Big Data Management

10/06/2021

Micha Kaiser

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

EFMD EQUIS ACCREDITED
AACSB ACCREDITED
ASSOCIATION OF MBAS ACCREDITED
CEMS
PIM

Learning objectives

- After this lecture you should…
  - …understand how a decision tree for classification works
  - …understand how a decision tree for regression works
  - …understand how to acquire the data for your final project

Agenda

- Decision tree for classification
- Decision tree for regression
- Various useful data sources

Decision tree for regression

AtBat	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	CHmRun	CRuns	CRBI	CWalks	League	Division	PutOuts	Assists	Errors	Salary	NewLeague
293	66	1	30	29	14	1	293	66	1	30	29	14	A	E	446	33	20	NA	A
315	81	7	24	38	39	14	3449	835	69	321	414	375	N	W	632	43	10	475	N
479	130	18	66	72	76	3	1624	457	63	224	266	263	A	W	880	82	14	480	A
496	141	20	65	78	37	11	5628	1575	225	828	838	354	N	E	200	11	3	500	N
321	87	10	39	42	30	2	396	101	12	48	46	33	N	E	805	40	4	91.5	N
594	169	4	74	51	35	11	4408	1133	19	501	336	194	A	W	282	421	25	750	A
185	37	1	23	8	21	2	214	42	1	30	9	24	N	E	76	127	7	70	A
298	73	0	24	24	7	3	509	108	0	41	37	12	A	W	121	283	9	100	A
323	81	6	26	32	8	2	341	86	6	32	34	8	N	W	143	290	19	75	N
401	92	17	49	66	65	13	5206	1332	253	784	890	866	A	E	0	0	0	1100	A
574	159	21	107	75	59	10	4631	1300	90	702	504	488	A	E	238	445	22	517.143	A
202	53	4	31	26	27	9	1876	467	15	192	186	161	N	W	304	45	11	512.5	N
418	113	13	48	61	47	4	1512	392	41	205	204	203	N	E	211	11	7	550	N
239	60	0	30	11	22	6	1941	510	4	309	103	207	A	E	121	151	6	700	A
196	43	7	29	27	30	13	3231	825	36	376	290	238	N	E	80	45	8	240	N
183	39	3	20	15	11	3	201	42	3	20	16	11	A	W	118	0	0	NA	A
568	158	20	89	75	73	15	8068	2273	177	1045	993	732	N	W	105	290	10	775	N
190	46	2	24	8	15	5	479	102	5	65	23	39	A	W	102	177	16	175	A
407	104	6	57	43	65	12	5233	1478	100	643	658	653	A	W	912	88	9	NA	A
127	32	8	16	22	14	8	727	180	24	67	82	56	N	W	202	22	2	135	N
413	92	16	72	48	65	1	413	92	16	72	48	65	N	E	280	9	5	100	N
426	109	3	55	43	62	1	426	109	3	55	43	62	A	W	361	22	2	115	N
22	10	1	4	2	1	6	84	26	2	9	9	3	A	W	812	84	11	NA	A
472	116	16	60	62	74	6	1924	489	67	242	251	240	N	W	518	55	3	600	N
629	168	18	73	102	40	18	8424	2464	164	1008	1072	402	A	E	1067	157	14	776.667	A

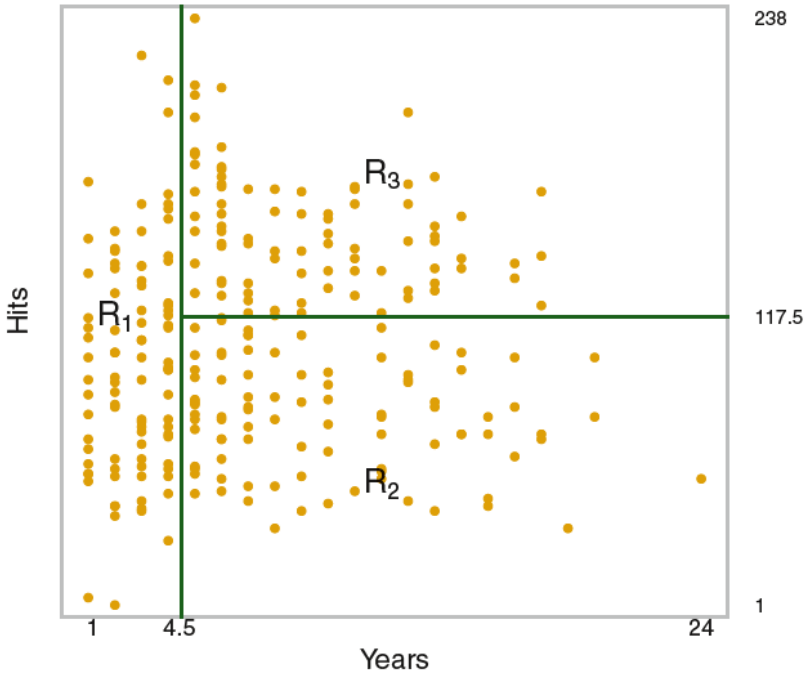
Source: https://gist.github.com/keeganhines/59974f1ebef97bbaa44fb19143f90bad#file-hitters-csv

