

COMPSCI361: Machine Learning

Data Preprocessing

Katerina Taskova and Jörg Simon Wicker
The University of Auckland

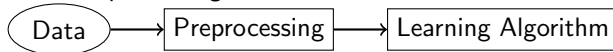


SCIENCE
SCHOOL OF COMPUTER SCIENCE

Week 5-8

- In weeks 5-8, we will cover:

- Data Preprocessing



Week 5-8



SCIENCE
SCHOOL OF COMPUTER SCIENCE

- In weeks 5-8, we will cover:
 - Bayes Learning

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Week 5-8

- In weeks 5-8, we will cover:
 - Clustering



Week 5-8

- In weeks 5-8, we will cover:

- Association Rules

If X buys *bread*, then X buys *milk* [support 50 %, confidence = 100 %]

Bread	Eggs	Milk	Oranges
1	1	1	0
0	0	1	0
1	0	1	0
0	1	0	1

Data Preprocessing

This week we will cover



Data Preprocessing

- Data Cleaning

- Missing Data

- Preprocessing and Evaluation

- Data Reduction

- Noisy Data

- Data Transformation and Data Discretization

- Imbalanced Data

Why preprocess?

- Preprocessing means to transform the data before we feed it to a learning algorithm
- Why would we do that?
- What would we for example do?

```
time, CO2, m14.0626, m15.0238, m15.9942, m16.0201, m17.0260, m18.0338, m19.0418, m20.0500, m21.0582, m22.0664, m23.0746, m24.0828, m25.0910, m26.0992, m27.1074, m28.1156, m29.1238, m30.1320, m31.1402, m32.1484, m33.1566, m34.1648, m35.1730, m36.1812, m37.1894, m38.1976, m39.2058, m40.2140, m41.2222, m42.2304, m43.2386, m44.2468, m45.2550, m46.2632, m47.2714, m48.2796, m49.2878, m50.2960, m51.3042, m52.3124, m53.3206, m54.3288, m55.3370, m56.3452, m57.3534, m58.3616, m59.3698, m60.3780, m61.3862, m62.3944, m63.4026, m64.4108, m65.4190, m66.4272, m67.4354, m68.4436, m69.4518, m70.4600, m71.4682, m72.4764, m73.4846, m74.4928, m75.5010, m76.5092, m77.5174, m78.5256, m79.5338, m80.5420, m81.5502, m82.5584, m83.5666, m84.5748, m85.5830, m86.5912, m87.5994, m88.6076, m89.6158, m90.6240, m91.6322, m92.6404, m93.6486, m94.6568, m95.6650, m96.6732, m97.6814, m98.6896, m99.6978, m100.7060, m101.7142, m102.7224, m103.7306, m104.7388, m105.7470, m106.7552, m107.7634, m108.7716, m109.7798, m110.7880, m111.7962, m112.8044, m113.8126, m114.8208, m115.8290, m116.8372, m117.8454, m118.8536, m119.8618, m120.8700, m121.8782, m122.8864, m123.8946, m124.9028, m125.9110, m126.9192, m127.9274, m128.9356, m129.9438, m130.9520, m131.9602, m132.9684, m133.9766, m134.9848, m135.9930, m136.0012, m137.0094, m138.0176, m139.0258, m140.0340, m141.0422, m142.0504, m143.0586, m144.0668, m145.0750, m146.0832, m147.0914, m148.0996, m149.1078, m150.1160, m151.1242, m152.1324, m153.1406, m154.1488, m155.1570, m156.1652, m157.1734, m158.1816, m159.1898, m160.1980, m161.2062, m162.2144, m163.2226, m164.2308, m165.2390, m166.2472, m167.2554, m168.2636, m169.2718, m170.2800, m171.2882, m172.2964, m173.3046, m174.3128, m175.3210, m176.3292, m177.3374, m178