Aprenentatge Computacional

MD3: Kaggle

GPU Kernel
Performance Dataset



Mario González 1566235



Índice de contenidos

- Explicación de la BBDD
- Análisis de los datos
- Preprocesado de los datos
- Modelos y búsqueda del número óptimo de variables
- Conclusiones

Explicación de la BBDD

Explicación de la BBDD

- Mesures de temps d'execució de multiplicacions de matrius
 2048*2048 en un kernel parametrizable
- 241600 posibles combinaciones de parámetros

Atributos de la BBDD

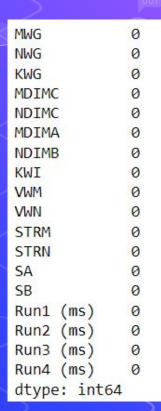
Tiempo de ejecución	Atributos parametrizables del kernel
Run (ms) (integer)	 MWG, NWG: {16, 32, 64, 128} (integer) KWG: {16, 32} (integer) MDIMC, NDIMC: {8, 16, 32} (integer) MDIMA, NDIMB: {8, 16, 32} (integer) KWI: {2, 8} (integer) VWM, VWN: {1, 2, 4, 8} (integer) STRM, STRN: {0, 1} (categorical) SA, SB: {0, 1} (categorical)

Análisis de los datos

Busqueda de valores nulos

No disponemos de valores NaN o null:

	MWG	NWG	KWG	MDIMC	NDIMC	MDIMA	NDIMB	KWI	VWM	VWN	STRM	STRN	SA	SB	Run1 (ms)	Run2 (ms)	Run3 (ms)	Run4 (ms)
0	16	16	16	8	8	8	8	2	1	1	0	0	0	0	115.26	115.87	118.55	115.80
1	16	16	16	8	8	8	8	2	1	1	0	0	0	1	78.13	78.25	79.25	79.19
2	16	16	16	8	8	8	8	2	1	1	0	0	1	0	79.84	80.69	80.76	80.97
3	16	16	16	8	8	8	8	2	1	1	0	0	1	1	84.32	89.90	86.75	85.58
4	16	16	16	8	8	8	8	2	1	1	0	1	0	0	115.13	121.98	122.73	114.81



Preprocesado de los datos

Transformación del target

○ 4 ejecuciones independientes → 1 sola (media)

	MWG	NWG	KWG	MDIMC	NDIMC	MDIMA	NDIMB	KWI	VWM	VWN	STRM	STRN	SA	SB	Run (ms)
0	16	16	16	8	8	8	8	2	1	1	0	0	0	0	116.3700
1	16	16	16	8	8	8	8	2	1	1	0	0	0	1	78.7050
2	16	16	16	8	8	8	8	2	1	1	0	0	1	0	80.5650
3	16	16	16	8	8	8	8	2	1	1	0	0	1	1	86.6375
4	16	16	16	8	8	8	8	2	1	1	0	1	0	0	118.6625

Eliminacion de outliers

Objetivo: reducir la dispersion de valores

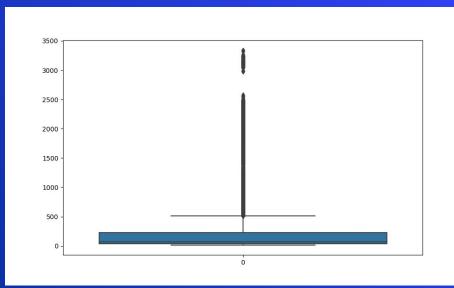
 Descartamos aquellos valores que mas se alejan de los más repetidos