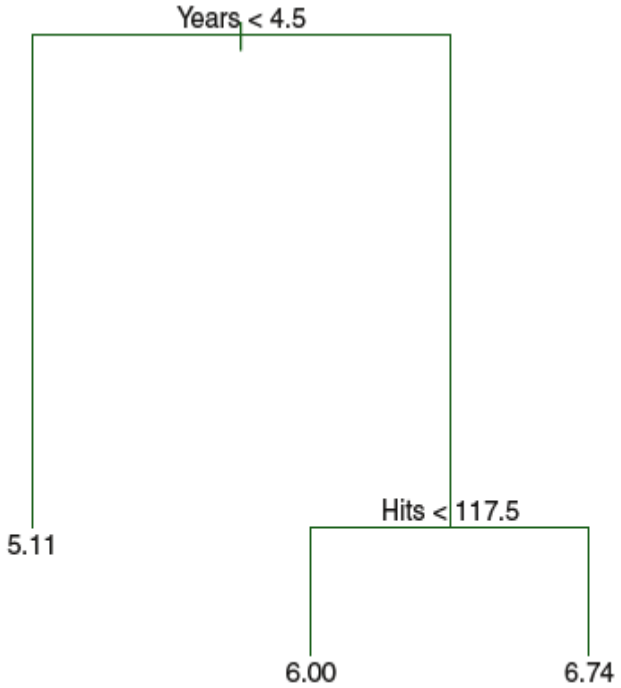
Decision tree for regression

Numeric!

AtBat	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	CHmRun	CRuns	CRBI	CWalks	League	Division	PutOuts	Assists	Errors	Salary	NewLeague
293	66	1	30	29	14	1	293	66	1	30	29	14	A	E	446	33	20	NA	A
315	81	7	24	38	39	14	3449	835	69	321	414	375	N	W	632	43	10	475	N
479	130	18	66	72	76	3	1624	457	63	224	266	263	A	W	880	82	14	480	A
496	141	20	65	78	37	11	5628	1575	225	828	838	354	N	E	200	11	3	500	N
321	87	10	39	42	30	2	396	101	12	48	46	33	N	E	805	40	4	91.5	N
594	169	4	74	51	35	11	4408	1133	19	501	336	194	A	W	282	421	25	750	A
185	37	1	23	8	21	2	214	42	1	30	9	24	N	E	76	127	7	70	A
298	73	0	24	24	7	3	509	108	0	41	37	12	A	W	121	283	9	100	A
323	81	6	26	32	8	2	341	86	6	32	34	8	N	W	143	290	19	75	N
401	92	17	49	66	65	13	5206	1332	253	784	890	866	A	E	0	0	0	1100	A
574	159	21	107	75	59	10	4631	1300	90	702	504	488	A	E	238	445	22	517.143	A
202	53	4	31	26	27	9	1876	467	15	192	186	161	N	W	304	45	11	512.5	N
418	113	13	48	61	47	4	1512	392	41	205	204	203	N	E	211	11	7	550	N
239	60	0	30	11	22	6	1941	510	4	309	103	207	A	E	121	151	6	700	A
196	43	7	29	27	30	13	3231	825	36	376	290	238	N	E	80	45	8	240	N
183	39	3	20	15	11	3	201	42	3	20	16	11	A	W	118	0	0	NA	A
568	158	20	89	75	73	15	8068	2273	177	1045	993	732	N	W	105	290	10	775	N
190	46	2	24	8	15	5	479	102	5	65	23	39	A	W	102	177	16	175	A
407	104	6	57	43	65	12	5233	1478	100	643	658	653	A	W	912	88	9	NA	A
127	32	8	16	22	14	8	727	180	24	67	82	56	N	W	202	22	2	135	N
413	92	16	72	48	65	1	413	92	16	72	48	65	N	E	280	9	5	100	N
426	109	3	55	43	62	1	426	109	3	55	43	62	A	W	361	22	2	115	N
22	10	1	4	2	1	6	84	26	2	9	9	3	A	W	812	84	11	NA	A
472	116	16	60	62	74	6	1924	489	67	242	251	240	N	W	518	55	3	600	N
629	168	18	73	102	40	18	8424	2464	164	1008	1072	402	A	E	1067	157	14	776.667	A