.3456, m179.3538, m180.3620, m181.3702, m182.3784, m183.3866, m184.3948, m185.4030, m186.4112, m187.4194, m188.4276, m189.4358, m190.4440, m191.4522, m192.4604, m193.4686, m194.4768, m195.4850, m196.4932, m197.5014, m198.5096, m199.5178, m200.5260, m201.5342, m202.5424, m203.5506, m204.5588, m205.5670, m206.5752, m207.5834, m208.5916, m209.6000, m210.6082, m211.6164, m212.6246, m213.6328, m214.6410, m215.6492, m216.6574, m217.6656, m218.6738, m219.6820, m220.6902, m221.6984, m222.7066, m223.7148, m224.7230, m225.7312, m226.7394, m227.7476, m228.7558, m229.7640, m230.7722, m231.7804, m232.7886, m233.7968, m234.8050, m235.8132, m236.8214, m237.8296, m238.8378, m239.8460, m240.8542, m241.8624, m242.8706, m243.8788, m244.8870, m245.8952, m246.9034, m247.9116, m248.9198, m249.9280, m250.9362, m251.9444, m252.9526, m253.9608, m254.9690, m255.9772, m256.9854, m257.9936, m258.0018, m259.0100, m260.0182, m261.0264, m262.0346, m263.0428, m264.0510, m265.0592, m266.0674, m267.0756, m268.0838, m269.0920, m270.0999, m271.1081, m272.1163, m273.1245, m274.1327, m275.1409, m276.1491, m277.1573, m278.1655, m279.1737, m280.1819, m281.1901, m282.1983, m283.2065, m284.2147, m285.2229, m286.2311, m287.2393, m288.2475, m289.2557, m290.2639, m291.2721, m292.2803, m293.2885, m294.2967, m295.3049, m296.3131, m297.3213, m298.3295, m299.3377, m300.3459, m301.3541, m302.3623, m303.3705, m304.3787, m305.3869, m306.3951, m307.4033, m308.4115, m309.4197, m310.4279, m311.4361, m312.4443, m313.4525, m314.4607, m315.4689, m316.4771, m317.4853, m318.4935, m319.5017, m320.5099, m321.5181, m322.5263, m323.5345, m324.5427, m325.5509, m326.5591, m327.5673, m328.5755, m329.5837, m330.5919, m331.6001, m332.6083, m333.6165, m334.6247, m335.6329, m336.6411, m337.6493, m338.6575, m339.6657, m340.6739, m341.6821, m342.6903, m343.6985, m344.7067, m345.7149, m346.7231, m347.7313, m348.7395, m349.7477, m350.7559, m351.7641, m352.7723, m353.7805, m354.7887, m355.7969, m356.8051, m357.8133, m358.8215, m359.8297, m360.8379, m361.8461, m362.8543, m363.8625, m364.8707, m365.8789, m366.8871, m367.8953, m368.9035, m369.9117, m370.9199, m371.9281, m372.9363, m373.9445, m374.9527, m375.9609, m376.9691, m377.9773, m378.9855, m379.9937, m380.0019, m381.0101, m382.0183, m383.0265, m384.0347, m385.0429, m386.0511, m387.0593, m388.0675, m389.0757, m390.0839, m391.0921, m392.1003, m393.1085, m394.1167, m395.1249, m396.1331, m397.1413, m398.1495, m399.1577, m400.1659, m401.1741, m402.1823, m403.1905, m404.1987, m405.2069, m406.2151, m407.2233, m408.2315, m409.2397, m410.2479, m411.2561, m412.2643, m413.2725, m414.2807, m415.2889, m416.2971, m417.3053, m418.3135, m419.3217, m420.3299, m421.3381, m422.3463, m423.3545, m424.3627, m425.3709, m426.3791, m427.3873, m428.3955, m429.4037, m430.4
```


This week we will...



