,p}^{[r]} - \hat{x}_{s,a,p}^{[r]}) \cdot \hat{x}_{s,a,p}^{[r]} \cdot (1 - \hat{x}_{s,a,p}^{[r]}) \cdot s_s \cdot T_a \quad (2.27)$$

Luego, para construir el paso de actualización del caso 1 se sustituyen las ecuaciones (2.21) en la ecuación (2.18), (2.22) en (2.19) y por último, (2.23) en (2.20).

Y, para el paso de actualización según el caso 2 se sustituyen las ecuaciones (2.25) en la ecuación (2.18), (2.26) en (2.19); y, (2.27) en (2.20).

Para finalizar, se han construido las expresiones que indican cuál es el paso de actualización de cada parámetro, para minimizar \mathcal{L}_1 .

CAPÍTULO 3

METODOLOGÍA

Nunca alguna gran obra se ha hecho de
prisa. Lograr un gran descubrimiento
científico, imprimir una excelente
fotografía, escribir un poema inmortal...
hacer cualquier gran logro requiere
tiempo, paciencia y perseverancia

W. J. Wilmont Buxton

En esta sección se detallan los datos extraídos del curso CO6612, la transformación aplicada a los mismos, las condiciones de convergencia del entrenamiento y la calibración de los hiperparámetros.

3.1. Datos del curso CO6612

Los datos a los que se les aplicó el modelo MVKM fueron extraídos del curso Redes Neuronales, código CO6612, del Departamento de Cómputo Científico de la Universidad Simón Bolívar, dictado por la profesora Minaya Villasana de los trimestres abril-julio 2020 y septiembre-diciembre 2021.

El curso contó con un total de 17 estudiantes, 2 tipos de materiales educativos -evaluados y no evaluados-, 12 recursos del tipo de material evaluativo y 8 recursos del tipo no evaluativo; para un total de 20 recursos disponibles para cada estudiante.

Los recursos educativos son los siguientes:

- Materiales no evaluados
 - Evaluación Formativa Actividad Formativa
 - Evaluación Formativa el Perceptrón
 - Evaluación Formativa el Adaline
 - Evaluación Formativa el Perceptrón Multicapas
 - Evaluación Formativa el RBF
 - Evaluación Formativa el SVM
 - Discusión El Perceptrón

- Discusión Adaline
- Materiales evaluados
 - Quiz Evaluativo el Perceptrón
 - Quiz Evaluativo el Adaline
 - Quiz Evaluativo el Perceptrón Multicapas
 - Quiz Evaluativo el RBF
 - Quiz Evaluativo el SVM
 - Tarea 1 (La Neurona)
 - Tarea 2 (Perceptrón)
 - Tarea 3 (Adaline)
 - Tarea 4 (PM)
 - Tarea 5 (RBF)
 - Tarea 6 (SVM)
 - Proyecto Final

Para propósitos de una mejor legibilidad del trabajo, a partir de ahora se abreviará Evaluación Formativa como EF y Quiz Evaluativo como QE.

Las evaluaciones formativas tienen la intención de que los alumnos pongan en práctica los conocimientos adquiridos; la misma es corta y consiste en una selección simple no calificada. Al finalizar la evaluación, el estudiante verifica cuáles respuestas fueron acertadas y cuáles no. Si la respuesta era incorrecta, se le notificaba cuál era la opción correcta. A su vez, los 20 recursos involucraban 5 conceptos macros, el Perceptrón, el Adaline, el Perceptrón Multicapas (PM), Las Máquinas de Soporte Vectorial (SVM) y las Funciones de Base Radial (RBF).

La correspondencia entre los conceptos y los recursos es la siguiente:

- El Perceptrón: EF el Perceptrón, QE el Perceptrón, tarea 1 (la neurona), tarea 2 (Perceptrón), discusión el Perceptrón.
- El Adaline: EF Adaline, QE Adaline, tarea 3 (Adaline), discusión el Adaline.
- El Perceptrón Multicapas: EF Perceptrón Multicapas, QE Perceptrón Multicapas, tarea 4 (PM)

- RBF: EF RBF, QE RBF, tarea 5 (RBF)
- El SVM: EF SVM, QE SVM, tarea 6 (SVM)

Se puede destacar que cada recurso de aprendizaje cubre algún o algunos conceptos presentados en el curso. Cada recurso de aprendizaje posee su propio nivel de dificultad. El curso presenta una secuencia en la forma en la que los estudiantes interactúan con los recursos, pero no es necesario que sigan dicha secuencia tal y como fue establecida. Es decir, un estudiante puede realizar en primer lugar la tarea asociada a un concepto, luego la evaluación formativa y por último el examen; así como otro estudiante puede decidir ejecutar primero la evaluación formativa, luego el examen y al final la tarea.

A medida que los estudiantes interactúan con los recursos adquieren mayor conocimiento; sin embargo, es posible que olviden fragmentos de lo aprendido. Todos los estudiantes poseen un nivel de conocimiento distinto, puesto que es factible que estén en distintas etapas de su carrera universitaria y cada uno trae consigo diferentes habilidades de aprendizaje. El resultado en un recurso dependerá de los conceptos, del conocimiento del estudiante en los conceptos, y de la habilidad general del estudiante y finalmente, el conocimiento adquirido en un recurso de aprendizaje es transferible para usarlo en otros recursos.

3.2. Transformación de los datos

Los datos extraídos del curso tienen la siguiente forma:

- Identificación del alumno
- Recurso con el que interactuó
- Fecha de interacción
- Nota obtenida

Para garantizar la anonimidad de los estudiantes, se le asignó un identificador único a cada uno (ID), lo mismo se hizo con los recursos. Todas las calificaciones fueron escaladas, de modo que las notas obtenidas están entre 0 y 1. Los materiales evaluados se denotan con el índice 0; en cambio, los materiales no evaluados se indican con el índice 1.

Cabe mencionar que los estudiantes no estaban en la obligación de llevar a cabo todas las actividades ofrecidas en el curso ni seguir un orden preestablecido; es por ello, que, se pueden ver

casos en que un estudiante tenga 10 intentos y otro tenga 15 intentos, y que en el mismo intento hayan efectuado un recurso distinto; por ejemplo, en el intento 3, el estudiante 4 realizó la pregunta 6; mientras que en el intento 3, el estudiante 5 realizó la pregunta 10.

La tabla 3.1 presenta la asociación entre cada uno de los recursos de aprendizaje y el número de pregunta asignado.

Gracias a la fecha de interacción del estudiante con el recurso se conoce el orden en el que el estudiante efectuó la interacción y permite asignar el valor correspondiente a esa interacción; por ejemplo, supóngase que el estudiante identificado por el número 2 comenzó sus estudios el día 20 de julio y ese mismo día usó el recurso 1 y el día 21 de julio practicó con el recurso 8; en ese caso, el día 20 de julio será su intento 0 y el día 21 de julio será el intento 1, y así sucesivamente.

La tabla 3.2 ofrece una mirada final a algunos de los datos transformados del curso CO6612. En total se reunieron 259 datos de los 17 estudiantes; de los cuales 188 son del tipo material evaluado y 71 son tipo material no evaluado. Adicionalmente, existieron otros tipos de material no evaluados (videos y lecturas recomendadas); pero no es posible conocer si fueron usados y el momento en el que esto habría ocurrido. Por tanto, esta información no entró en el conjunto.

ID Estudiante	Id del Recurso	Nota Obtenida	Intento	Tipo de Material
1	0	1	0	1
16	1	0.25	1	0
4	3	1	15	1
3	4	1	13	1
13	3	0.85	11	0
3	10	1	15	0
10	11	0.9	16	0

Tabla 3.2: Datos transformados del curso CO6612.

3.3. Elección de conjuntos para entrenar y probar

La estrategia que se emplea para la elección y presentación de los datos al modelo es la explicada en la sección 2.2.

Para los datos de prueba se seleccionan, de forma aleatoria, el 20 % de los datos totales; es decir, 51 datos. Del 80 % restante, perteneciente al conjunto para entrenar, se particiona el conjunto para entrenar en datos de entrenamiento y validación. Los datos de validación constan del 20 % del conjunto de entrenamiento, solo cuando el tipo de material es evaluativo. Por lo cual, el entrenamiento

Recurso	Id del Recurso	Tipo de Material
EF Actividad Formativa	0	1
EF Percetrón	1	1
EF Adaline	2	1
EF PM	3	1
EF RBF	4	1
EF SVM	5	1
Discusión Adaline	6	1
Discusión Perceptrón	7	1
QE Adaline	0	0
QE Perceptrón	1	0
QE PM	2	0
QE RFB	3	0
QE SVM	4	0
QE Proyecto Final	5	0
Tarea 1 (La neurona)	6	0
Tarea 2 (Perceptrón)	7	0
Tarea 3 (Adaline)	8	0
Tarea 4 (PM)	9	0
Tarea 5 (RBF)	10	0
Tarea 6 (SVM)	11	0

Tabla 3.1: Número de identificación de cada recurso.

es llevado a cabo con 167 datos.

