Machine Learning FS2019

Alex Neher

June 15, 2019

Contents

1 Introduction

There are two popular definitions of Machine Learning:

> "Field of study that gives computers the ability to earn without being explicitly programmed" (Arthur Samuel, IBM, 1959)

> "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E" (Tom Mitchell, 1998)

So summarizing these two quotes, it can be said, that machine learning is defined as **the process in which machines learn something (mostly) on their own**.

1.1 Disciplines

There are different disciplines in machine learning:

Supervised Learning: The algorithm is given **labeled training data** and learns to **predict** the **labels** of yet unseen examples.

Unsupervised Learning: The algorithm is given **unlabeled data** and **creates labels by itself** based on the structure of the given data

Semi-Supervised Learning: A **mixture** of supervised and unsupervised learning. This approach is usually chosen if there is only **very little labeled test data**

Reinforcement Learning: No data is availabe, but the algorithm is **being rewarded**. The algorithm searches the ideal behaviour that maximizes its reward (Not subject of this lecture)

These classifications can be subdivided even more:

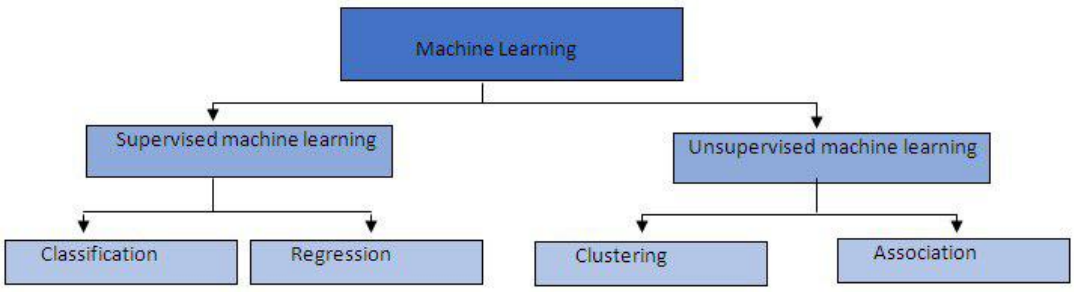


Figure 1.1: Distinction between supervised and unsupervised learning

The main difference between **classification** and **regression** is that when using classification, the result is **categorical**, whereas regression returns **numerical** results.

Clustering is similar to classificiation. However, while classification algorithms sort the given data into given groups, clustering algorithms determine these groups **by themself**. This means, you can give a clustering algorithm a seemingly random dataset and the algorithm finds some kind of structure in it.

2 Data Quality

Data is categorized into **numerical** and **categorical** data.

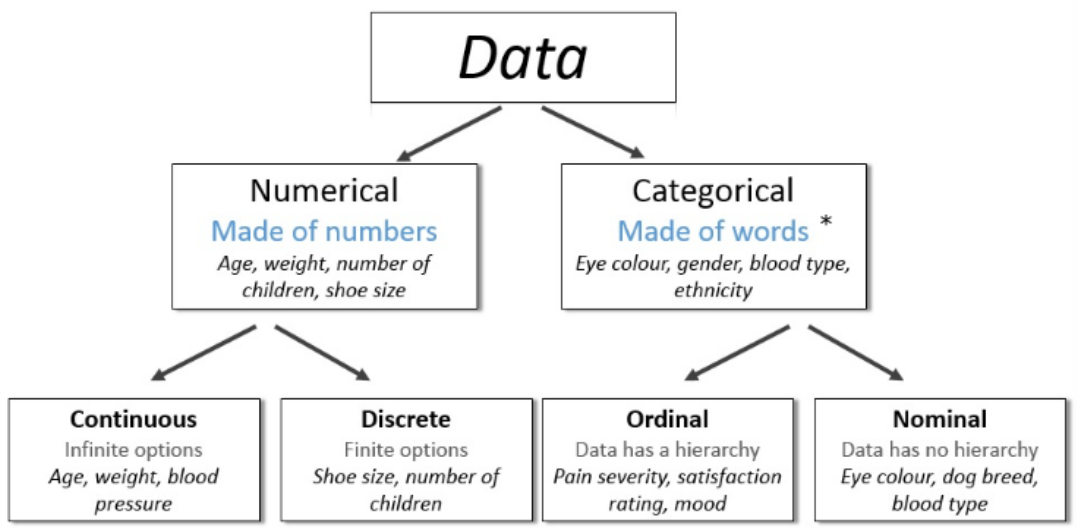


Figure 2.1: Classification of Data

Before any machine learning can take place, the quality of the given data has to be assessed and in some cases improved. Because every prediction made by machine learning algorithms is shit if the data quality is shit.

There are many reasons why the data quality could be poor:

- Ill-designed, inadequate or inconsistent data formats
- Programming errors or technical issues (e.g. sensor outage)
- Data decay (e.g. outdated e-mail addresses)
- Poorly designed data entry forms (e.g. data fields without verification)
- Human errors in data export or data pre-processing
- Deliberate errors and false information (e.g. due to privacy concerns everybody is called Hans Muster and lives at Musterstrasse 123)

2.1 Data Quality Assessment

Before even starting to assess the data-quality, it is seldomly a bad idea to **clean** the data first.

1. Identify and remove duplicates
2. Replace null-values (do not delete them because that might falsify the mean and median of the data)
3. Make data formats more machine-friendly (so-called *data-wrangling* e.g. store the gender as boolean)

If you change anything from the original data set, you should always

- Document all the changes
- Use a SVN (e.g. git)
- Let the data provider know that his data quality is shit (maybe they'll improve in the future)
- Investigate the origins of the poor data quality

2.2 Approaches to Data Quality Assessment

Identify data sources and their trustworthiness

Interpret statical key figures: See following sections

Visualize selected portions of the data: e.g. with Pair Plots (See Abb. 2.2)

Manually check data ranges Negative Salaries, People more than 200 years old...

Validate plausibility of attribute correlation: e.g. are mileage and number of seats in a core correlated? Can one of the columns be removed for redundancy?