SCIENCE
SCHOOL OF COMPUTER SCIENCE

- Talk about problems that can appear in data
- Introduce strategies to solve these problems
- Talk about feature selection, a very important technique in machine learning

Major Tasks in Data Preprocessing

- Data cleaning
 - Missing values
 - Noisy data
 - Outliers
- Data reduction
 - Dimensionality reduction
 - Numerosity reduction
 - Data compression
- Transformation and discretization
 - Normalization
 - Hierarchy generation

Data Cleaning

- Basic assumption in machine learning?
- But, real-world data are, in most cases, dirty
- This can lead to problems, if data are
 - Incomplete** lacking attribute values, certain attributes, or containing only aggregate data
 - Noisy** containing noise, errors, or outliers
 - Inconsistent** containing discrepancies in codes or names
 - Intentially wrong** for example, there are a lot of pictures with a GPS location just a bit west of Africa

Incomplete (Missing) Data

- Data are not always available
 - Many tuples have no recorded value for several attributes
 - E.g. customer income in sales data
- Missing data may be due to
 - Equipment malfunction
 - Inconsistent with other recorded data and thus deleted
 - Data not entered due to misunderstanding
 - Certain data may not be considered important at the time of entry
 - Data history or changes of the data not recorded
- Missing data may need to be inferred
 - When, for example?

What to Consider When Handling Missing Data?

- Missing completely at random (MCAR)
 - Completely unrelated to the data

Name	Country	Income
Jane	NZ	\$50k
Kate	NZ	\$75k
Tom	US	\$53k
George	UK	\$64k
Mark	UK	\$77k
Philippe	US	\$80k

MCAR

Name	Country	Income
Jane	NZ	
	NZ	\$75k
Tom	US	
George		\$64k
	UK	\$77k
Philippe	US	\$80k

-
- Potential problem? Small sample size

What to Consider When Handling Missing Data?

■ Missing at random (MAR)

- The fact the data are missing is related not to the missing attribute, but to some other data in the data set

Name	Country	Income
Jane	NZ	\$50k
Kate	NZ	\$75k
Tom	US	\$53k
George	UK	\$64k
Mark	UK	\$77k
Philippe	US	\$80k

MAR

Name	Country	Income
Jane	NZ	\$50k
Kate	NZ	\$75k
Tom	US	\$53k
George	UK	
Mark	UK	
Philippe	US	\$80k

-
- Potential problem? Bias due to row-wise deletion

What to Consider When Handling Missing Data?

■ Missing not at random (MNAR)

- There is a reason the data are missing and it is related to the attribute itself

Name	Country	Income
Jane	NZ	\$50k
Kate	NZ	\$75k
Tom	US	\$53k
George	UK	\$64k
Mark	UK	\$77k
Philippe	US	\$80k

MNAR →

Name	Country	Income
Jane	NZ	
Kate	NZ	\$75k
Tom	US	
George	UK	
Mark	UK	\$77k
Philippe	US	\$80k

-
- Potential problem? Bias due to row-wise deletion

How to Handle Missing Data – Imputation

■ Ignore the tuple

X						X'				
0	1	1	1	...		0	1	1	1	...
?	?	?	1
1	0	?	?	...	→	1	0	1	0	...
...						
1	0	1	0	...						

- Usually done when the class label is missing (classification)
- Not effective when the fraction of missing values varies considerably

How to Handle Missing Data – Imputation

- Fill in the missing data manually

X						X'				
0	1	1	1	...	→	0	1	1	1	...
?	?	?	1	...		1	0	0	1	...
1	0	?	?	...		1	0	1	1	...
...
1	0	1	0	...		1	0	1	0	...

- Tedious and sometimes infeasible

How to Handle Missing Data – Imputation

- Fill in automatically
 - A global constant

X						X'				
sunny	warm	Mon	May	...		sunny	warm	Mon	May	...
cloudy	?	?	July	...		cloudy	missing	missing	July	...
sunny	cold	?	?	...		sunny	cold	missing	missing	...
...
overcast	cold	Sat	June	...		overcast	cold	Sat	June	...

