Fundação Vanzolini

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

Aula 3

LexisNexis® RISK SOLUTIONS

RELX Group

PRO

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO

USP UNIVERSIDADE DE SÃO PAULO

Bem-vindo! – Agenda da aula 3

- ✓Desafio Lending Club
- ✓Regressão Linear
- ✓Regressão Logística binomial

Exercício prático:

Faça o perfilamento do dataset do Lending Club

- Utilize a bilbioteca DataPatterns

Fundação Vanzolini

Regressão linear

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).

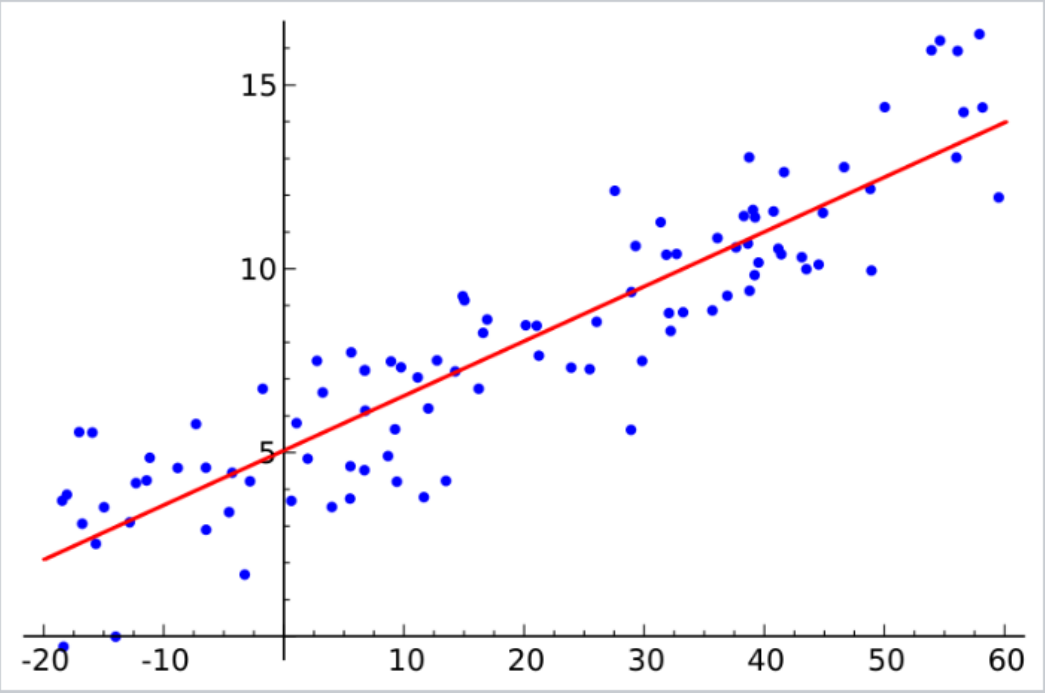
Modelos quantitativos e qualitativos

- Aprendizado supervisionado suporta dois tipos principais de modelos:
  - Quantitativo – Ex.: determinar a idade da árvore
  - Qualitativo – Ex.: determinar a espécie da árvore
- O processo de obtenção de um modelo quantitativo é também conhecido como “Regressão”.
- O processo de obtenção de um modelo qualitativo é denominado “Classificação”.

Regressão linear

Algoritmo de regressão que assume que a variável dependente é uma função linear das variáveis independentes (https://github.com/hpcc-systems/LinearRegression.git)