3.4. Proceso de entrenamiento

Definidos los datos y la estrategia para entrenar se aplica el modelo MVKM y se ajustan los parámetros S , T y Q . El ajuste de las matrices S , T y Q se lleva a cabo cada vez que el algoritmo completa un ciclo compuesto por 40 iteraciones máximo. Dentro de cada una de las iteraciones se procede a entrenar las matrices, reemplazando sus valores a través de los pasos de actualización definidos en las ecuaciones (2.18), (2.19) y (2.20), respectivamente. El proceso se detiene si se

cumple alguna de las siguientes condiciones:

- Se alcanza el máximo de iteraciones.
- Se realizan mínimo 35 iteraciones y el RMSE de la diferencia entre las predicciones de los datos de validación de la última y la penúltima iteración es menor que una cierta tolerancia.
- Se ejecutan mínimo 35 iteraciones y el RMSE de los datos de validación en la n -ésima iteración es mayor que el promedio del RMSE cometido en los datos de validación en las últimas 3 iteraciones.

CAPÍTULO 4

EXPERIMENTACIÓN Y DISCUSIÓN

Si buscas resultados distintos no hagas
siempre lo mismo

Albert Einstein

En este capítulo se presentan los resultados de la aplicación del modelo MVKM a los datos del curso CO6612. Se analiza el nivel del conocimiento de los estudiantes en cada concepto, se agrupan y caracterizan a los estudiantes según sus habilidades de aprendizaje, se analiza el aporte al conocimiento por parte de cada recurso y se detalla la correspondencia entre los recursos, del tipo evaluativo, y los conceptos.

Para el ajuste de los parámetros al modelo MVKM se realizan los siguientes casos:

- Caso 1: Se minimiza únicamente la función \mathcal{L}_1 sin tomar en cuenta el efecto de la penalización por parte de \mathcal{L}_2 . Esto quiere decir, que para el paso de actualización solo se consideran las derivadas parciales sobre la función \mathcal{L}_1 y se asume que $\omega = 0$ anulando a \mathcal{L}_2 en el paso de actualización de la ecuación (2.19).
- Caso 2: Se minimiza la función \mathcal{L} considerando la penalización sobre la disminución del conocimiento.

Cabe acotar que, para cada uno de los casos, los parámetros e hiperparámetros a ser ajustados son los mismos, a excepción de cuando $\omega = 0$.

4.1. Calibración de hiperparámetros

El problema de optimización contiene muchos hiperparámetros o constantes que definen el error a minimizar usando el proceso de entrenamiento descrito anteriormente. A priori, no se conocen los valores de los hiperparámetros que definen la mejor función de costo para nuestro problema; por tanto sus valores óptimos deben ser descubiertos también. De esta forma, se lleva a cabo una búsqueda exhaustiva para hallar los mejores valores de los hiperparámetros o constantes K , λ_s , λ_t , $\gamma^{[r]}$, ω y α .

Para disminuir la cantidad de parámetros que se deben calibrar en la función objetivo se escala \mathcal{L}_1 usando $\gamma^{[1]}$ y redefiniendo $\gamma = \frac{\gamma^{[2]}}{\gamma^{[1]}}$, $\lambda_s^* = \lambda_s / \gamma^{[1]}$. Para simplicidad de notación y escritura se elimina el $*$.

En principio, se hace la calibración de los hiperparámetros K , λ_s , λ_t , γ y ω , asignando un valor relativamente pequeño para la tasa de aprendizaje α que garantice convergencia. Seguidamente se gradúa el valor de α para los mejores hiperparámetros de la función de costo encontrados.

Para poner en práctica la búsqueda exhaustiva se aplica la técnica de búsqueda en cuadrícula. Esta técnica lleva a cabo todas las posibles combinaciones para cada uno de los posibles valores de cada hiperparámetro, luego, a través del proceso de entrenamiento, se computan los errores RMSE y MAE y se ordenan las combinaciones en orden ascendente, según lo obtenido en el RMSE.

En la tabla 4.1 se presentan los posibles valores de cada hiperparámetro, para cada uno de los casos antes mencionados, con los que se hizo la búsqueda en cuadrícula. Los valores acá propuestos están dentro de los rangos sugeridos en la literatura de *Modeling Knowledge Acquisition from Multiple Learning Resource Types*, [1].

Caso	Hiperparámetro	Valores
Caso 1	K	3; 5; 10; 15
	λ_s	0; 0.01; 0.05
	λ_t	0; 0.01; 0.05
	ω	0
	γ	0.1; 0.2; 0.5; 1
	α	0.005; 0.01; 0.02; 0.1
Caso 2	K	3; 5; 10
	λ_s	0; 0.01; 0.05
	λ_t	0; 0.01; 0.05
	ω	0; 0.0001; 0.001; 0.01; 0.1
	γ	0.1; 0.2; 0.5; 1
	α	0.0001; 0.001; 0.005; 0.01; 0.02; 0.1

Tabla 4.1: Posibles valores de los hiperparámetros para la búsqueda en cuadrícula.

La tabla 4.2, muestra los mejores hiperparámetros encontrados para cada uno de los casos.

Observando ambas tablas, 4.1 y 4.2, se detalla que los valores para K , λ_s y λ_t son iguales para todos los casos.

Para este estudio en particular, se encuentra que λ_s y λ_t son iguales a 0, debido a que, su efecto no es determinante en el control del aumento en norma de las matrices S , T y Q , es decir, los parámetros λ_s y λ_t no juegan un papel en la regularización del crecimiento de las matrices. Por esta razón, se admite su anulación. Cabe destacar que los autores de la propuesta incluyeron este

Caso	Hiperparámetro	Valores
Caso 1	K	3
	λ_s	0
	λ_t	0
	ω	0
	γ	1
	α	0.01
Caso 2	K	3
	λ_s	0
	λ_t	0
	ω	0.01
	γ	0.2
	α	0.005

Tabla 4.2: Valores óptimos de los hiperparámetros para cada caso.

término puesto que en algunos casos los datos eran generados aleatoriamente y había mucha variabilidad y poca consistencia en la información, mientras que al entrenar con datos provenientes de cursos reales los autores también encontraban que los términos de regularización eran innecesarios (ver tabla 4 de *Modeling Knowledge Acquisition from Multiple Learning Resource*, [1]).

En última instancia, la tabla 4.3, expone los resultados según el RMSE y MAE que se obtuvieron al aplicar el modelo a los datos para cada uno de los casos según sus mejores hiperparámetros los cuales están dentro de los rangos reportados en *Modeling Knowledge Acquisition from Multiple Learning Resource*, [1], al aplicar el MVKM a conjuntos de datos distintos.

Caso	RMSE	MAE
Caso 1	0.1863	0.1337
Caso 2	0.2583	0.2177

Tabla 4.3: RMSE y MAE para cada caso.

Los valores hallados para K , λ_s , λ_t , γ y α en combinación con la optimización de la función \mathcal{L}_1 o \mathcal{L}_2 -por medio del descenso del gradiente estocástico- y la estrategia de entrenamiento determinan los valores de las matrices S , T y Q , para cada caso, con las que se desarrollan los experimentos y conclusiones sobre la aplicación del MVKM a los datos del curso.

4.2. Aporte de cada recurso al nivel de conocimiento de los estudiantes

Las figuras 4.1 y 4.2 constan de un gráfico de barras que cuantifica el nivel de conocimiento promedio aportado por cada recurso evaluativo correspondientes al caso 1 y caso 2, respectivamente.

En la figura 4.1 se observa que los recursos QE PM, QE Adaline y QE Proyecto Final contribuyen en mayor medida al conocimiento; seguido de todas las tareas. Finalmente, se encuentran los recursos QE Perceptrón, QE RBF y QE SVM. Cabe destacar que el orden de contribución de los recursos es independiente del promedio de las notas obtenidas por los estudiantes en cada recurso.

La figura 4.2 indica que los recursos tarea 6 (SVM), QE Proyecto Final y QE PM contribuyen en mayor medida al conocimiento; seguido de la tarea 4 (PM) y tarea 3 (Adaline) QE. Los recursos que menor aportan al conocimiento son QE Perceptrón y QE RBF.

Al efectuar la comparativa con lo obtenido en ambas figuras, se aprecia que los recursos QE PM y QE Proyecto Final son los que aportan mayor conocimiento al curso, además, en ambos casos se coincide con que los recursos QE Perceptrón y QE RBF son los que en menor proporción contribuyen al conocimiento de los estudiantes.