- E.g. “missing”
- A new class

How to Handle Missing Data – Imputation

- Fill in automatically
 - The attribute mean

X						X'				
12	2	22	38	...	→	12	2	22	37	...
11	?	?	90	...		11	12	38	90	...
2	23	?	?	...		2	23	38	30	...
...
9	11	54	23	...		9	11	54	23	...

- Done automatically by many implementations
- Changes relationship with other variables \Rightarrow bias in data

How to Handle Missing Data – Imputation

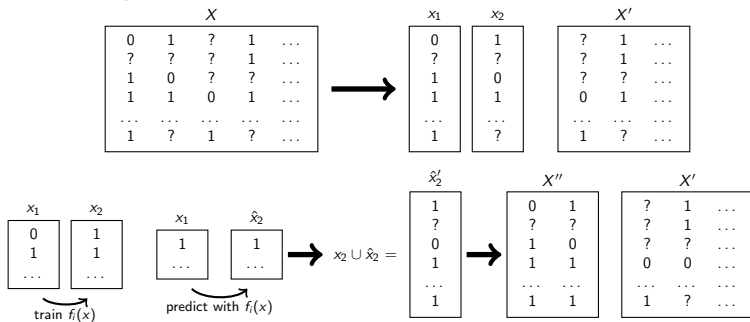
- Fill in automatically
 - The attribute mean of the samples belonging to the same class

$X Y$							$X' Y$					
12	2	22	38	...	1	→	12	2	22	38	...	1
11	?	?	90	...	0		11	11	54	90	...	0
2	23	?	?	...	1		2	23	22	38	...	1
...
9	11	54	23	...	0		9	11	54	23	...	0

- Might change relationship with other variables other than class \Rightarrow bias in data

How to Handle Missing Data – Imputation

- Fill in automatically
 - The most probable value



- Inference-based such as Bayesian formula, decision tree, nearest neighbour,...

More on Imputation

- Matrix decomposition approaches
 - Decompose matrix using, e.g, Singular Value Decomposition
 - Decompose the data matrix X such that $X = U\Lambda V^T$
 - Create imputed matrix X' by multiplying $U \times \Lambda \times V^T$

$$\begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{bmatrix} \approx \begin{bmatrix} u_{11} & \cdots & u_{1k} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nk} \end{bmatrix} \begin{bmatrix} \lambda_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_{nk} \end{bmatrix} \begin{bmatrix} v_{11} & \cdots & v_{1d} \\ \vdots & \ddots & \vdots \\ v_{k1} & \cdots & v_{kd} \end{bmatrix}$$

- Minimize the sum of squared errors

$$\min_{U, \Lambda, V} \sum_{x_{ij} \in X} (x_{ij} - [U\Lambda V]_{ij})^2$$

Even More on Imputation

- EM imputation

- Expectation Maximization
- Use other variables to impute the values (Expectation)
- Check if value is most probable (Maximization)

- Multiple imputation (e.g. MICE)

1. Impute missing values using appropriate model (for example using classifier / regression model to predict the missing value)
2. Repeat the step multiple times (3-5)
3. Carry out required full analysis of data (e.g. build classifier and evaluate)
4. Average the results (predictions or evaluation)

- So what is the best approach?

Preprocessing and Evaluation



- So now we know a preprocessing example
- Where would you put the preprocessing step in the evaluation?
- For example, for imputation:
 - Impute the values before splitting in train and test?
 - Impute the values in the training set – then how about the test set?

Conclusion



- Preprocessing is an important part in machine learning and data analysis
- Missing values can be caused by various reasons depending on what the reasons are, they must be addressed differently
- Various imputation approaches exist, they use the information of other instances and values to impute the missing values

Literature

- Material in Chapter 3 in Han's *Data Mining*

Thank you for your attention!

`https://ml.acukland.ac.nz`