Decision tree for regression





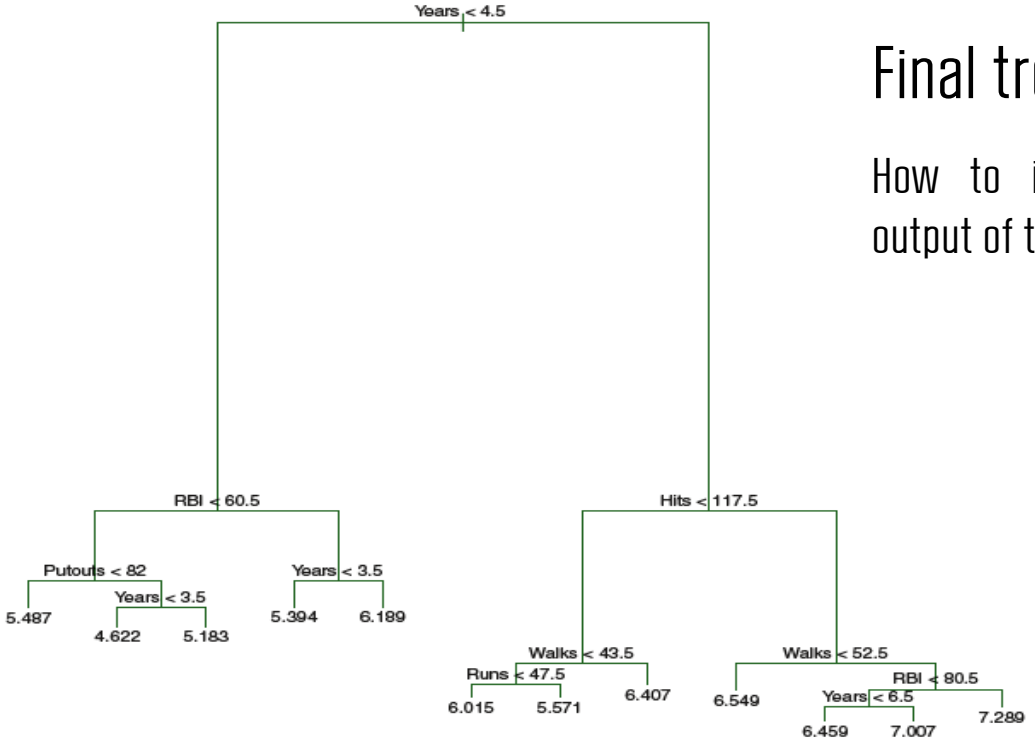
Decision tree for regression

Roughly speaking, there are two steps of building a regression tree:

1. We divide the predictor into J distinct and non-overlapping regions, R_1, R_2, R_3,...,R_J by *recursive binary splitting*
2. For every observation that falls into Region R_J we make the same prediction, which is simply the mean of the response values for the training observations in R_J

The main goal is to minimize the RSS: $$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

Decision tree for regression



Final tree

How to interpret the final output of the regression tree?

Decision tree for regression

- Advantages:
  - It is easy to interpret and easy to explain to people
  - Some people believe that decision trees more closely mirror human decision-making
  - Has a nice graphical representation
  - Perform well if the underlying data generating process is complex and highly non-linear
- Disadvantages
  - Overfitting (if not pruned) and hence poor performance on the test data
  - Oversimplification

Decision tree for classification

- Classification trees are very similar to regression trees
- Instead of predicting a quantitative response we predict a qualitative response
- Instead of minimizing the RSS we want to minimize
  - The classification error rate
  - The Gini index
  - The entropy
- Here, $\hat{p}_{mk}$ presents the proportion of training observations in the *m*th region that are from the *k*th class

$$E = 1 - \max_k(\hat{p}_{mk})$$

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

Decision tree for classification

Small example

ID	A	B	C	Classification
1	T	T	T	F
2	T	T	T	T
3	T	F	F	F
4	F	T	T	T
5	F	F	T	T

Decision tree for classification

1. Entropy before first split:

$$H(T) = -\frac{3}{5} * log_2\left(\frac{3}{5}\right) - \frac{2}{5} * log_2\left(\frac{2}{5}\right) = 0.971$$