Measure data redundancy: Can certain columns be removed due to not adding any real value to the data

Check for anomalies in syntax and semantics: Outliers can really distort a dataset and render the whole algorithm useless. Can be prevented by e.g. normalization of the data or removal of the outlier

Replace NULL Values and remove duplicate values

There are different ways to cope with NULL variables, but they have to be addressed, as most machine learning algorithms do not play well with them.

- Delete all rows with NULL values
  Might be the easiest way if you have loads of data
- Fill in the missing values manually (e.g. from other sources)
  Might be the hardest way if you have loads of data
- Fill in a global constant like N/A, UNKNOWN
- Use a measure for central tendency
  e.g. take the mean if your data is symmetric or take the median if its skewed
- Use a measure for central tendency per class
  e.g. take different values for healthy and sick people
- Use e.g Regression to 'guess' the missing values

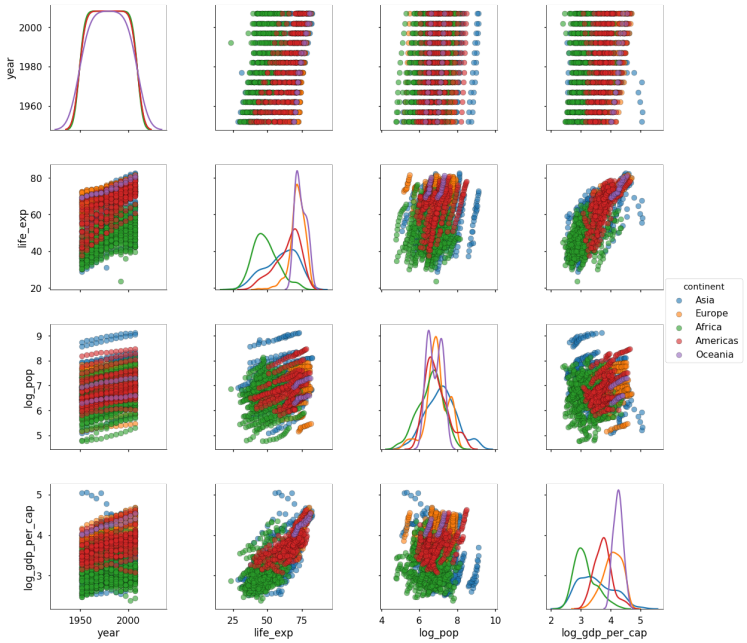


Figure 2.2: Visualisation of Data with Pair Plots

2.3 Statistical Key Figures

These figures can give you a rough overview about the whereabouts of your data-magnitude.

2.3.1 Central Tendency

Mean

This is the averge in a set of numeric data. You add all data and divide it by the number of data points

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

Mode

This is the value that occurs the most in a given set of data

Median

This is the middlemost value of a sorted set of data. In contrast to the Mean, the Median can give information concerning the distribution of the data.

Given a dataset of $1, 2, 3, 4, 5$, the median and mean are both 3. However, if we have $1, 2, 3, 1000, 10000$, the mean is 2201.2 whereas the mean is still 3

2.3.2 Skewdness

All of these values can give information concerning the datas **skewness**

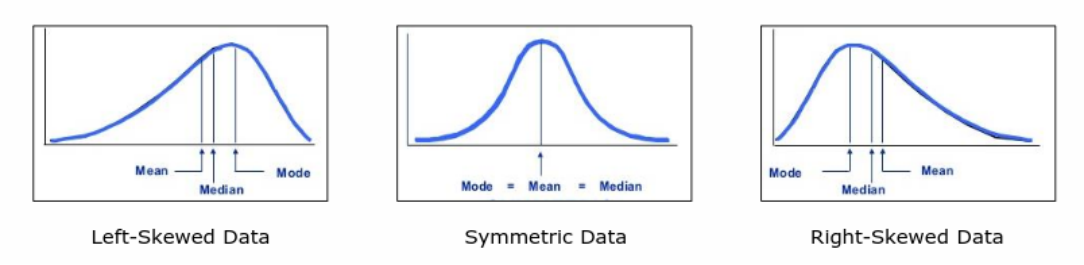


Figure 2.3: Skewness of data

$Mean - Mode > 0 \rightarrow$ Negative skewness / Left-skewed data
$Mean - Mode = 0 \rightarrow$ Symmetric Data
$Mean - Mode > 0 \rightarrow$ Positive skewness / Right-skewed data

2.3.3 Quartile & Interquartile Range (IQR)

The three quartiles divide your data into four equal-sized, cnsecutive subsets.

To calculate $Q1$, take the median of your data and then again the madian of the left half of the data.

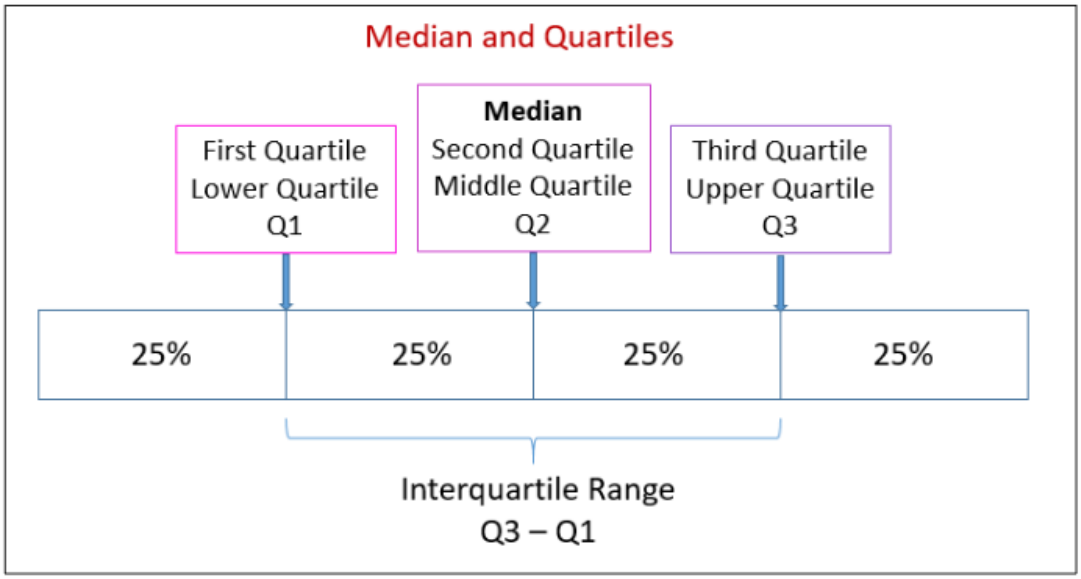


Figure 2.4: Quartiles of a dataset

2.3.4 Five Number Summary

With this method, you can get a pretty good overview of your data. The **Five Number Summary** of a dataset consists of:

- Median Q2
- Quartiles Q1 and Q2
- Smallest individual Value
- Largest individual Value

```
1  import numpy as np
2  import panas as pd
3
4  s = pd.Series(np.random.rand(100))
5  s.describe()
```

Listing 2.1: Five Number Summary in Python

```
1  mean       0.524559
2  std        0.285565
3  min        0.003933
4  25%        0.298367
5  50%        0.530632
6  75%        0.765907
7  max        0.993293
8  dtype: float64
```

Listing 2.2: Output

2.3.5 Boxplot

This plot is a **visual representation of the five number summary** and can also give information on potential outliers.

Values $1.5 \cdot IQR$ above the 3rd or below the 1st Quartile can be considered outliers and are displayed with small circles.

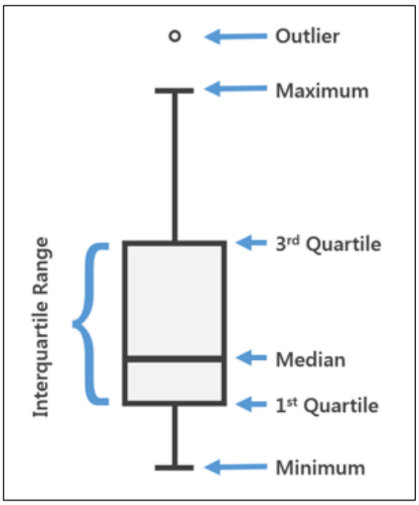


Figure 2.5: Boxplot

2.3.6 Variance

The variance shows **how much the values are spread on average**. This is measured by squaring the sum of all deviations from the mean

$$\frac{1}{1-n}\sum_{i=1}^{n}(x_i - \mu x)^2 \qquad (2)$$

The standard deviation σ is calculated as $\sqrt{variance}$

2.3.7 Covariance

The covariance is used to determine whether two variables are **connected** to each other.

If both variables are on the same side of the mean, the variance is **positive**, the variables are probably connected. Meaning if the value of one variable is rising, the other one will most likely rise as well.

If one is above and one is below the mean, the variance is **negative**, the variables are most likely **inversely connected** to each other. Meaning if the value of one variable is rising, the other is most likely falling.

If the variables are **independent** from each other, the covariance is zero, as they both cancel each other out.

$$Cov(x, y) = \frac{1}{1-n} \sum_{i=1}^{n} (x_i - \mu x)(y_i - \mu y) \tag{3}$$

The **covariance matrix** shows the covariance from all X with all Y. As $Cov(x, x) = Var(x)$, the covariance matrix has the variance of X in its diagonal

2.3.8 Pearson Correlation

Both the covariance and the variance are connected to the scale of the dataset, so the covariance of $X = [1, 2, 3, 4, 5]/Y = [6, 7, 8, 9, 10]$ is 2.5, whereas the covariance of $X = [1000, 2000, 3000, 4000, 5000]/Y = [6000, 7000, 8000, 9000, 10000]$ is 2'500'000'000. However, the Perason Correlation is 1 in both examples.

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_x \sigma_y} = \frac{\frac{1}{1-n} \sum_{i=1}^{n} (x_i - \mu x)(y_i - \mu y)}{\sqrt{\frac{1}{1-n} \sum_{i=1}^{n} (x_i - \mu x)^2} \cdot \sqrt{\frac{1}{1-n} \sum_{i=1}^{n} (y_i - \mu y)^2}} \tag{4}$$

The Pearson Correlation is always between 1 and -1.
1 means the data is perfectly correlated, whereas -1 means that the data is perfectly incorrelated

2.4 Normalization

It is immensely important that all data is normalized before we run a machine learning algorithm over it. Considering the data in figure 3.2, 'Mileage' and 'Price' are in a completely different scale. If the mileage of the first shown car goes up 500 miles, it's not really a big deal. However, a price increase by 500 would double the car's price.

Such differently scaled data can (and will) falsify the result of every machine learning algorithm you could find. Therefore, **normalization is really important**.

There are two popular normalization approaches: The Min-Max and the Z-Score normalization.

Min-Max normalization
All data is condensed to a value between 0 and 1. The smallest value becomes 0 and the largest one becomes 1.

$$x \to \frac{x - min_x}{max_x - min_x} \tag{5}$$

Z-Score Normalization
The dataset is transformed in such a way, that the mean becomes 0 (so-called *mean-centering*) and the standard deviation is 1

$$x \to \frac{x - \mu_x}{\sigma_x} \tag{6}$$

3 Geometry of Data

3.1 Feature Engineering

Sometimes, data has to be modified to be better accessible/processable for machine learning algorithms. These algorithmus can work the best with simple numbers, so that's the data we should be striving for:

Free	Date	Time
283	2015-09-27 00:00:00	06:26:46
282	2015-09-11 00:00:00	05:18:55
280	2015-09-20 00:00:00	21:14:49
283	2015-09-25 00:00:00	01:22:47
0	2015-10-15 00:00:00	08:12:35
0	2015-10-27 00:00:00	10:02:28
281	2015-09-13 00:00:00	12:20:54
168	2015-10-14 00:00:00	08:07:35
283	2015-09-25 00:00:00	05:42:47
283	2015-09-18 00:00:00	22:57:50
279	2015-09-10 00:00:00	20:26:55
279	2015-10-04 00:00:00	18:37:40
84	2015-09-17 00:00:00	17:17:51
86	2015-09-11 00:00:00	08:28:55
3	2015-10-26 00:00:00	13:51:28
281	2015-09-30 00:00:00	00:44:44
252	2015-10-15 00:00:00	07:19:35
280	2015-09-15 00:00:00	00:41:52
282	2015-09-09 00:00:00	06:05:56
0	2015-10-29 00:00:00	12:16:27

Free	Hour	Minute	Year	Month	Day
283	6	26	2015	9	27
282	5	18	2015	9	11
280	21	14	2015	9	20
283	1	22	2015	9	25
0	8	12	2015	10	15
0	10	2	2015	10	27
281	12	20	2015	9	13
168	8	7	2015	10	14
283	5	42	2015	9	25
283	22	57	2015	9	18
279	20	26	2015	9	10
279	18	37	2015	10	4
84	17	17	2015	9	17
86	8	28	2015	9	11
3	13	51	2015	10	26
281	0	44	2015	9	30
252	7	19	2015	10	15
280	0	41	2015	9	15
282	6	5	2015	9	9
0	12	16	2015	10	29

Figure 3.1: Turn 'complicated' data into easier data for better results

3.2 Vector Space Model

As described before, machine learning algorithms work best with **numeric** data. However, the real world isn't that easy and mostly throws categorical data at you. Therefore, you have to convert categorical data to numerical data.

This transformed data can also be visualized in a coordinate system and we can do math with it.

Name	Price	Mileage	Color
ALFA ROMEO 145 1.4 TS 16V L	500	187000	schwarz
ALFA ROMEO 145 1.8 TS 16V L	2600	182510	rot
ALFA ROMEO 145 1.9 JTD	3500	116000	grau
ALFA ROMEO 145 2.0 TS 16V Quadrifog.	4900	181000	rot
ALFA ROMEO 145 2.0 TS 16V Quadrifog.	800	121000	rot
ALFA ROMEO 145 2.0 TS 16V Quadrifog.	3200	156000	schwarz
ALFA ROMEO 146 2.0 Ti 16V	770	158000	grau
ALFA ROMEO 146 2.0 Ti 16V	1200	119000	rot
ALFA ROMEO 146 2.0 Ti 16V	4900	166000	schwarz
ALFA ROMEO 146 2.0 Ti 16V	4900	102000	silber
ALFA ROMEO 146 2.0 Ti 16V Kit Sport	5800	165000	schwarz
ALFA ROMEO 147 1.6 16V Blackline	11500	46230	braun

Name	Price	Mileage	braun	gelb	grau	grün	rot	schwarz	silber	weiss
ALFA ROMEO 145 1.4 TS 16V L	500	187000	0	0	0	0	0	1	0	0
ALFA ROMEO 145 1.8 TS 16V L	2600	182510	0	0	0	0	1	0	0	0
ALFA ROMEO 145 1.9 JTD	3500	116000	0	0	1	0	0	0	0	0
ALFA ROMEO 145 2.0 TS 16V Quadrifog.	4900	181000	0	0	0	0	1	0	0	0
ALFA ROMEO 145 2.0 TS 16V Quadrifog.	800	121000	0	0	0	0	1	0	0	0
ALFA ROMEO 145 2.0 TS 16V Quadrifog.	3200	156000	0	0	0	0	0	1	0	0
ALFA ROMEO 146 2.0 Ti 16V	770	158000	0	0	1	0	0	0	0	0
ALFA ROMEO 146 2.0 Ti 16V	1200	119000	0	0	0	0	1	0	0	0
ALFA ROMEO 146 2.0 Ti 16V	4900	166000	0	0	0	0	0	1	0	0
ALFA ROMEO 146 2.0 Ti 16V	4900	102000	0	0	0	0	0	0	1	0
ALFA ROMEO 146 2.0 Ti 16V Kit Sport	5800	165000	0	0	0	0	0	1	0	0
ALFA ROMEO 147 1.6 16V Blackline	11500	46230	1	0	0	0	0	0	0	0

Figure 3.2: Turn categorical data into numerical data with the vector space model

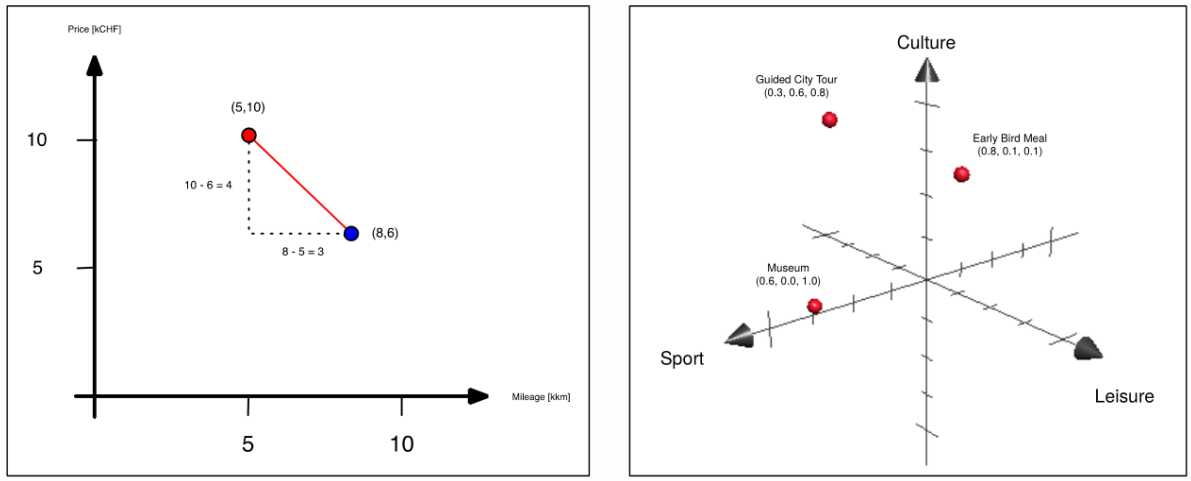


Figure 3.3: Transformed into vector space, data points can be interpreted as geometric points

3.3 Similarity of Data

The math we want to do is not even overly complicated: We just want to measure the distance between different points. Because **the smaller the distance between two points, the more similar they are**.