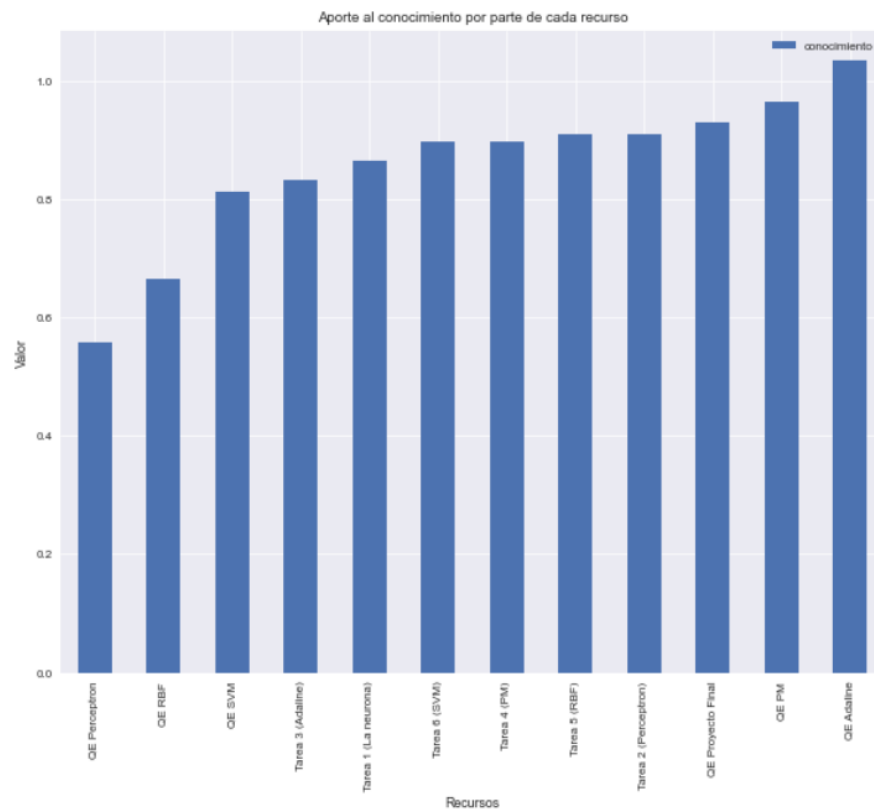


Figura 4.1: Aporte al conocimiento por parte de cada recurso, caso 1.

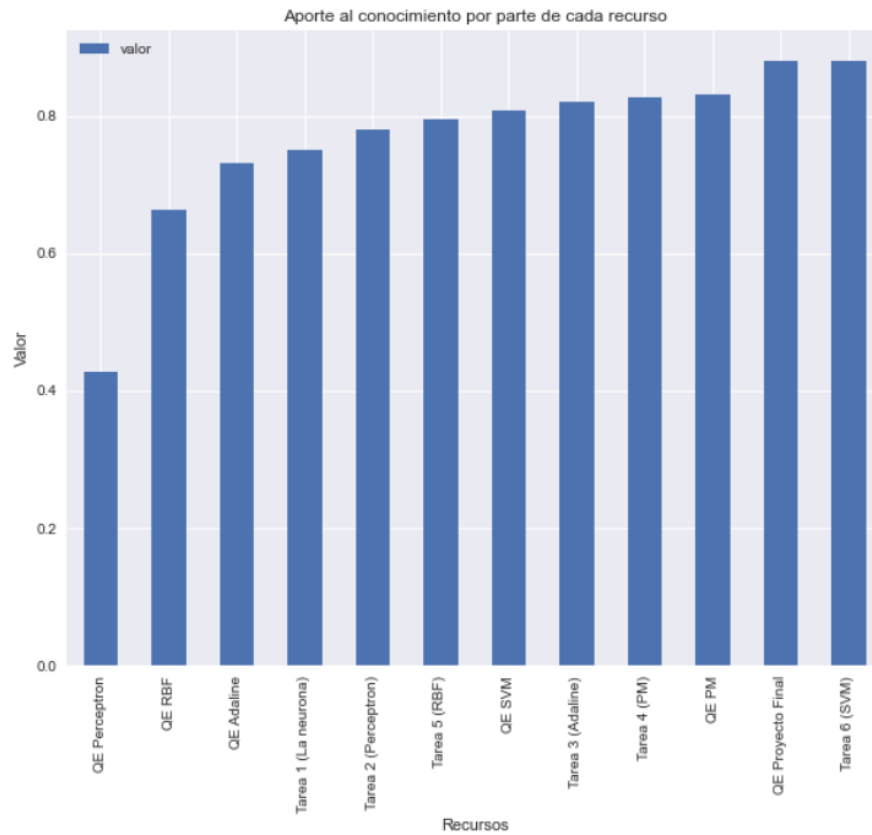


Figura 4.2: Aporte al conocimiento por parte de cada recurso, caso 2.

4.3. Correspondencia entre los recursos y los conceptos

Se desea ver la correspondencia que mantiene el modelo MVKM para los datos del curso. Para ello, detallaremos la matriz $Q^{[r]}$ cuando r pertenece a los materiales del tipo evaluativo.

En las figuras 4.3 y 4.4 se visualiza un gráfico de barras apiladas horizontales correspondientes al caso 1 y caso 2, respectivamente. En el eje y se encuentran cada uno de los recursos y en el eje x los pesos relativos de cada concepto para el recurso. La suma de los pesos relativos de los conceptos para cada recurso es siempre igual a 1.

Para la figura 4.3, el recurso QE Proyecto Final muestra que todos los conceptos influyen al menos en un 15 %. Luego en el QE SVM hay una representación del 100 % por parte del concepto 0 y este también se encuentra por encima del 30 % para los recursos QE PM y Adaline. En el caso del concepto 1, tiene un peso del 100 % para la tarea 5 (RBF) y su peso relativo en los demás recursos, exceptuando el QE Proyecto Final, no supera el 15 %. Por su parte el concepto 2 tiene gran peso en los recursos Tarea 1 (la neurona), Tarea 2 (Perceptrón), Tarea 3 (Adaline) y QE PM. Por otro lado, el concepto 3 se ve involucrado en los recursos QE RBF y Perceptrón y en la tarea 6 (SVM). Finalmente, el concepto 4 tiene una importancia en los recursos QE Adaline, Tarea 4 (PM) y Tarea

6 (SVM).

De esta forma, cada concepto se compone de los siguientes recursos:

- Concepto 0: QE SVM, QE PM, QE Adaline y QE Proyecto Final
- Concepto 1: Tarea 5 (RBF) y QE Proyecto Final
- Concepto 2: Tarea 1 (la neurona), Tarea 2 (Perceptrón) y Tarea 3 (Adaline), QE PM y QE Proyecto Final
- Concepto 3: Tarea 6 (SVM), QE RBF, QE Perceptrón y QE Proyecto Final
- Concepto 4: Tarea 4 (PM), Tarea 6 (SVM), QE Adaline y QE Proyecto Final

En cambio, en la figura 4.4 se observa -a excepción de los recursos QE Perceptrón y QE RBF que además son los que menos aporte generan al conocimiento- como cada una de las barras de distintos colores miden prácticamente lo mismo, es decir, cada recurso tiene la misma importancia en cada uno de los conceptos.

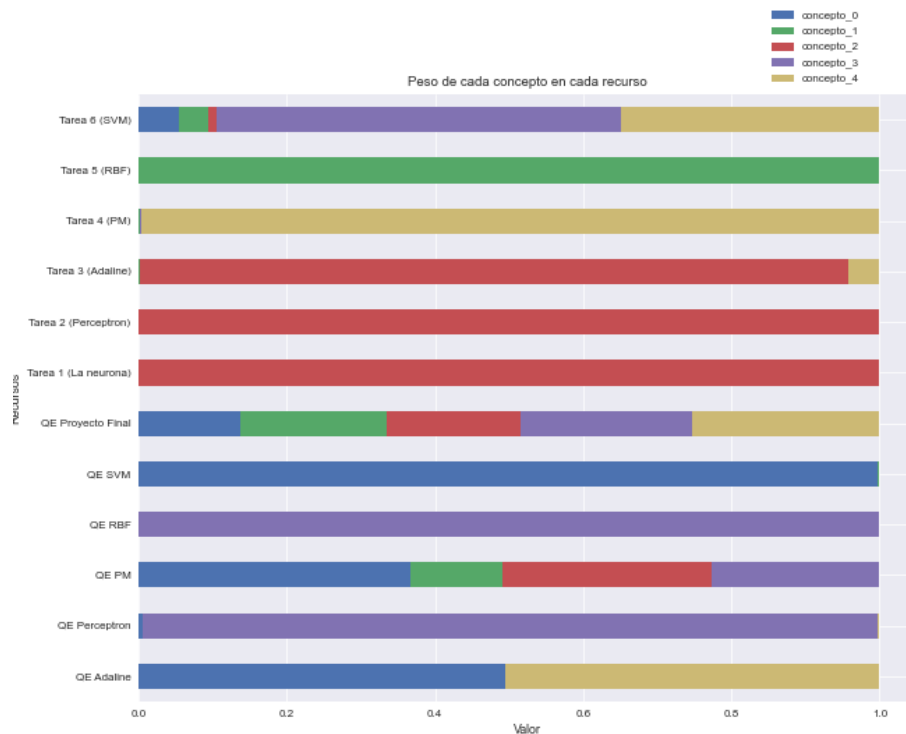


Figura 4.3: Peso relativo de cada concepto en cada recurso, caso 1.

