Fundação Vanzolini

Dominando Big Data com o uso de Plataformas Gratuitas (nível intermediário)

Aula 2

LexisNexis® RISK SOLUTIONS

RELX Group

PRO

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO

USP UNIVERSIDADE DE SÃO PAULO

Bem-vindo! – Agenda da aula 2

- ✓Desafio Lending Club
- ✓Introdução ao Machine Learning
- ✓Intervalo (20 min)
- ✓Tutorial de preparação de dados

Exercício prático:

Faça a extração do dataset do Lending Club

- Spray
- Estrutura RECORD
- Declaração DATASET

Introdução ao Machine Learning

O que é Machine Learning?

- *"O estudo científico de algoritmos e modelos estatísticos que sistemas de computador usam para realizar uma tarefa específica sem usar instruções explícitas, baseando-se em padrões e inferência"*

- **Supervisionado -** quando apresentamos ao algoritmo dados de entrada e as respectivas saídas

- **Não supervisionado -** quando apresentamos somente os dados de entrada e o algoritmo descobre as saídas

- **Por reforço, profundo, etc -** o algoritmo utiliza tentativa e erro para encontrar uma solução para o problema, múltiplas camadas de aprendizado com dados complexos (imagens, vídeo, áudio), etc

Terminologia de ML

- Exemplo de aprendizado supervisionado:

  Dada uma amostra de registros:

```
Record1: Field1, Field2, Field3, … , FieldM
Record2: Field1, Field2, Field3, … , FieldM
…
RecordN: Field1, Field2, Field3, …, FieldM
```

Variáveis “Independentes”

  E um conjunto de valores a serem determinados,

```
Record1: TargetValue
Record2: TargetValue
…
RecordN: TargetValue
```

Variáveis “Dependentes”

  Aprenda a predizer valores para novas amostras.

Exemplo prático de ML

- Dado o conjunto de dados sobre árvores em uma floresta:

Altura	Diâmetro	Altitude	Pluviosidade	Idade
50	8	5000	12	80
56	9	4400	10	75
72	12	6500	18	60
47	10	5200	14	53

- Obtenha um modelo que determine a idade de uma árvore (variável dependente) a partir da sua altura, diâmetro, altura e pluviosidade do local (variáveis independentes).

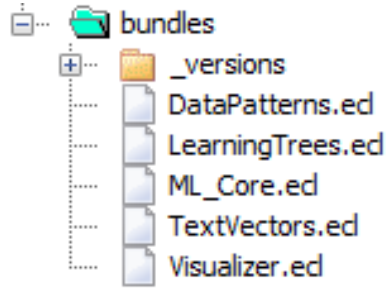
Fluxo de aprendizagem de máquina

1. Definição do problema
2. Extração dos dados
3. Preparação dos dados
4. Conversão e segregação das variáveis
5. Treinamento e avaliação do modelo
6. Implantação do modelo

Aprendizado de máquina no HPCC Systems

- Bundle validado, suportado e otimizado para desempenho na plataforma (https://hpccsystems.com/download/free-modules/machine-learning-library)

- Processo de instalação:
  - Fácil e independente da versão da plataforma
  - *ecl bundle install https://github.com/hpcc-systems/<nome>.git*



- Curso online:
  - https://learn.lexisnexis.com/Activity/2553

Bundle de ML

- Base:
  - ML_Core: Machine Learning Core (https://github.com/hpcc-systems/ML_Core.git)
  - PBblas: Paralell Block Basic Linear Algebra Subsystem (https://github.com/hpcc-systems/PBblas.git)

- Algoritmos supervisionados
  - LinearRegression: OLS (https://github.com/hpcc-systems/LinearRegression.git)
  - LogisticRegression: binomial/multinomial (https://github.com/hpcc-systems/LogisticRegression.git)
  - GLM: General Linear Model (https://github.com/hpcc-systems/GLM.git)
  - SVM: Support Vector Machines (https://github.com/hpcc-systems/SupportVectorMachines.git)
  - LearningTrees: Árvores de decisão (https://github.com/hpcc-systems/LearningTrees.git)

Bundle de ML (cont.)

- Algoritmos não-supervisionados
  - K-Means: clusterização de Big Data (https://github.com/hpcc-systems/KMeans.git)
  - DBSCAN: Scalable Paralell Density-Based Spatial Clustering of Applications with Noise (https://github.com/hpcc-systems/dbscan.git)
  - TextVectors: Vetorização de palavras, frases e sentenças (https://github.com/hpcc-systems/TextVectors.git)

- Aprendizagem profunda
  - GNN: Generalized Neural Network (https://github.com/hpcc-systems/GNN.git)

Tutorial de prepação de dados

1. Definição do problema

“Dado um conjunto de atributos de uma propriedade (localização, metragem, ano de construção), como predizer o seu valor?"