3.3.1 Euclidean Distance

The distance between two points is most easily calculated using the **euclidean distance**:

$$dist(X, Y) = \sqrt{\sum_{i=1}^{n}(x_i - yi)^2} \tag{7}$$

So the distance between the points (5/10) and (8/6) can be calculated as

$$\begin{aligned} &\sqrt{(5-8)^2 + (10-6)^2} \\ &\sqrt{-3^2 + 4^2} \\ &\sqrt{9 + 16} \\ &\sqrt{25} = 5 \end{aligned}$$

3.3.2 Cosine Similarity

If you want to to compare two points that appear to be on a line (Pearson Correlation close to 1), but the euclidean distance is high, then the cosine similarity is probably pretty low.

The cosine similarity looks at the **angle** between point A and point B. However, it does also take the euclidean distance into consideration.

The cosine similarity is essentially just the scalar product of the two points.

$$sim(X,Y) = \frac{\sum_{i=i}^{n} x_i y_i}{\sqrt{\sum_{i=i}^{n} {x_i}^2}\sqrt{\sum_{i=i}^{n} {y_i}^2}} \quad (8)$$

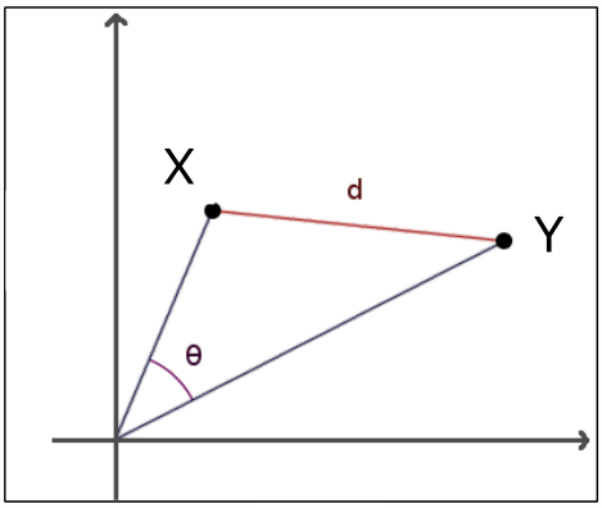


Figure 3.4: Cosine Similarity

$$dist(X,Y) = 1 - sim(X,Y) \quad (9)$$

3.3.3 Levenshtein / Edit Distance for Strings

Count the minimal number of changes necessary to turn one string into another:

- count +1 when deleting a character [d]
- count +1 when adding a character [a]
- count +2 when changing a character [c]

1. Word	2. Word	Levenshtein Distance
Hello	Yellow	1 [c] + 1 [a] = 3
MacDonald	McDonalds	1 [d] + 1 [a] = 2
banana	ananas	? d+a=2

Figure 3.5: Examples for Levenshtein Distance

4 Supervised Machine Learning

4.1 Regression and classification algorithms

Regression

- Linear Regression
- Polynomial Regression
- k-NN Regression
- Support Vector Regression
- Neural Networks
- Regression Trees

Classification

- Logistic Regression
- Naïve Bayes
- k-NN
- Support Vector Machines
- Neural Networks
- Decision Trees