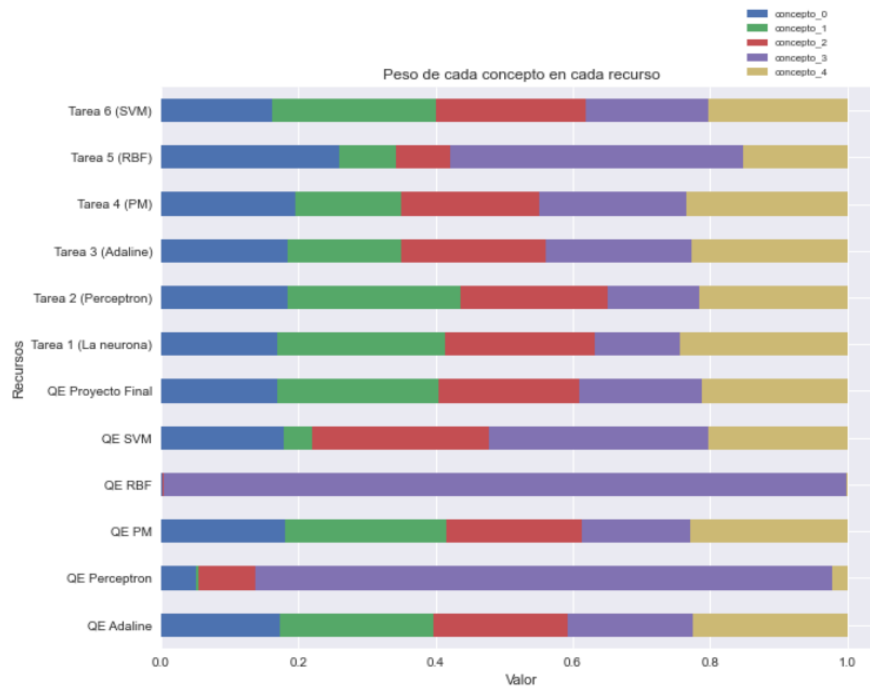


Figura 4.4: Peso relativo de cada concepto en cada recurso, caso 2.

4.4. Análisis de los estudiantes

Luego de entrenar el modelo para el caso 1, es decir, sin considerar restricciones sobre el crecimiento del conocimiento adquirido, se puede visualizar en la figura 4.5 el panorama general del nivel de conocimiento promedio adquirido de los estudiantes. La figura está compuesta por un diagrama de caja que permite ver la distribución del conocimiento de los estudiantes en cada intento.

Lo primero que se evidencia es una fluctuación a lo largo de los intentos en la adquisición de conocimiento de los estudiantes. En general, entre los intentos 4 a 8 el conocimiento adquirido promedio varía menos que en otros instantes del curso. Esto se debe a que los estudiantes, en su mayoría, realizaron los mismos QE y las EF por lo que pudiese conjeturar que el aporte de esas actividades al conocimiento adquirido por cada estudiante es similar. Luego, para el intento 9 efectúan las tareas y el nivel de conocimiento adquirido decae. Similarmente ocurre entre los intentos 10 al 12. Los estudiantes asimilan nuevos conceptos y a la hora de llevarlos a la práctica se produce una disminución del conocimiento lo cual puede resultar contraintuitivo.

Por otro lado, se aprecian algunos puntos en la parte baja y alta de los diagramas, conocidos como datos atípicos, que representan a estudiantes con comportamientos distintos al grupo. Esto da a entender que los estudiantes poseen habilidades de aprendizaje distintas.

Por el contrario, la figura 4.6, calculada utilizando el caso 2, exhibe una adquisición del conocimiento constante independiente del recurso con el que interactuó cada estudiante. En este caso, los

diagramas de caja se muestran con una forma irregular y con dispersión en su parte baja; indicando, una diferenciación en la adquisición de conocimiento por parte de los estudiantes.

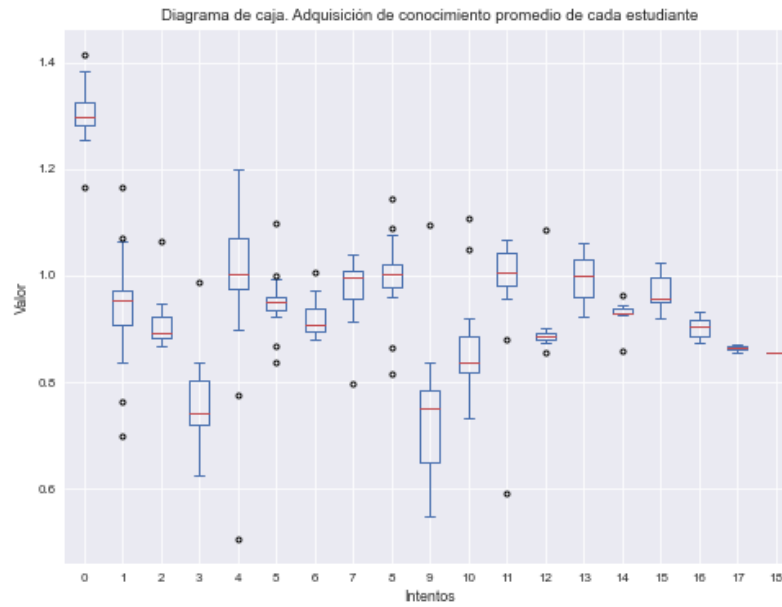


Figura 4.5: Adquisición de conocimiento de los estudiantes, caso 1.

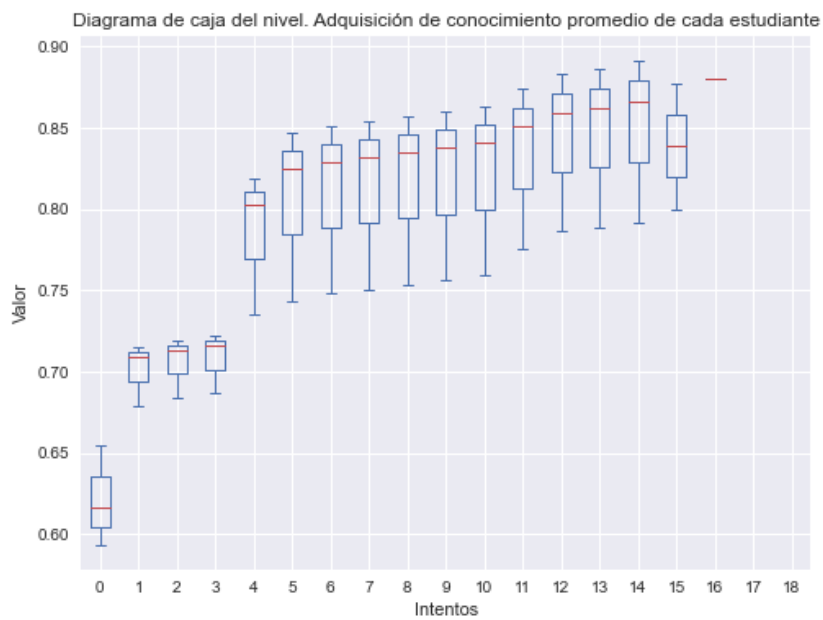


Figura 4.6: Adquisición de conocimiento de los estudiantes, caso 2.

4.5. Construcción de los grupos según las habilidades de aprendizajes de cada estudiante

Por lo visto anteriormente, se puede considerar agrupar a los estudiantes según sus habilidades de aprendizaje. Para ello, se aplica el algoritmo de conglomeración K -medias. Según Gareth, J., Daniela, W., Trevor, H., y Robert, T., [5]:

La agrupación en clústeres de k -medias es un enfoque simple y elegante para dividir un conjunto de datos en k grupos distintos que no se superponen. Para realizar el agrupamiento de k -medias, primero debemos especificar el número deseado de agrupamientos k ; entonces el algoritmo de k -medias asignará cada observación exactamente a uno de los k conglomerados.