propertyid	house_numbe	house_nu	predir	street	streett	postdir	apt	city	state	zip	total_value	assessed_value	year_acquired	land_square_foot	living_square_fee	bedrooms	full_baths
828195	144			MCKIERNAN	DR			WALNUT CREEK	CA	94597	62614	62614	2006	20418	2485	3	2
1144455	281			CENTER	ST			BALTIMORE	MD	21136	105500	10550	2007	4807	1368	0	0
1494347	483			NEWTON	RD			FLAGSTAFF	AZ	86011	2220	2220	0	5654	1011	3	1
1910847	802			HATCHERY	CT			WOODLAND	WA	98674	356000	356000	0	6094	0	2	1
4267562	5007		E	ROY ROGERS	RD			TROY	MI	48085	327253	327253	2007	3484	0	3	0
4888602	7607			PEBBLESTONE	DR		000009	KERNVILLE	CA	93238	732179	732179	2010	19597	6132	6	6
48725	4			LONG	AVE			SUNRISE	FL	33323	271000	271000	2008	6880	2392	4	2
83528	6			TRILLUM	LN			WAYLAND	MA	02193	79889	79889	2007	7657	1657	4	1
94604	7			PARMENTER	AVE			PLYMOUTH	MN	55441	23800	23800	2005	19994	1754	3	2
220326	17			TIMBER	RD			LOS ANGELES	CA	90063	89000	89000	2008	7840	954	3	1
994609	212			FREYER	DR	NE		PHILOMONT	VA	20131	59800	59800	2009	11199	1241	3	0
1836173	724			EASTER	ST			ALLENTOWN	PA	18102	191600	191600	0	9100	2534	4	2
2910797	1903			SADDLE BROOK	DR			CLIO	CA	96106	61610	61610	2007	0	0	0	0
3083959	2158			RIVERSIDE	DR			UPPER MORELA…	PA	19006	90300	0	0	0	1235	3	2
3952189	4040			GRAND VIEW	BLVD		000054	RIO LINDA	CA	95673	0	0	0	2700720	0	0	0
4186238	4726			LAS PALMAS	CT			WAELDER	TX	78959	18816	18816	2009	2159	1320	0	0
4597143	6213			WILSON	RD			ZOLFO SPRINGS	FL	33890	72600	0	0	8496	0	3	1
4624905	6321			STONEWALL	LN			PATERSON	NJ	07514	139880	139880	2008	10454	1391	4	2
92326	7			KNOLLCREST	DR			NARANJA	FL	33032	76214	76214	2008	4800	930	2	0
1792852	704			ERIN	DR			TRABUCO	CA	92678	28010	28010	2007	5200	0	3	1
1843977	728		S	ARLINGTON HE…	RD			BLOOMING GRO…	TX	76626	130400	130400	2007	36154	1629	3	1
4214872	4821			MYRTLE OAK	DR		000025	SAN BERNARDI	CA	92376	33250	0	2007	93654	0	0	0

2. Extração dos dados

Estrutura RECORD e declaração DATASET

```
EXPORT File_Property := MODULE
     EXPORT Layout := RECORD
          INTEGER8      propertyid;
          STRING5       streettype;
          STRING40      city;
          STRING2       state;
          STRING5       zip;
          UNSIGNED4     total_value;
          UNSIGNED4     assessed_value;
          UNSIGNED2     year_acquired;
          UNSIGNED4     land_square_footage;
          UNSIGNED4     living_square_feet;
          UNSIGNED2     bedrooms;
          UNSIGNED2     full_baths;
          UNSIGNED2     half_baths;
          UNSIGNED2     year_built;
     END;
     EXPORT File := DATASET('~online::hmw::AdvECL::property',Layout,THOR);
END;
```

```
EXPORT MLProp := RECORD
          UNSIGNED8 PropertyID;
          UNSIGNED3 zip;                          //categorica
          UNSIGNED4 assessed_value;
          UNSIGNED2 year_acquired;
          UNSIGNED4 land_square_footage;
          UNSIGNED4 living_square_feet;
          UNSIGNED2 bedrooms;
          UNSIGNED2 full_baths;
          UNSIGNED2 half_baths;
          UNSIGNED2 year_built;
          UNSIGNED4 total_value;       //Variavel dependente
     END;
```

3. Preparação dos dados

Função PROJECT(), RANDOM() e SORT()

```
// Clean the data and assign a random number to each record

CleanFilter := Property.zip <> '' AND Property.assessed_value <> 0 AND Property.year_acquired <> 0
               AND Property.land_square_footage <> 0 AND Property.living_square_feet <> 0
               AND  Property.bedrooms <> 0 AND Property.year_Built <> 0;

MLPropExt := RECORD(ML_Prop)
     UNSIGNED4 rnd; // A random number
END;

EXPORT myDataE := PROJECT(Property(CleanFilter), TRANSFORM(MLPropExt,
                                                           SELF.rnd := RANDOM(),
                                                           SELF.Zip := (UNSIGNED3)LEFT.Zip,
                                                           SELF := LEFT)) ;

// Shuffle your data by sorting on the random field
SHARED myDataES := SORT(myDataE, rnd);

// Treat first 5000 as training data.  Transform back to the original format.
EXPORT myTrainData := PROJECT(myDataES[1..5000], ML_Prop);

// Treat next 2000 as test data
EXPORT myTestData  := PROJECT(myDataES[5001..7000], ML_Prop);
```