4.2 Decision Boundaries

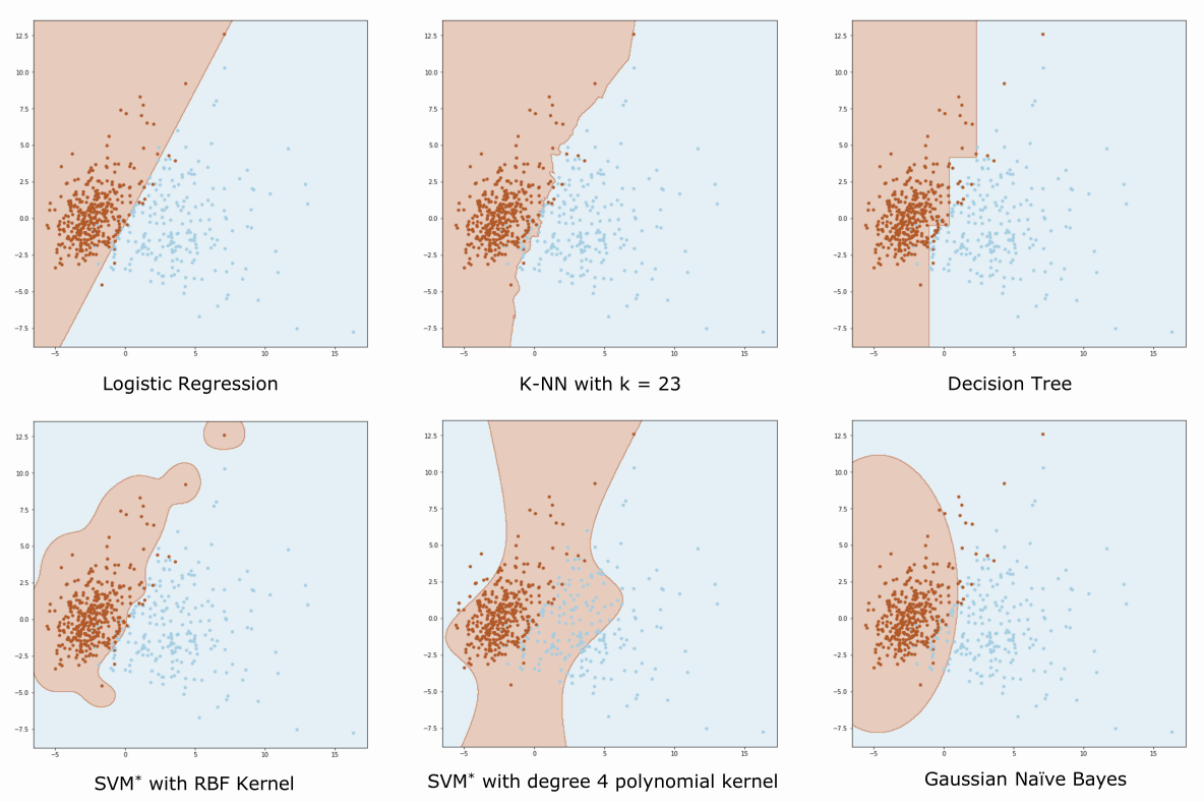


Figure 4.1: Decision Boundaries for different classification approaches

Classifications usually end in somithing like Figure 4.1. The example shows a classification whether a tumor is benign or cancerous. Brown means cancerous and blue means benign. Even though the data points are the same in all pictures, different approaches yield different results.

The goal of a 'good' classification-algorithm is to produce as few false-positive (algorithm says is cancer, but is actually not) and false-negatives (algorithm says its benign but is actually cancerous) as possible.

4.2.1 Kernel-Trick

The data in Fig. 4.1 is still theoretically linearly separable. But in case it is not, you could use the so-called 'kernel-trick', where you simply add a dimension and change your point of view (see Fig. 4.2)

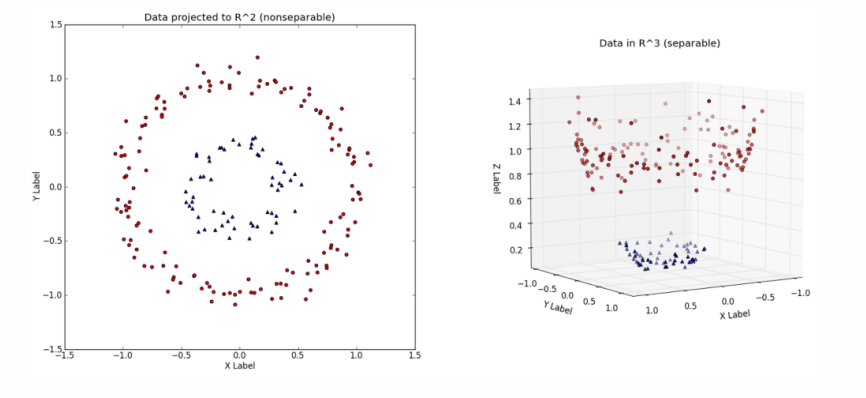
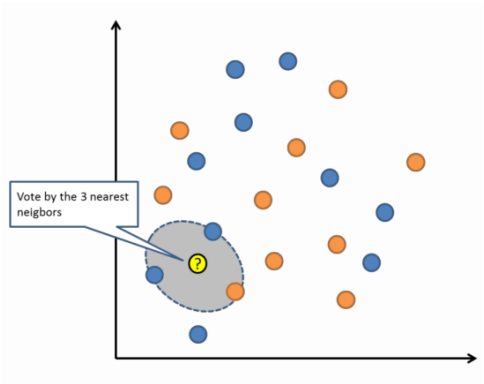


Figure 4.2: Kernel Trick to linearly separate data that is not linearly separable

4.3 k-Nearest-Neighbor

```
1  from sklearn.neighbors import
       KNeighborsClassifier
2
3  knn = KNeighborsClassifier(n_neighbors=3)
4  knn.fit(X_train, y_train)
5  y_pred = knn.predict(X_test)
6  acc = accuracy_score(y_test, y_pred)
7
8  print("Test Set Accuracy for k=3" + ":
       (:.2f)".format(acc))
```



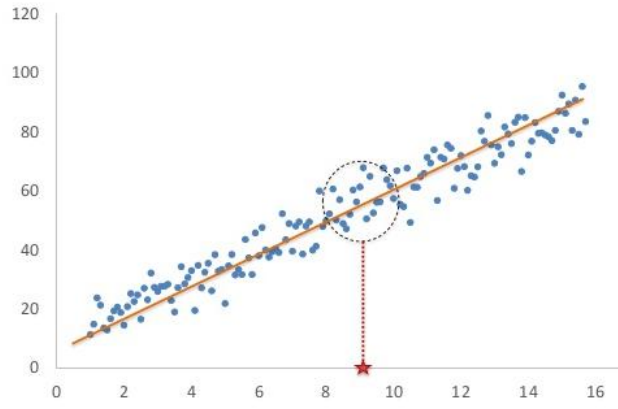


Figure 4.3: Regression with k-NN

The k-neeares-Neighbor algorithm assigns the label of its nearest datapoint to the sample datapoint. If `n_neighbors` is > 1, it assigns the label of the most neighbours (via majority voting).

Even though the K-NN algorithm is pretty slow compared to other algorithms, it is often a good choice if you only have a limited amount of data, because other, more performant algorithms usually require a lot more data to get decent predictions.

Usually, k-NN treats all neighbors as equals. However, you could assign a weight to each datapoint. This weight depends on the distance d from the sample point. This is especially useful if you want to use k-NN to e.g. form a regression line.

You can use k-NN for regression by simply assigning the mean of all k neighbors as label to the sample data.

4.4 Training- and Testdata

If you use the same data to train and test your algorithm, it might occur that the algorithm is 'memorizing' the data and gives you brilliant results. However, if you release it into the wild, where it encounters different data, it will perform really poorly. This is called **overfitting**

To counter overfitting, you usually split your data into **trainingdata** and **testdata**. You train the algorithm with the training data and test it with the testdata (who would've thought...).