La asignación de cada observación se determina según la distancia entre los puntos y los respectivos centroides de los grupos. Si un punto x está más cerca del centroide x_1 que del x_2 entonces x es asignado al mismo grupo de x_1 . El proceso se repite para cada uno de los puntos.

La distancia que se utiliza para medir la diferencia entre los puntos de la matriz S y los centroides es la distancia Euclídea.

El número de grupos k para los datos se verifica gracias al método del codo. En este método se calcula la suma de las distancias al cuadrado de los puntos a su centro de conglomerado más cercano; dicha diferencia es conocida como inercia. Los resultados son graficados y se determina el número óptimo de conglomerados cuando la inercia disminuye significativamente. Es importante mencionar que el método del codo es un método referencial para ayudar en la decisión final sobre la cantidad de grupos a considerar.

La representación gráfica de los puntos de la matriz S para cada caso se pueden ver en las figuras 4.7 y 4.8.

Véase las figuras 4.9 y 4.10 para visualizar las gráficas obtenidas por el método del codo. Considerando ambas figuras se puede elegir $k = 2$ o $k = 3$ como número óptimo de conglomerados. Sin embargo, al computar los conglomerados con $k = 3$, tanto para el caso 1 como para el caso 2, un grupo queda conformado por un solo integrante (estudiante 8). Por esta razón, se aplica el algoritmo de k -medias con $k = 2$ para construir los grupos en ambos casos. Véase las figuras 4.11 y 4.12 en donde cada color indica un grupo y el punto en forma de estrella muestra el centroide de cada uno de los grupos. La posición del estudiante 8 indica un comportamiento atípico.

La tabla 4.4 resume las agrupaciones realizadas por el algoritmo. Analizando la tabla, es claro que los estudiantes 1, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14 y 15 siguen formando parte de un mismo grupo a su vez que, los estudiantes 2 y 16, permanecen juntos en ambos casos. Por su parte, la

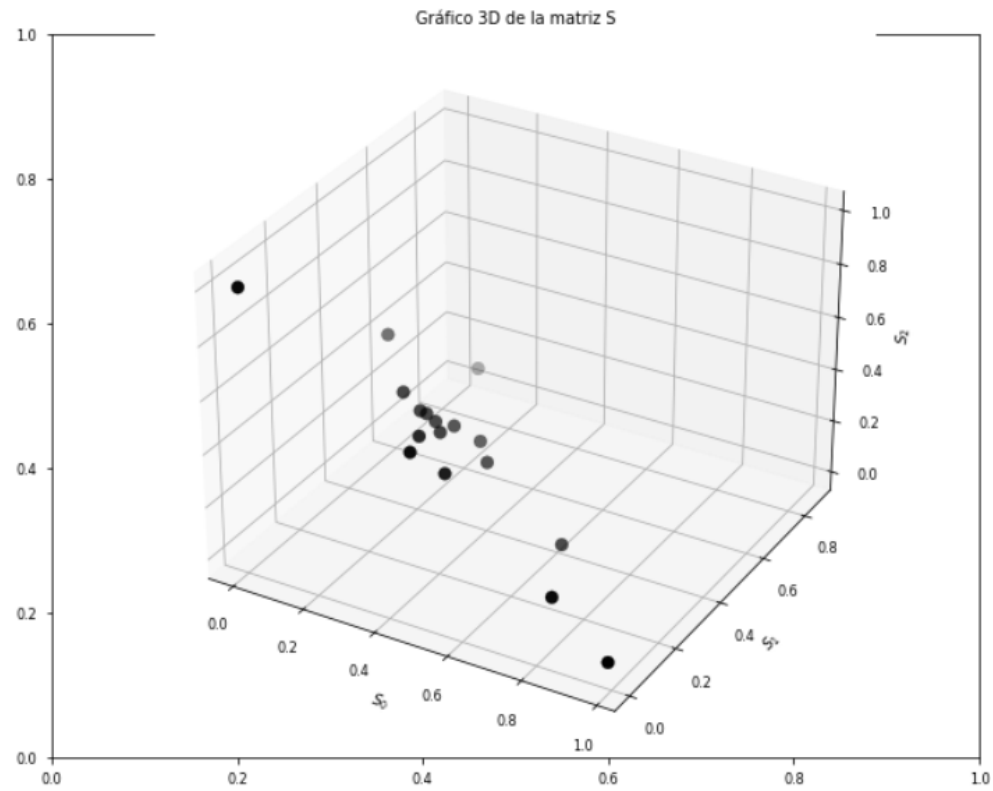


Figura 4.7: Visualización de los datos de la matriz S , caso 1.

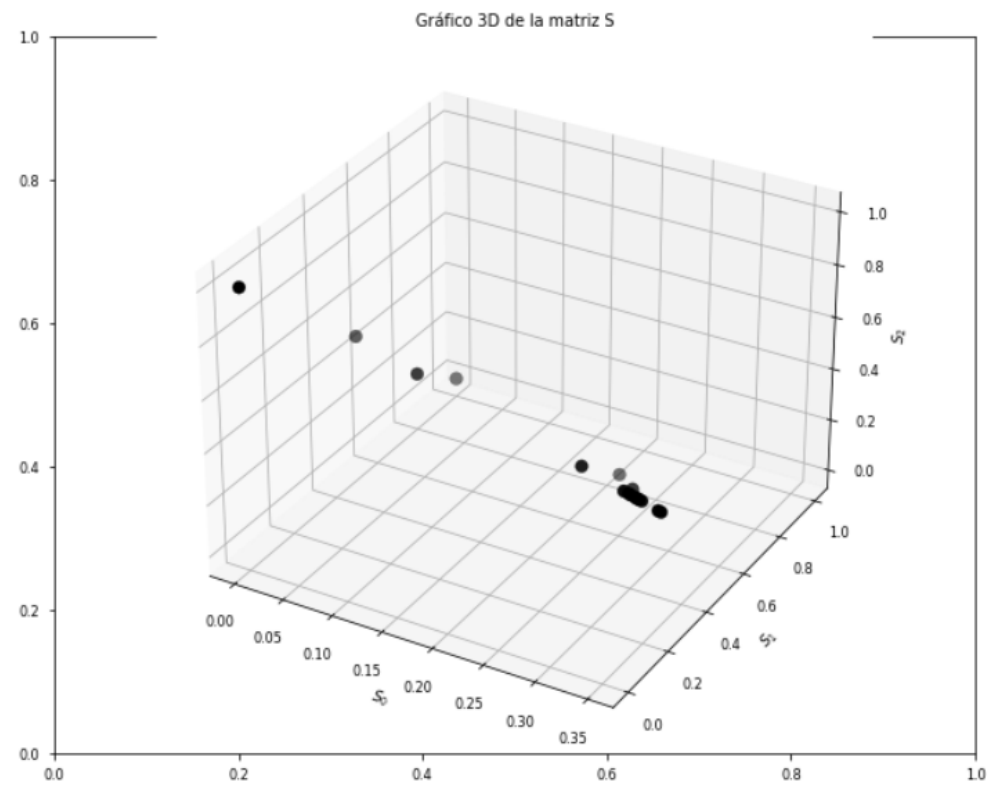


Figura 4.8: Visualización de los datos de la matriz S , caso 2.

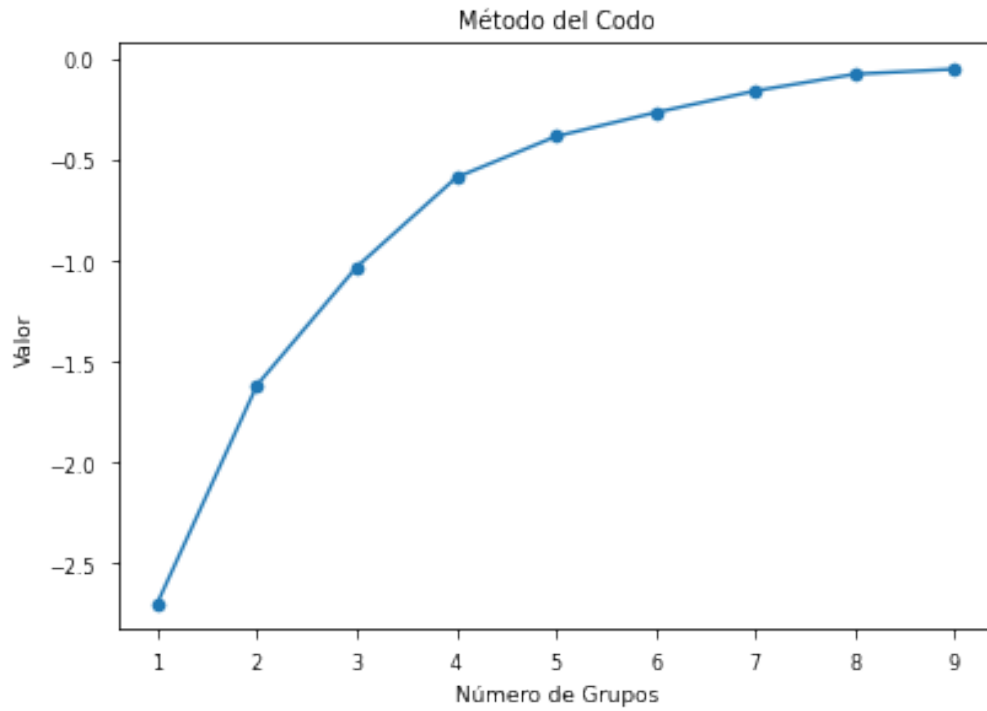


Figura 4.9: Método del codo para la asignación óptima de los grupos, caso 1.

