Propuesta de Trabajo Fin de Grado en Matemáticas (curso 2021-2022)

Responsable de tutorización: Julián Luengo Martín
Departamento: Ciencias de la Computación e Inteligencia Artificial
Correo electrónico: julianlm@ugr.es

Responsable de cotutorización:
Departamento:
Correo electrónico:

(Rellenar sólo en caso de que la propuesta esté realizada a través de un estudiante)
Estudiante que propone el trabajo:

Título del trabajo: Descomposición analítica de matrices kernel para aprendizaje máquina

Tipología del trabajo (marcar una o varias de las siguientes casillas):

☐ Complementario de profundización
☐ Divulgación de las Matemáticas
☐ Docencia e innovación
☒ Herramientas informáticas
☒ Iniciación a la investigación

Materias del grado relacionadas con el trabajo:
Informática I
Informática II
Métodos numéricos I y II
Análisis Matemático I
Geometría I y II

Descripción y resumen de contenidos:

En el aprendizaje automático, se utilizan las funciones kernel para realizar transformaciones en el espacio de entrada con el objetivo de facilitar la separabilidad o el agrupamiento de diferentes puntos de interés en el espacio [Sch2018]. Gracias al Teorema de Cover, las probabilidades de mejorar la separabilidad de los puntos en una alta dimensionalidad son favorables [Cov65].
En la Figura 1 se muestra un ejemplo de transformación no lineal que permite separar puntos de diferentes tipologías gracias al uso de funciones kernel (de https://people.eecs.berkeley.edu/~jordan/courses/281B-spring04/lectures/lec3.pdf)

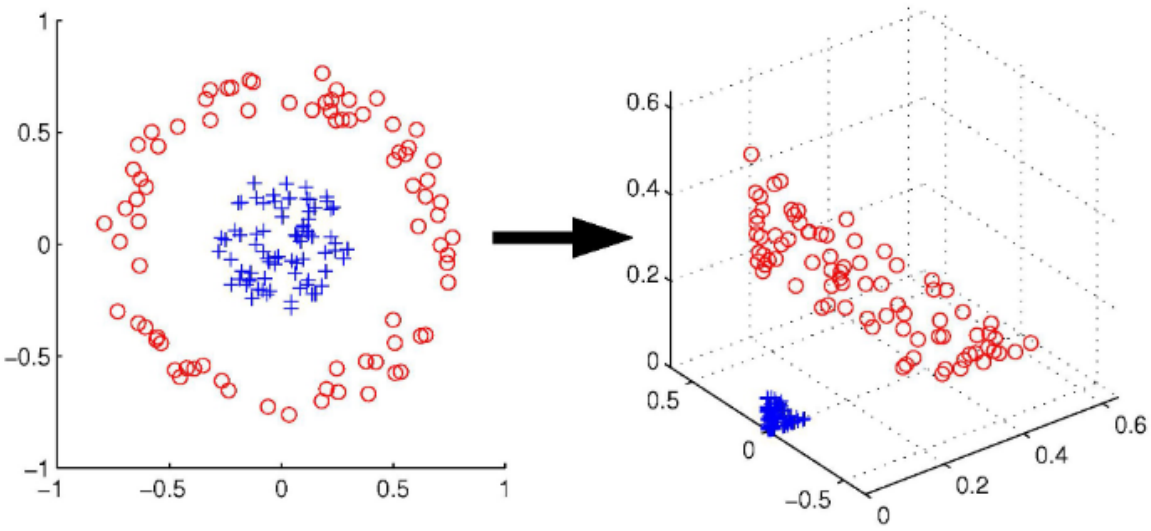

Figure 1: Transforming the data can make it linearly separable

Existen múltiples funciones kernel bien estudiadas que permiten trabajar a métodos de aprendizaje automático (como las Máquinas de vectores soporte -SVM-) aplicando funciones de separabilidad sencillas (como hiperplanos) en espacios transformados con buenos resultados. Esto es debido a:

1. El resultado de la función de separabilidad se puede expresar como el producto escalar de los puntos transformados.
2. Sólo algunos de los puntos de interés son necesarios para realizar la separación.