$Y = mX + b$



Ref: https://en.wikipedia.org/wiki/Linear_regression

Tutorial de regressão linear

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

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

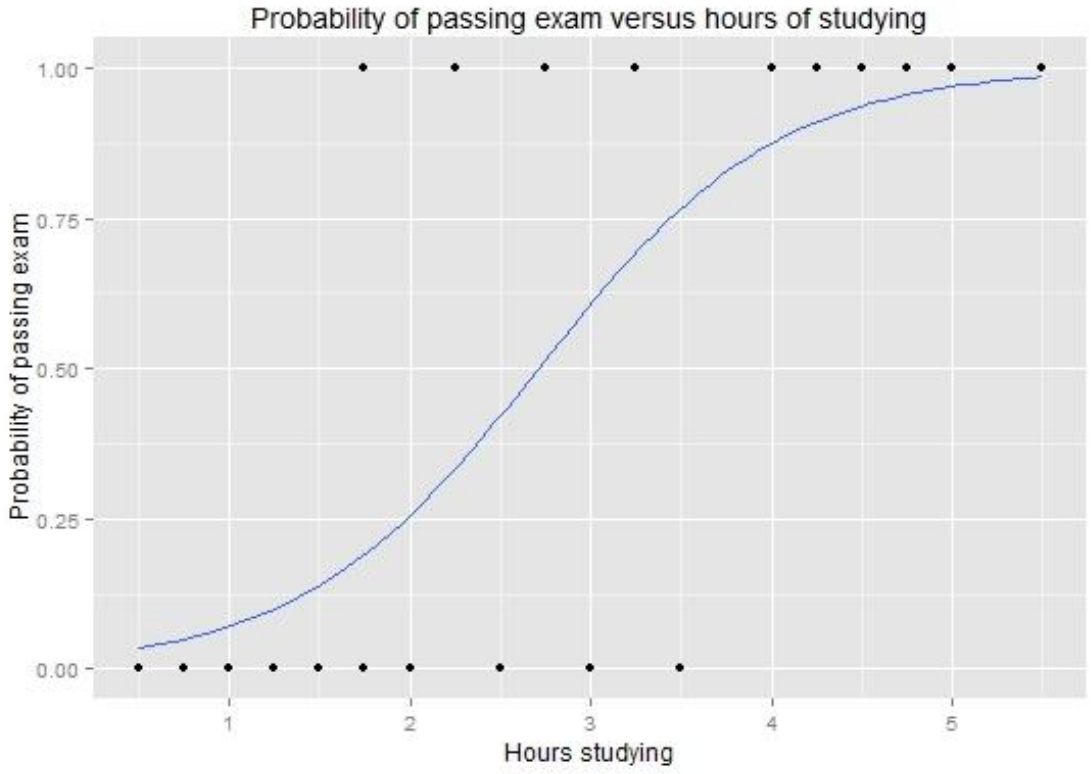
Fundação Vanzolini

Regressão logística binomial

Regressão logística binomial

Algoritmo de classificação que estabelece uma relação de dependência entre variáveis preditoras e uma variavel dependente binária.

(https://github.com/hpcc-systems/LogisticRegression.git)



Ref: By Michaelg2015 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=42442194

Tutorial de regressão logística

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

1. Definição do problema

“Dado um conjunto de atributos de um histórico de contatos telefonicos feitos por um banco a seus clientes, como predizer se o cliente fará um investimento?"

age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp_var_rate	cons_price_idx	cons_conf_idx	euribor3m	nr_employed	y
44	blue-collar	married	basic.4y	unknown	yes	no	cellular	aug	thu	210	1	999	0	nonexistent	1.4	93.444	-36.1	4.963	5228.1	0
53	technician	married	unknown	no	no	no	cellular	nov	fri	138	1	999	0	nonexistent	-0.1	93.2	-42	4.021	5195.8	0
28	management	single	university.degree	no	yes	no	cellular	jun	thu	339	3	6	2	success	-1.7	94.055	-39.8	0.729	4991.6	1
39	services	married	high.school	no	no	no	cellular	apr	fri	185	2	999	0	nonexistent	-1.8	93.075	-47.1	1.405	5099.1	0
55	retired	married	basic.4y	no	yes	no	cellular	aug	fri	137	1	3	1	success	-2.9	92.201	-31.4	0.869	5076.2	1
30	management	divorced	basic.4y	no	yes	no	cellular	jul	tue	68	8	999	0	nonexistent	1.4	93.918	-42.7	4.961	5228.1	0
37	blue-collar	married	basic.4y	no	yes	no	cellular	may	thu	204	1	999	0	nonexistent	-1.8	92.893	-46.2	1.327	5099.1	0
39	blue-collar	divorced	basic.9y	no	yes	no	cellular	may	fri	191	1	999	0	nonexistent	-1.8	92.893	-46.2	1.313	5099.1	0
36	admin.	married	university.degree	no	no	no	cellular	jun	mon	174	1	3	1	success	-2.9	92.963	-40.8	1.266	5076.2	1
27	blue-collar	single	basic.4y	no	yes	no	cellular	apr	thu	191	2	999	1	failure	-1.8	93.075	-47.1	1.41	5099.1	0
34	housemaid	single	university.degree	no	no	no	telephone	may	fri	62	2	999	0	nonexistent	1.1	93.994	-36.4	4.864	5191	0
41	management	married	university.degree	no	yes	no	cellular	aug	thu	789	1	999	0	nonexistent	1.4	93.444	-36.1	4.964	5228.1	0
55	management	married	university.degree	no	no	no	cellular	aug	mon	372	3	999	0	nonexistent	1.4	93.444	-36.1	4.965	5228.1	1
33	services	divorced	high.school	no	yes	no	cellular	may	tue	75	5	999	0	nonexistent	-1.8	92.893	-46.2	1.291	5099.1	0
26	admin.	married	high.school	no	no	yes	telephone	jun	mon	1021	1	999	0	nonexistent	1.4	94.465	-41.8	4.96	5228.1	0
52	services	married	high.school	unknown	yes	no	cellular	jul	thu	117	2	999	0	nonexistent	1.4	93.918	-42.7	4.962	5228.1	0

Desafio: Lending Club

Exercício prático:

Prepare o dataset do Lending Club

- Considere a aplicação de aprendizagem supervisionada
- Se baseie nos resultados do perfilamento de dados

Fundação Vanzolini

Até a próxima aula!!!