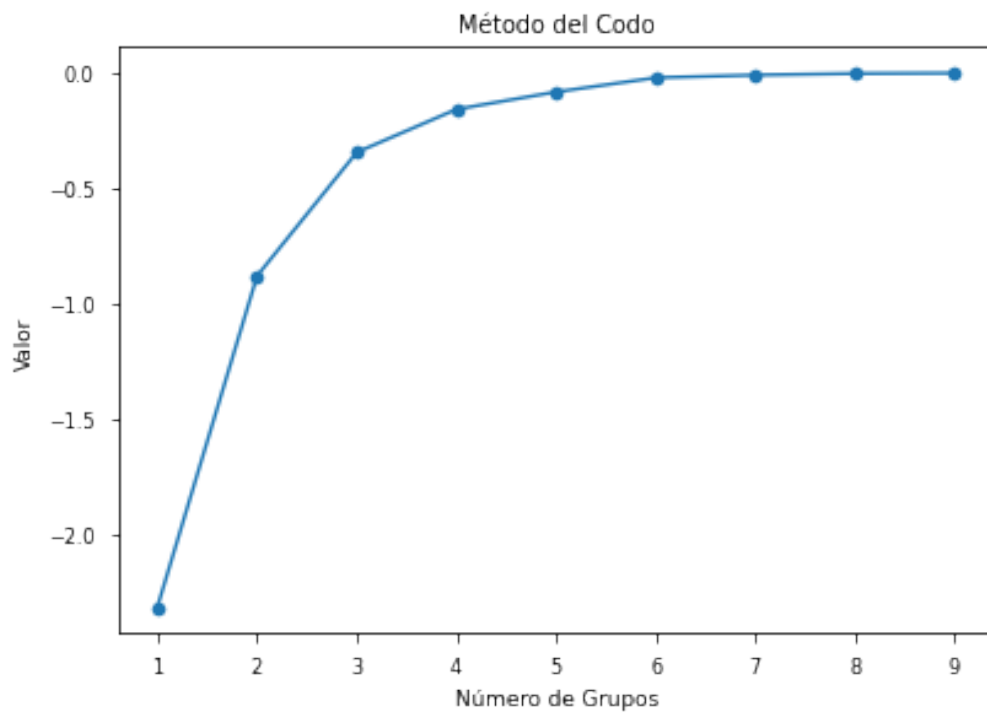


Figura 4.10: Método del codo para la asignación óptima de los grupos, caso 2.

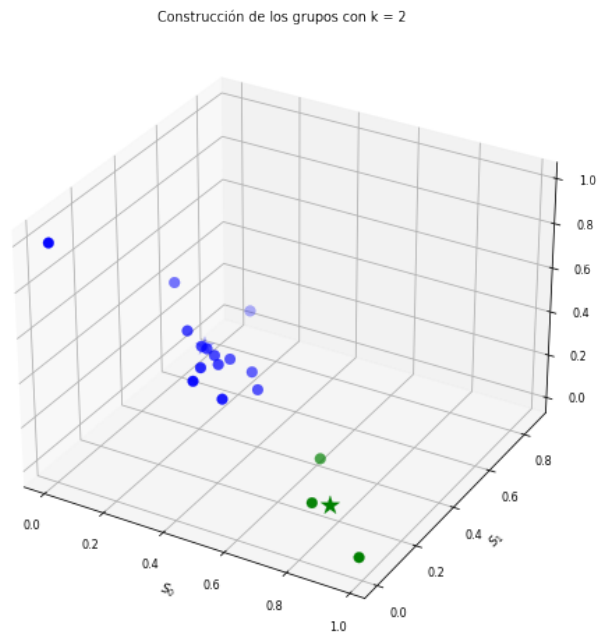


Figura 4.11: Visualización de los grupos con $k = 2$ según el algoritmo k -medias, caso 1.

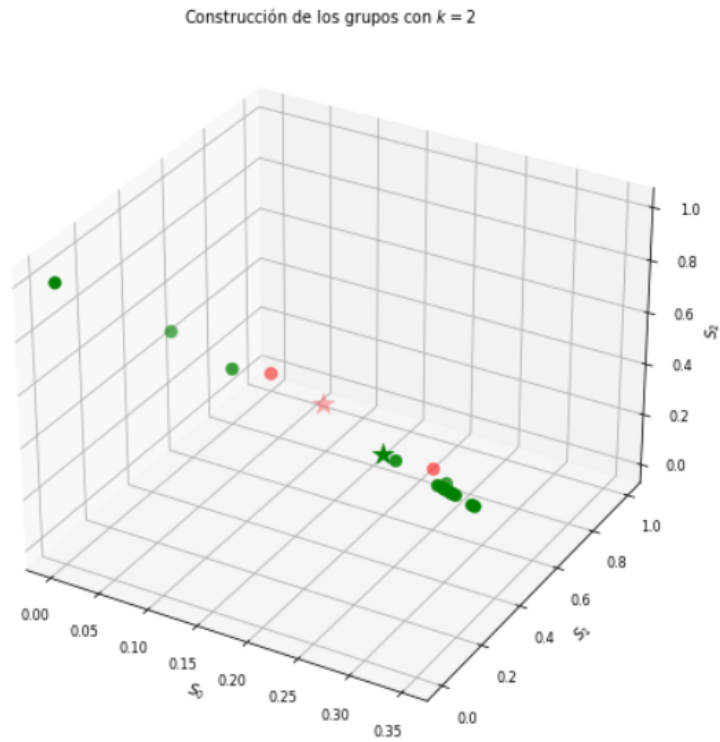


Figura 4.12: Visualización de los grupos con $k = 2$ según el algoritmo k -medias, caso 2.

mayor diferencia se evidencia en los estudiantes 0 y 7. A pesar de ello, estos estudiantes cambian de grupo al tomar un valor mayor para k .

Id Estudiante	Grupos Caso 1	Grupos Caso 2
0	0	1
1	1	1
2	0	0
3	1	1
4	1	1
5	1	1
6	1	1
7	1	0
8	1	1
9	1	1
10	1	1
11	1	1
12	1	1
13	1	1
14	1	1
15	1	1
16	0	0

Tabla 4.4: Agrupaciones de los estudiantes.

Por lo tanto, a pesar de que la representación inicial de los estudiantes según la matriz S en las figuras 4.7 y 4.8, no son semejantes, las distintas versiones aplicadas en el modelo preservan las mismas características subyacentes de los datos y se producen agrupaciones similares.

4.6. Análisis de cada grupo de estudiantes en la adquisición de conocimiento

Para el análisis de cada uno de los grupos y para una visualización más clara se muestran los estudiantes 2, 4 y 5 para el caso 1 y caso 2. Véase las figuras 4.13 y 4.14, respectivamente.

Según la figura 4.13, todos los estudiantes tienen un nivel de adquisición de conocimiento inicial mayor al final. El estudiante 5, tiene una disminución hasta el intento 3, luego, desde el intento 4 hasta el 11 presenta ganancias y pérdidas del conocimiento pero con menores variaciones de conocimiento que la de sus compañeros. Por otro lado, el estudiante 2, presenta cambios abruptos en su conocimiento desde el intento 2 al 9. Finalmente, su nivel de conocimiento queda por encima del estudiante 4 y 5. Por su parte, el estudiante 4 demuestra un comportamiento semejante al del estudiante 5.