En computación, suele expresarse en forma de matriz el resultado de aplicar el producto escalar entre cada par de puntos muestreados en el espacio transformado, generando la matriz kernel. Una matriz kernel es válida sí y sólo sí es semidefinida positiva.

En la realidad, y debido al auge del Big Data, no es posible en la mayoría de los casos representar la matriz kernel para todas las parejas de puntos, ya que no cabe en la memoria de un solo computador. Existen estrategias de creación de submatrices que, a su vez, generan pequeñas superficies de decisión para diferentes subregiones del espacio de entrada [Gra2004]. Estas subsoluciones son unificadas de formas sencillas, usando métodos de agregación simples (medias o votaciones).

El objetivo de este TFG es explorar los casos y condiciones (si existen) en los cuales se pueden realizar divisiones iterativas o paralelizadas de los datos (colecciones de puntos en el espacio de entrada original) que puedan generar submatrices que constituyan una matriz de kernel válida. A partir del teorema de Mercer [Mer1909], es conocido que las matrices de Gram son semidefinidas positivas y siempre válidas como funciones kernel. Existen propuestas empíricas que se acercan a este concepto [Jia2005, Do2006, Alh2013], pero sin fundamentación teórica de funcionamiento para matrices completas.

Actividades a desarrollar:

1. Analizar de forma teórica la posibilidad de descomposición y reconstrucción de la matriz kernel a partir de submatrices (dictaminadas por los subconjuntos de puntos usados).

2. Evaluar las condiciones de separación y reconstrucción que puedan ser óptimas.
3. Evaluar de forma empírica (al menos) las hipótesis formuladas en los puntos anteriores. Esto implica realizar ajustes a métodos de análisis numérico en el cálculo de las matrices y el ajuste de los hiperplanos en algunos algoritmos de aprendizaje automático (SVMs).

Objetivos matemáticos planteados
Para algunas funciones kernel conocidas, estudiar si cualquier descomposición de los datos permite obtener submatrices (bloques) válidos (positivas semidefinidas).
Demostrar si cualquier subdivisón en submatrices permite generar una matriz kernel completa o no para las funciones conocidas.
Analizar las condiciones generales de funciones kernels conocidas o no que permiten la reconstrucción de la matriz kernel completa a partir de submatrices.
Evaluar qué funciones kernel son más susceptibles de ser divididas o no de forma empírica y, a ser posible, teórica.

Bibliografía para el desarrollo matemático de la propuesta:

[Mer1909] Mercer, J. (1909), "Functions of positive and negative type and their connection with the theory of integral equations", Philosophical Transactions of the Royal Society A, 209 (441–458): 415–446
[Cov65] T. M. Cover, "Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition," in *IEEE Transactions on Electronic Computers*, vol. EC-14, no. 3, pp. 326-334, June 1965, doi: 10.1109/PGEC.1965.264137.

Otras referencias (si procede):

[Jia2005] Jian-xiong Dong, A. Krzyzak and C. Y. Suen, "Fast SVM training algorithm with decomposition on very large data sets," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 4, pp. 603-618, April 2005
[Do2006] Do, T. N., & Poulet, F. (2006, February). Classifying one billion data with a new distributed svm algorithm. In *RIVF* (pp. 59-66).
[Gra2004] Graf, H., Cosatto, E., Bottou, L., Dourdanovic, I., & Vapnik, V. (2004). Parallel support vector machines: The cascade svm. *Advances in neural information processing systems*, *17*, 521-528.
[Alh2013] Alham, N. K., Li, M., Liu, Y., & Qi, M. (2013). A MapReduce-based distributed SVM ensemble for scalable image classification and annotation. *Computers & Mathematics with Applications*, *66*(10), 1920-1934.
[Sch2018] Schölkopf, B.; Smola, A. J.; Bach, F. (2018). Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond

Firma del estudiante
(solo para trabajos propuestos por alumnos)

Firma del responsable de tutorización
(solo para trabajos propuestos por estudiantes)

Firma del responsable de cotutorización
(solo para trabajos propuestos por estudiantes)

En, Granada, a 21 de mayo de 2021