2. The conditional entropy of split 1 at A:

$$H(T|A) = \frac{3}{5} * \left(-\frac{1}{3} * log_2\left(\frac{1}{3}\right) - \frac{2}{3} * log_2\left(\frac{2}{3}\right)\right) + \frac{2}{5} * \left(-1 * log_2(1)\right) = 0.551$$

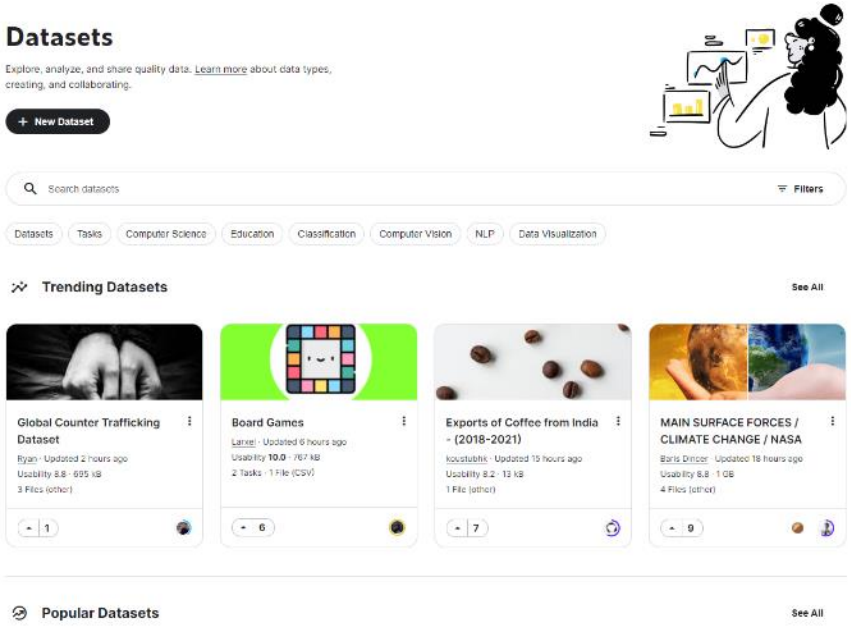
3. Calculate information gain:

$$\mathrm{IG}(T, A) = H(T) - H(T|A) = 0.42$$

4. Repeat these steps for all candidate splits and pick the one with the highest information gain

Various useful data sources

- www.kaggle.com
- Online community for data scientists
- All kind of data sets, covering various topics



Various useful data sources

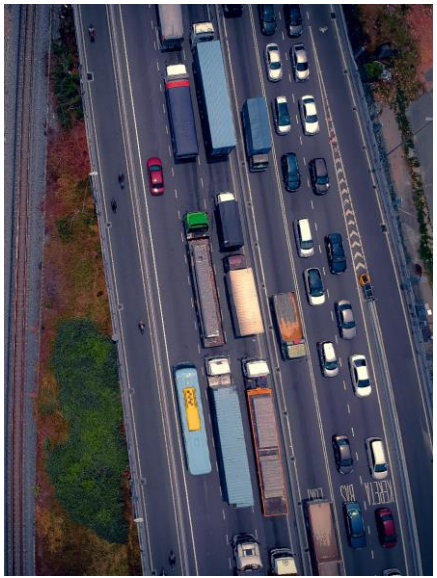
Survey data - examples

- German Socio Economic Panel (GSOEP)
- British Household Panel Survey (BHPS)
- Panel Study of Income Dynamics (PSID)
- Household, Income and Labor Dynamics (HILDA)
- Korean Labor and Income Panel Study (KLIPS)
- Russian Longitudinal Monitoring Survey (RLMS)
- Swiss Household Panel (SHP)
- …and many more…

Various useful data sources

Public data/Census data/official statistics

- https://www.dst.dk/en (Denmark)
- https://www.census.gov/ (USA)
- https://www.usa.gov/statistics (USA)
- https://www.destatis.de/EN/Home/_node.html (Germany)
- https://www.ons.gov.uk/census (UK)
- https://ec.europa.eu/eurostat/data/database (EU)



Literature

- James, Gareth, et al. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013. **Chapter 8.1.**