4. Conversão e segregação de variáveis

Função ML_Core.ToField()

```
IMPORT $;
IMPORT ML_Core;

myTrainData := $.Prep01.myTrainData;
myTestData  := $.Prep01.myTestData;

//Numeric Field Matrix conversion
ML_Core.ToField(myTrainData, myTrainDataNF);
ML_Core.ToField(myTestData, myTestDataNF);

EXPORT Convert02 := MODULE
  EXPORT myIndTrainDataNF := myTrainDataNF(number < 10); //

  EXPORT myDepTrainDataNF := PROJECT(myTrainDataNF(number = 10),
                                     TRANSFORM(RECORDOF(LEFT),
                                               SELF.number := 1,
                                               SELF := LEFT));

  EXPORT myIndTestDataNF := myTestDataNF(number < 10);

  EXPORT myDepTestDataNF := PROJECT(myTestDataNF(number = 10),
                                    TRANSFORM(RECORDOF(LEFT),
                                              SELF.number := 1,
                                              SELF := LEFT));
END;
```

wi	id	number	value
1	160350	1	20706
1	160350	2	18020
1	160350	3	2007
1	160350	4	4610
1	160350	5	2594
1	160350	6	2
1	160350	7	2
1	160350	8	0
1	160350	9	1916
1	82569	1	60527
1	82569	2	78477
1	82569	3	2007
1	82569	4	6098
1	82569	5	1032
1	82569	6	3
1	82569	7	2
1	82569	8	0
1	82569	9	1992

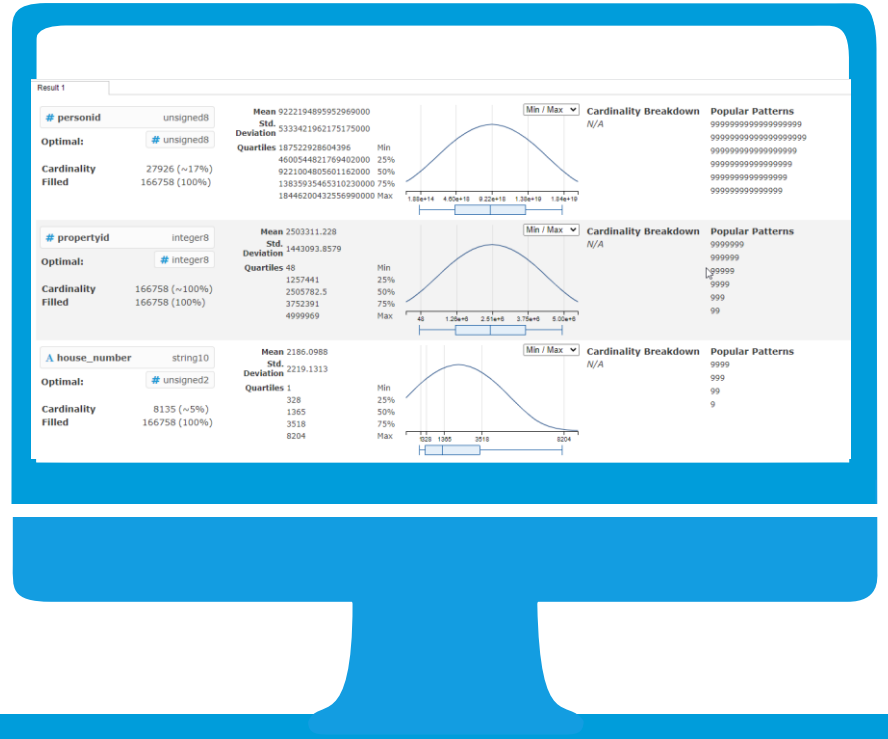
wi	id	number	value
1	160350	1	185000
1	82569	1	78477
1	2192898	1	79290
1	2223942	1	45511
1	4648854	1	39900
1	2367580	1	108610
1	607178	1	31072
1	1584497	1	88284
1	3615520	1	341400
1	2103806	1	58520
1	2209348	1	87610
1	1298734	1	66175
1	1310023	1	301644
1	2840506	1	94200
1	3600022	1	262700
1	131449	1	16500
1	4649661	1	84000
1	1042629	1	38740
1	1732698	1	197700

Desafio: Lending Club

Exercício prático:

Faça o perfilamento do dataset do Lending Club

- Utilize a bilbioteca DataPatterns



Fundação Vanzolini

Até a próxima aula!!!