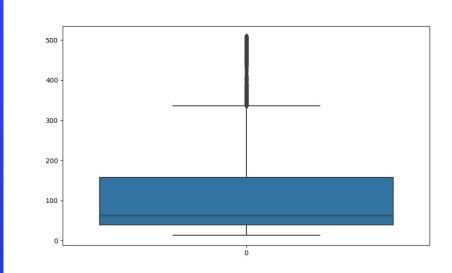
Eliminacion de outliers

001

Antes

Después





Matriz de correlación

Observación de la matriz de correlación

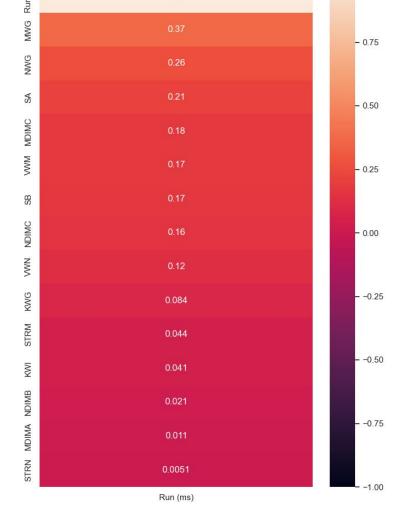
Atributos con una mayor

correlación: MWG, NWG, SA, VWN

																	- 1.0
MWG	1	0.0006	0.0093	0.11	-0.0086	0.16	0.015	-4.8e-15	0.35	-0.00084	-6e-15	-3.7e-15	-6.7e-15	-8.2e-20	0.35		
NWG	0.0006	1	0.0093	-0.0086	0.11	0.015	0.16	-2.3e-16	-0.00084	0.35	-1.9e-15	-1.5e-15	-4.6e-15	5.7e-20	0.32		
KWG	0.0093	0.0093	1	0.15	0.15	-0.035	-0.035	-7.2e-16	-0.012	-0.012	-6.9e-16	-1.1e-15	-3.1e-15	2.3e-19	0.011		- 0.8
MDIMC	0.11	-0.0086	0.15	1	-0.21	0.2	0.085	3.7e-16	-0.13	0.011	-6.5e-16	-6.4e-16	-1.1e-15	-3.7e-21	-0.22		
NDIMC	-0.0086	0.11	0.15	-0.21	1	0.085	0.2	2.5e-16	0.011	-0.13	-2.1e-16	-2.8e-16	-1.1e-15	1.7e-19	-0.21		– o.€
MDIMA	0.16	0.015	-0.035	0.2	0.085	1	0.088	9.3e-16	-0.2	-0.019	-7.2e-16	-2.8e-16	-8.1e-16	-2.7e-20	-0.007		0.0
NDIMB	0.015	0.16	-0.035	0.085	0.2	0.088	1	7.6e-17	-0.019	-0.2	-3.7e-16	-2.5e-16	-6.3e-16	-8.1e-20	-0.0087		
KWI	-4.8e-15	-2.3e-16	-7.2e-16	3.7e-16	2.5e-16	9.3e-16	7.6e-17	1	-8.6e-16	-1.3e-16	-5.8e-18	-9.1e-18	-1.2e-17	1.4e-19	0.033		- o.4
WWW	0.35	-0.00084	-0.012	-0.13	0.011	-0.2	-0.019	-8.6e-16	1	0.0012	-1.9e-15	-1e-15	-2.2e-15	-6.4e-20	0.16		
NWN	-0.00084	0.35	-0.012	0.011	-0.13	-0.019	-0.2	-1.3e-16	0.0012	1	-6.2e-16	-4.7e-16	-1.5e-15	1.9e-19	0.14		– o.:
STRM	-6e-15	-1.9e-15	-6.9e-16	-6.5e-16	-2.1e-16	-7.2e-16	-3.7e-16	-5.8e-18	-1.9e-15	-6.2e-16	1	-6.4e-19	-4.9e-19	-1e-19	-0.013		
STRN	-3.7e-15	-1.5e-15	-1.1e-15	-6.4e-16	-2.8e-16	-2.8e-16	-2.5e-16	-9.1e-18	-1e-15	-4.7e-16	-6.4e-19	1	-4.1e-19	-1.4e-19	-0.00011		
SA	-6.7e-15	-4.6e-15	-3.1e-15	-1.1e-15	-1.1e-15	-8.1e-16	-6.3e-16	-1.2e-17	-2.2e-15	-1.5e-15	-4.9e-19	-4.1e-19	1	1.3e-19	0.052		- 0.0
88	-8.2e-20	5.7e-20	2.3e-19	-3.7e-21	1.7e-19	-2.7e-20	-8.1e-20	1.4e-19	-6.4e-20	1.9e-19	-1e-19	-1.4e-19	1.3e-19	1	0.064		
Run (ms)	0.35	0.32	0.011	-0.22	-0.21	-0.007	-0.0087	0.033	0.16	0.14	-0.013	-0.00011	0.052	0.064	1		(
	MWG	NWG	KWG	MDIMC	NDIMC	MDIMA	NDIMB	KWI	VWM	VWN	STRM	STRN	SA	SB	Run (ms)		