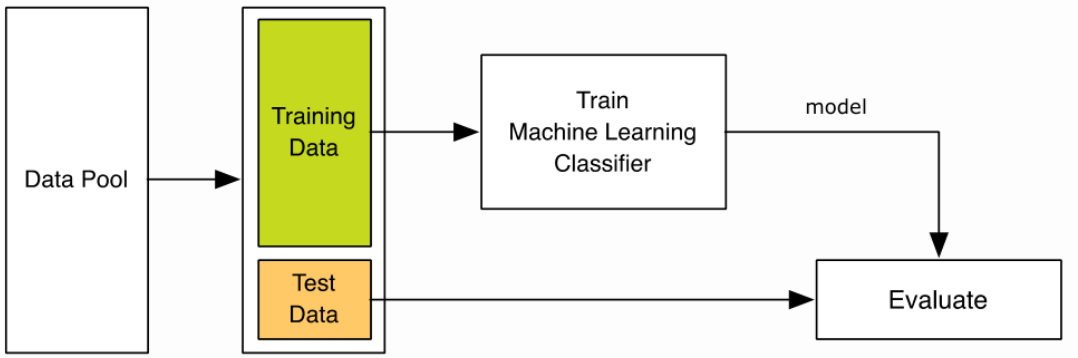


Figure 4.4: Split your data into training- and testdata

training means that you tweak the parameters of the algorithm to minimize the cost function.

testing means that you test the performance of the algorithm with those tweaked parameters on *yet unseen data*. If the algorithm has already seen the data, you might run into overfitting problems.

If you need to tweak your hyperparameters a lot (e.g. the 'k' in k-NN), you should probably use a more complex evaluation workflow. Because if you keep tweaking the hyperparameters and then testing them with the same data, you'll end up with the very same overfitting problem that I explained earlier (and will therefore probably get fired and have to live on the street).

Therefore, it is recommended that you add **validation data** to your workflow. You train your model with the training data, validate the results with the validatin data, and if the result is satisfactory, you can test it on entirely different test data.

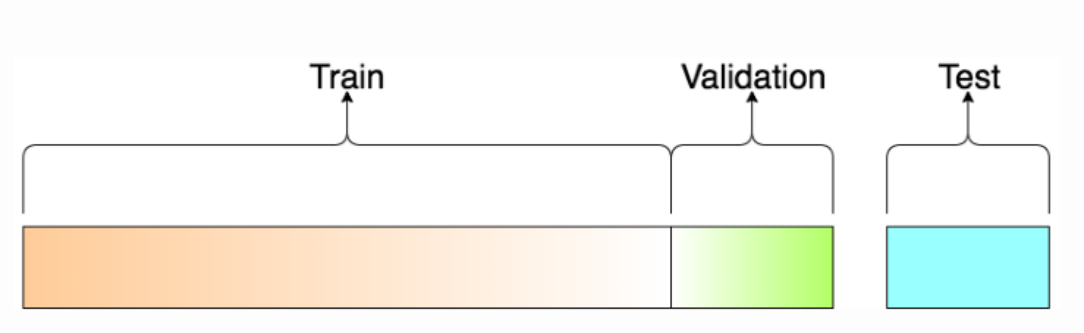


Figure 4.5: Add validation data to the mix

This method requires quite a lot of data. If you do not have the required amount of data, you could for example use **cross validation**. You still split your data into training- and testdata and then use a different 'slice' of your training data to validate the hyperparameter.

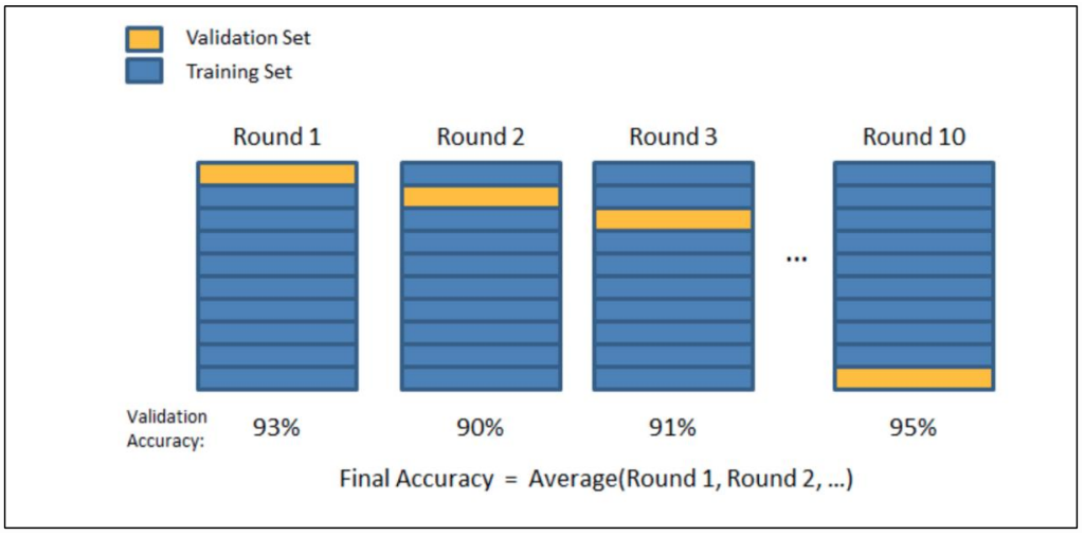


Figure 4.6: 10-fold cross validation

4.5 Measuring the performance of classification

To verify that your tweaked parameters are indeed within the margin that is acceptable, you need to do some quality assuranc first.

4.5.1 Confusion Matrix

n=165	Predicted YES	Predicted NO
Actual No	50	10
Actual YES	5	100

True Positive: Predicted Yes, Actual Yes

True Negative: Predicted No, Actual No

False Positive: Predicted Yes, Actual No

False Negative: Predicted No, Actual Yes

The confusion matrix shows, how many true/false positives and true/false negatives the algorithm produced.

With these values, one can calculate the algorithms **Accuracy** and **Error Rate**

4.5.2 Accuracy and Error Rate

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total} \tag{10}$$

$$Error\ Rate = \frac{False\ Positive + False\ Negative}{Total} = 1 - Accuracy \tag{11}$$

In our case the Accuracy would be $\frac{50+100}{50+100+5+10} = \frac{150}{165} = 0.91 = 91\%$ and the error rate therefore 9%

4.5.3 Sensitivity

Accuracy works great on balanced data, but it's not reliable on inbalanced data because Accuracy only checks how many times the classifier was right.