De esta manera, se concluye que la gráfica en la figura 4.13 presenta un comportamiento contraintuitivo en la adquisición de conocimiento de los estudiantes intento tras intento siendo complicado realizar un análisis de la conducta de la adquisición de conocimiento. Por ejemplo, se puede comentar que los estudiantes pertenecientes a ambos grupos exhiben un comportamiento parecido durante los primeros 5 intentos, sin embargo, el estudiante 2 alcanzan niveles de conocimiento mayores y/o menores en los intentos 4 y 9, por su parte, los estudiantes del grupo 4 y 5 demuestran menor variabilidad con olvidos no tan significativos.

Por otro lado, en la figura 4.14 se muestra que para todos los estudiantes la adquisición de conocimiento final es mayor a la inicial; los estudiantes 4 y 5 comienzan con un conocimiento previo menor que el estudiante 2 pero rápidamente lo superan; quedando por encima hasta el final del curso. Para todos los estudiantes el mayor salto en la adquisición del conocimiento ocurre del intento 3 al 4.

De esta forma, y según una combinación de distintos factores como habilidades de aprendizaje, conocimiento previo del estudiante y motivación para enfrentar el curso, los estudiantes 4 y 5, del grupo 1, se pueden caracterizar como aprendices rápidos en el que su conocimiento adquirido presenta mayores crecimientos a lo largo del curso.

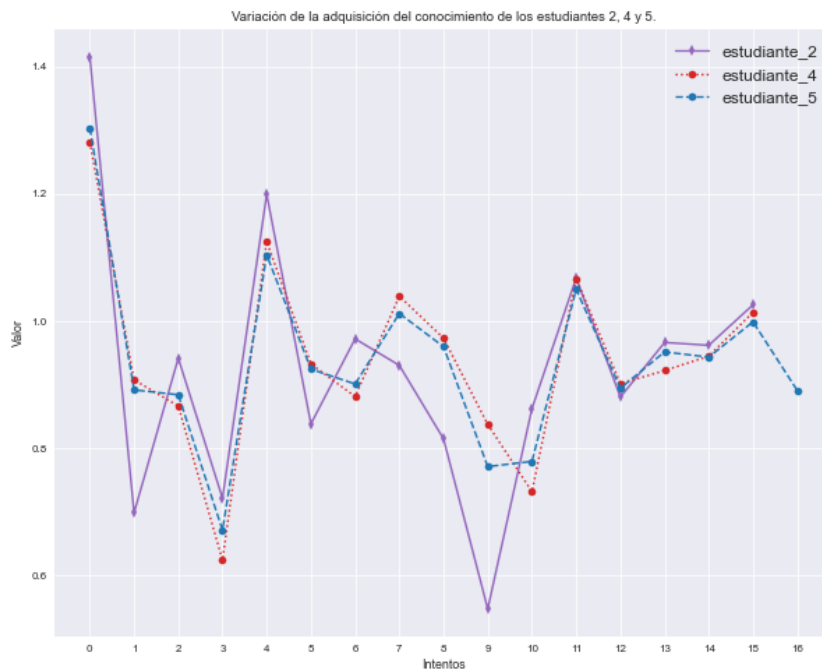


Figura 4.13: Conocimiento adquirido por los estudiantes en cada intento, caso 1.

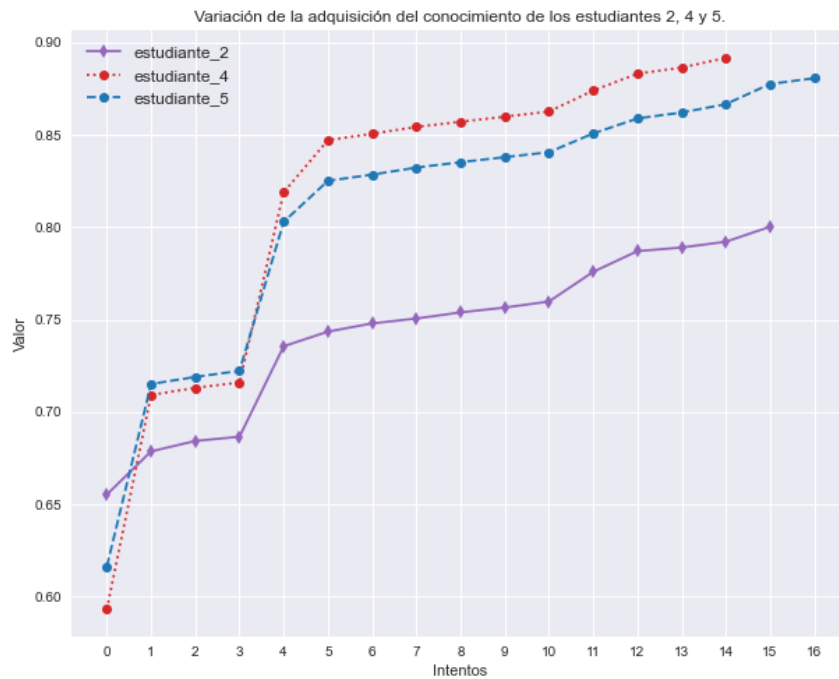


Figura 4.14: Conocimiento adquirido por los estudiantes en cada intento, caso 2.

4.7. Análisis del conocimiento adquirido por los estudiantes para cada concepto

Las figuras 4.15 y 4.16 contienen una gráfica por concepto en la que se visualiza el comportamiento del conocimiento promedio adquirido por los estudiantes en cada intento.

Analizando la figura 4.15, se observa nuevamente un comportamiento contraintuitivo a lo largo de los intentos ya que para todos los conceptos hay ganancias y pérdidas de conocimiento. El conocimiento adquirido promedio del concepto 4 es mayor que los demás conceptos y su conocimiento final posee un mayor valor. Los conceptos 0 y 1 presentan cambios en la adquisición de conocimiento con mayor frecuencia y pronunciación que los demás.

Mientras tanto, la figura 4.16, indica un crecimiento de conocimiento constante para cada concepto en cada intento. Se aprecia que en promedio el conocimiento de los estudiantes incrementa para todos los conceptos. El conocimiento de los estudiantes entre los intentos 0 y 1 muestra un aumento para los conceptos 0, 1, 2, y 4, mientras que, para el concepto 3 ocurre hasta el intento 3. El conocimiento inicial y final para los conceptos 1, 2 y 4 es mayor al de los conceptos 0 y 3.

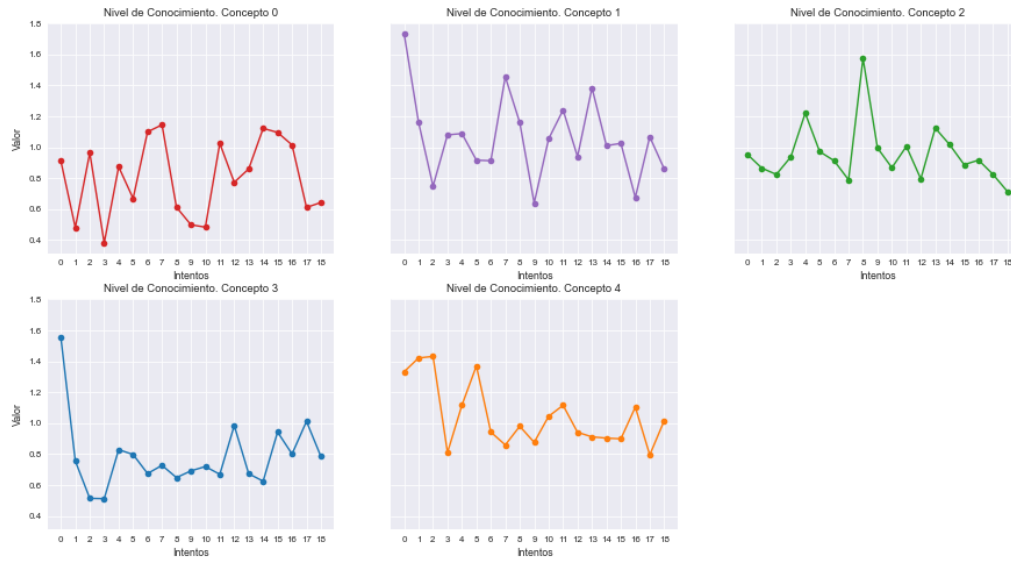


Figura 4.15: Conocimiento para cada concepto en cada intento, caso 1.