Matriz de correlación

 Representación de la correlación de cada atributo respecto al target



Escalado de los datos

 Creamos un dataset reducido con los atributos con una mayor correlación: MWG, NWG, SA, VWN

Valores:

	MWG	NWG	SA	VWN	Run (ms)
count	214833.000000	214833.000000	214833.000000	214833.000000	214833.000000
mean	75.688037	75.761619	0.486038	2.345156	114.554350
std	41.968313	41.997377	0.499806	1.862122	113.825481
min	16.000000	16.000000	0.000000	1.000000	13.317500
25%	32.000000	32.000000	0.000000	1.000000	39.095000
50%	64.000000	64.000000	0.000000	2.000000	61.790000
75%	128.000000	128.000000	1.000000	4.000000	157.892500
max	128.000000	128.000000	1.000000	8.000000	509.962500

Escalado de los datos

 Creamos un dataset reducido con los atributos con una mayor correlación: MWG, NWG, SA, VWN

Valores:

	MWG	NWG	SA	VWN	Run (ms)
count	214833.000000	214833.000000	214833.000000	214833.000000	214833.000000
mean	75.688037	75.761619	0.486038	2.345156	114.554350
std	41.968313	41.997377	0.499806	1.862122	113.825481
min	16.000000	16.000000	0.000000	1.000000	13.317500
25%	32.000000	32.000000	0.000000	1.000000	39.095000
50%	64.000000	64.000000	0.000000	2.000000	61.790000
75%	128.000000	128.000000	1.000000	4.000000	157.892500
max	128.000000	128.000000	1.000000	8.000000	509.962500

Escalado de los datos

 Mediante el escalado de los datos, los dejamos todos en una escala parecida.

Valores:

```
0.01595121]
[[-1.42222005 -1.42298787 -0.97245519 -0.7223801
 [-1.42222005 -1.42298787 -0.97245519 -0.7223801
                                                  -0.31495086
 [-1.42222005 -1.42298787 1.02832501 -0.7223801
                                                  -0.29861002]
   1.24646624
              1.24385153 -0.97245519
                                      0.8886892
                                                  -0.69751108
                                                  -0.75643915]
   1.24646624
              1.24385153 1.02832501
                                       0.8886892
                                                  -0.84987187]]
  1.24646624
              1.24385153 1.02832501
                                       0.8886892
```

Modelos y búsqueda de hiperparámetros



Selección de modelos

- Para abordar la regresion haremos uso de los modelos:
 - · Linear Regression
 - Decision Tree
 - SGD Regressor

Entrenamiento de modelos

- Rendimiento con las 4 mejores variables (correlación):
 - Score
 - MSE

Variables: MWG, NWG, SA, VWN

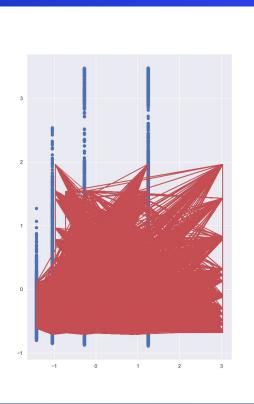
Entrenamiento de modelos - 4 mejores atributos

Resultados:

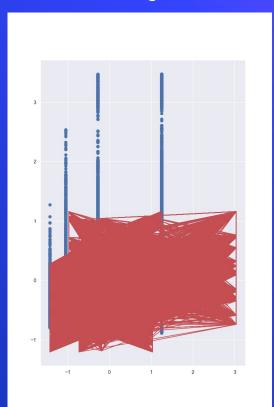
Modelo	Score	MSE
Linear Regression	0.272	0.724
Decision Tree	0.336	0.665
SGD Regressor	0.339	0.656

Predicciones de los modelos - 4 mejores atributos

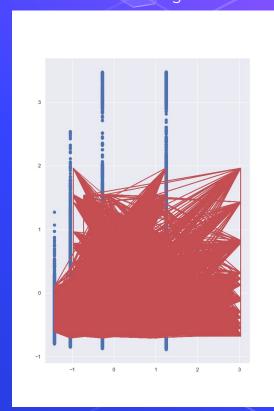
Decision Tree



Linear Regression



SGD Regressor



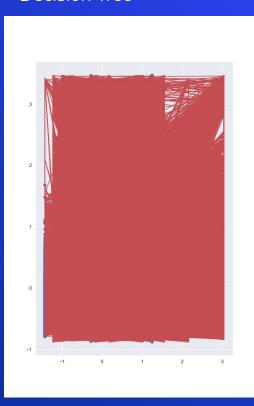
Entrenamiento de modelos - todos los atributos