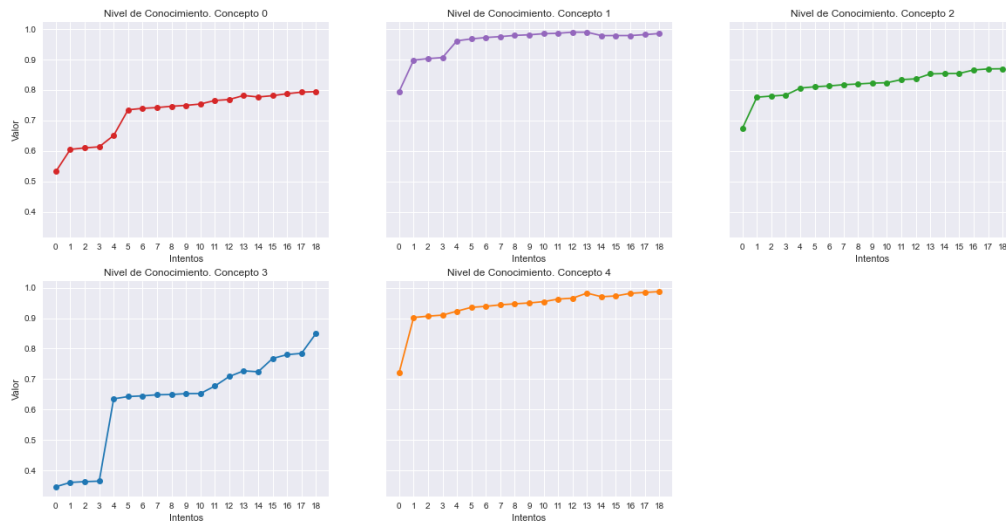


Figura 4.16: Conocimiento para cada concepto en cada intento, caso 2.

4.8. Análisis del conocimiento de los estudiantes para cada concepto

Dada las características contraintuitivas en el análisis del conocimiento adquirido por los estudiantes para cada concepto según el caso 1; en esta sección se propone realizar el análisis únicamente para caso 2 ya que se desea entender el comportamiento de cada grupo de estudiantes para

cada concepto.

Ahora, en la figura 4.17 se contrasta el conocimiento adquirido en cada una de los conceptos para cada grupo. Se observa que los estudiantes del grupo 0 finalizaron el conocimiento adquirido en cada concepto por debajo que los estudiantes del grupo 1. Sin embargo, su conocimiento para los conceptos 0 y 1 se posicionó por encima del grupo 1; pero, el conocimiento fue igualado y superado por el grupo 0 a partir del intento 4.

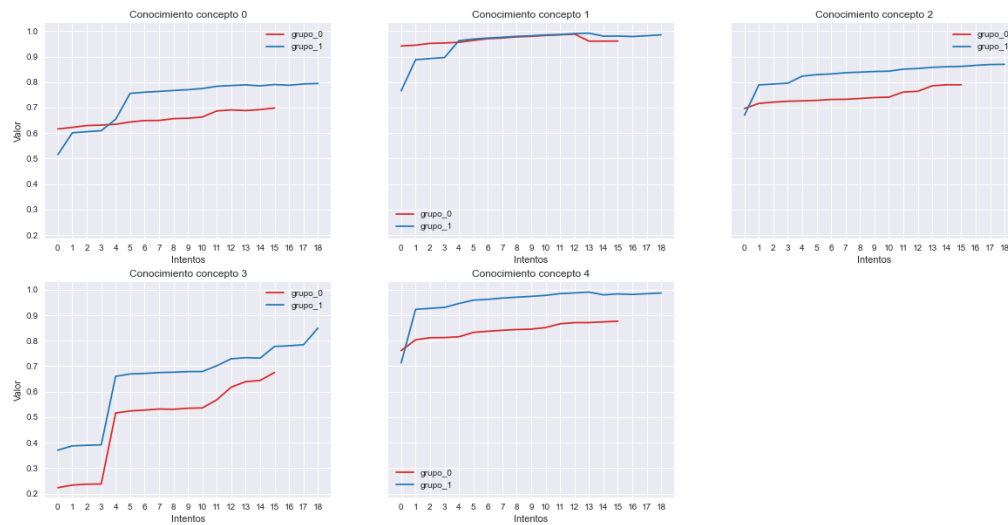


Figura 4.17: Conocimiento para cada concepto en cada intento por cada grupo, caso 2.

Finalmente, se puede concluir que al aplicar el modelo MVKM minimizando únicamente la función \mathcal{L}_1 , caso 1, se obtienen resultados complejos de comprender y analizar generando conclusiones contraintuitivas, a pesar de esto, el modelo consigue conservar las características subyacentes de los datos. Esto se verifica ya que la agrupación de los estudiantes mediante sus habilidades de aprendizaje y el aporte logrado por cada recurso de aprendizaje es similar en ambos casos. Por su parte, al aplicar el modelo MVKM penalizando el olvido de los estudiantes por medio de la matriz T , caso 2, se logran resultados que permiten llegar a conclusiones más acertadas y relevantes.

CONCLUSIONES

La web, tal como yo la imaginaba, todavía
no la hemos visto. El futuro sigue siendo
mucho más grande que el pasado

Tim Berners-Lee

Después de aplicar el modelo MVKM a los datos del curso CO6612 utilizando los métodos mencionados y explicados con anterioridad; se concluye que la aplicación del modelo logró ajustarse a los datos del curso obteniéndose predicciones razonablemente buenas según la literatura.

También se observó que al emplear la penalización del olvido de conocimiento sobre la función \mathcal{L}_1 -cuando no se cumple que la diferencia de adquisición de conocimiento entre los tiempos $a+1$ y a sea mayor que 0 para cada una de las entradas de la matriz $T_{a+1} - T_a$ - se consiguieron resultados sencillos de analizar en comparación a los resultados logrados cuando solo se minimizó la función \mathcal{L}_1 , pero también se observó que al minimizar únicamente la función \mathcal{L}_1 , se conservaron las características subyacentes de los datos y para ambos casos se consiguieron resultados similares.

Los estudiantes se agruparon bajo 2 conglomerados contruidos según sus habilidades de aprendizaje; en donde, el grupo conformado por los estudiantes 2, 7 y 16 se distingue por una adaptación y asimilación de nuevos conceptos de una manera más lenta que el resto de sus compañeros. El otro grupo, alcanzó mayores niveles de adquisición de conocimiento. Dicho comportamiento, se repite en el análisis del conocimiento adquirido en cada uno de los conceptos.

En última instancia, se analizaron los aportes de cada recurso al conocimiento y la correspondencia entre los conceptos y los recursos realizados por el modelo MVKM. Se identificó que los recursos que más contribuyeron fueron el QE PM, QE Proyecto Final seguido de las tareas. Mientras que, para la correspondencia entre los conceptos y los recursos se observó que sigue las relaciones naturales al momento de aprender los conceptos dictados en el curso.

El código que contiene las funciones y programas usados para realizar los cálculos y la minimización de la función objetivo fue elaborado en el lenguaje de programación Python. Dicho código se encuentra almacenado en un repositorio en la nube que puede ser consultado a través del siguiente enlace: <https://github.com/Leojmartin/MVKM.git>.

REFERENCIAS

- [1] S. Zhao, C. Wang y S. Sahebi. “Modeling knowledge acquisition from multiple learning resource types”. En: *arXiv preprint arXiv:2006.13390* (2020).
- [2] T. Mitchell. *Machine Learning*. McGraw-Hill Science/Engineering/Math, 1997.
- [3] D. Michie, D. J Spiegelhalter y C. C. Taylor. “Machine learning, neural and statistical classification”. En: (1994).
- [4] J. Brownlee. *Master Machine Learning Algorithms. Discover How They Work and Implement Them From Scratch*. 2016.
- [5] J. Gareth, W. Daniela, H. Trevor y T. Robert. *An introduction to statistical learning: with applications in R*. Springer, 2013.
- [6] S. Boyd, S.P. Boyd y L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [7] Chayakrit Krittanawong, Kipp W Johnson, Robert S Rosenson, Zhen Wang, Mehmet Aydar, Usman Baber, James K Min, W H Wilson Tang, Jonathan L Halperin y Sanjiv M Narayan. “Deep learning for cardiovascular medicine: a practical primer”. En: *European Heart Journal* 40.25 (feb. de 2019), págs. 2058-2073. ISSN: 0195-668X. DOI: 10.1093/eurheartj/ehz056. eprint: <https://academic.oup.com/eurheartj/article-pdf/40/25/2058/28882900/ehz056.pdf>.
- [8] S. Haykin. *Neural networks and learning machines, 3/E*. Pearson Education, India, 2009.
- [9] Thanh-Nam Doan y S. Sahebi. “Rank-based tensor factorization for student performance prediction”. En: *12th International Conference on Educational Data Mining (EDM)*. 2019.
- [10] J. González-Brenes. “Modeling skill acquisition over time with sequence and topic modeling”. En: *Artificial Intelligence and Statistics*. PMLR. 2015, págs. 296-305.
- [11] S.S. Shai y B. D. Shai. *Understanding Machine Learning from Theory to Algorithms*. Cambridge University Press, 2014.