Resultados:

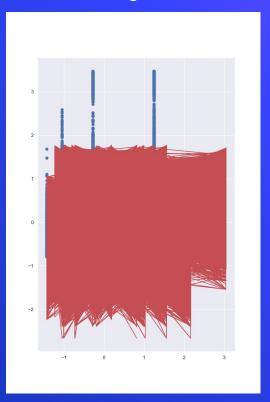
Modelo	Score	MSE
Linear Regression	0.469	0.530
Decision Tree	0.999	0.0006
SGD Regressor	0.467	0.531

Predicciones de los modelos - todos los atributos

Decision Tree



Linear Regression



SGD Regressor



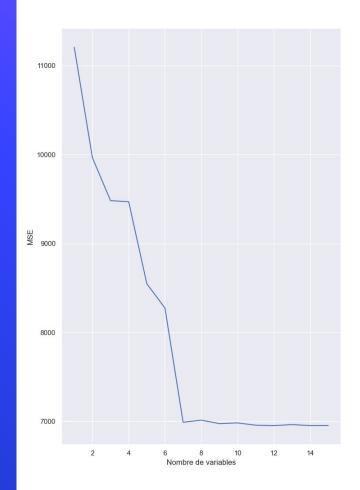
Entrenamiento de modelos - busqueda óptima de variables

○ Minimizar el nº de variables → overfitting

○ Minimizar MSE → mejores predicciones

Entrenamiento de modelos - busqueda óptima de variables

Estancamiento MSE al utilizar mas de 7 variables



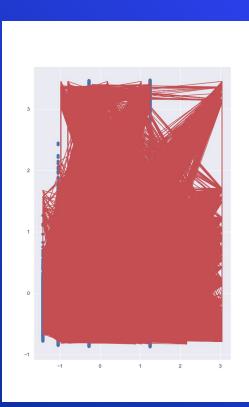
Entrenamiento de modelos numero óptimo de variables

Resultados:

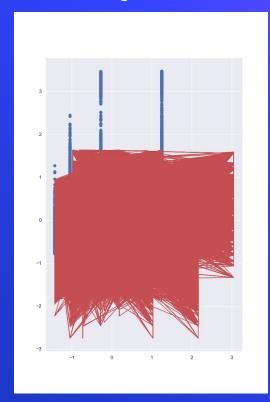
Modelo	Score	MSE
Linear Regression	0.457	0.537
Decision Tree	0.935	0.063
SGD Regressor	0.457	0.538

Predicciones de los modelos – nº optimo de atributos

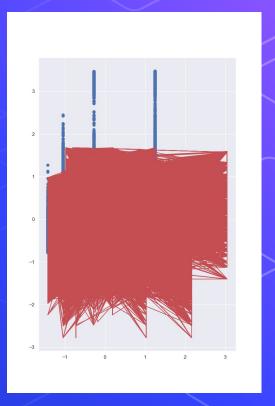
Decision Tree



Linear Regression



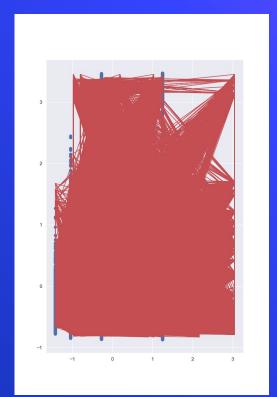
SGD Regressor



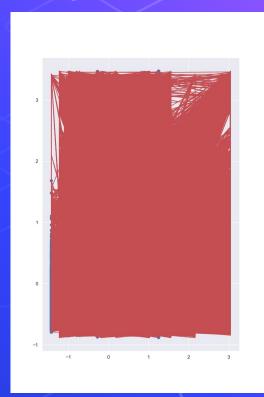
Modelos - Decision tree

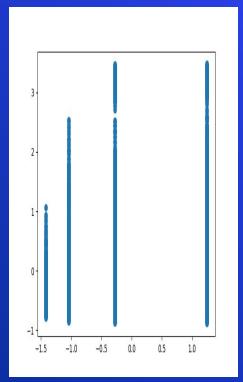
Valores test

Predicción nº óptimo de variables



Predicción con todas las variables





Conclusiones