If there were 5000 NO instances and 20 YES instances, a classifier that only returns NO would have an accuracy of over 99%.

The **Sensitivity** (also called 'Recall') counts how many true positives there are.

$$Sensitivity = \frac{True\ Positive}{Actual\ YES} = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{12}$$

In our confusion matrix from earlier, the Sensitivity would be $\frac{100}{100+5} = \frac{100}{105} = 0.95 = 95\%$

4.5.4 Sepcificity

This is the inverse of the Sensitivity. It counts how many NO the algorithm correctly predicted.o

$$Specificity = \frac{True\ Negative}{Actual\ NO} = \frac{True\ Negative}{True\ Negative + False\ Positive} \tag{13}$$

In our confusion matrix from earlier, the Specificity would be $\frac{50}{50+10} = \frac{50}{60} = 0.83 = 83\%$

4.5.5 Precision

Both of the preceding measures relied on the true negative. However, what would you do if you could not count the True Negatives? You use **Precision**. Precision shows how many times the algorithm is correct if it predicts YES.

$$Precision = \frac{True\ Positive}{Predicted\ YES} = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{14}$$

In our confusion matrix from earlier, the Precision would be $\frac{100}{100+10} = \frac{100}{110} = 0.91 = 91\%$

4.5.6 F1 Score

The F1 score is the harmonic mean between precision and recall/sensitivity

$$F1 = \frac{2 \cdot Precision \cdot Sensitivity}{Precision + Sensitivity} \tag{15}$$

In our confusion matrix from earlier, the Precision would be $\frac{2 \cdot 0.91 \cdot 0.95}{0.91+0.95} = \frac{1.73}{1.86} = 0.93 = 93\%$

Due to the fact that the F1 score does not take True Negatives into account, it tends to be strongly biased towards the worse score.

However, it is still one of the best methods to solve a classification problem with skewed data.

4.6 Measuring the performance of regression

4.6.1 Coefficient of Determination

$$R^2 = 1 - \frac{\frac{1}{m}\sum_{i=1}^{m}(y_i - f_i)^2}{\frac{1}{m}\sum_{i=1}^{m}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{m}(y_i - f_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{16}$$

- The sum of errors does not make sense (positive and negative errors cancel out)

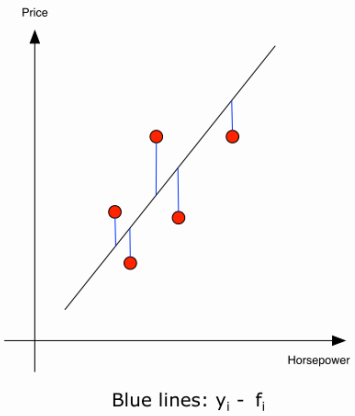
$$\frac{1}{m}\sum_{i=1}^{m}(y_i - f_i)$$

- Mean Absolute Error

$$\frac{1}{m}\sum_{i=1}^{m}|y_i - f_i|$$

- Mean Squared Error

$$\frac{1}{m}\sum_{i=1}^{m}(y_i - f_i)^2$$



R^2 is a staticstical measure of how well the predictions approximate the real data points. The top is the sum of squared errors (how much does the prediction deviate from the actual value) and the botom is the deviation of the mean.

$R^3 = 1$ is a perfect prediction-line

$R^3 = 0.53$ means that 53% of the of the predictions are correct and can be explained by the model.

5 Linear Regression

Linear regression is a **supervised machine learning algorithm** to predict values based on previous variables.

It can also be used to verify whether two variablex X and Y are **dependant** on each other.

Linear Regression produces a **straight line** (who would have thought) and each data point has a certain distance from that line. That distance is called **error** or **residual**.

These errors should cancel each other out, as the regression-line should be placed in the middle of all the points.

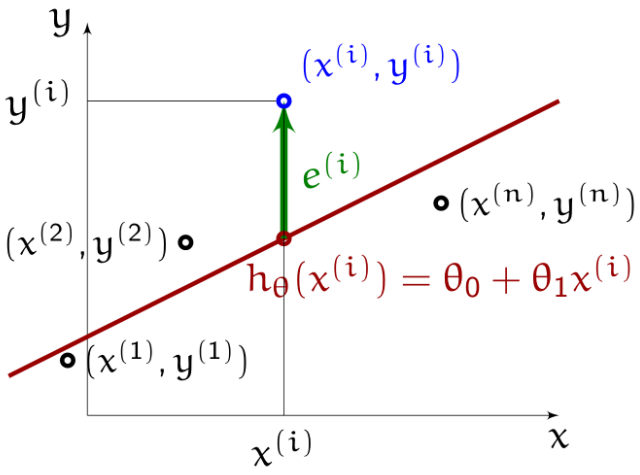


Figure 5.1: Linear Regression

The equation for a 'normal' line is $g(x) = m \cdot x + q$, where m symbolizes the slope of the line and q tells you where the line crosses the y-axis.

The equation for the regression line is very much like the 'normal' line-equation, but we use greek symbols, because why not...

$$h_\theta(x) = \theta_0 + \theta_1 \cdot x \tag{17}$$

As mentioned before, deviation from that line are called errors. Therefore, the error of a given point i can be calculated from the difference of the actual y_i from the $h_\theta(x)$ (the y on the line)

$$e_i = y_i - (\theta_0 + \theta_1 \cdot x_i) \tag{18}$$

The goal of linear regression is to find the **optimal line**, ergo the line that **produces the smallest errors**.

However, as mentioned earlier, the errors usually cancel each other out, as sometimes the line lies above the data point (negative error) and sometimes it lies below it (positive error). To solve that problem, the parameter we will use to determine the quality of the produced line will be **the sum of squared error**. By squaring the errors, we can't have negative errors (because squared numbers cannot be negative). Therefore, the smaller the sum of squared errors, the better the line.

The sum of squared errors can be calculated as follows:

$$J(\theta_0, \theta_1) = \frac{1}{2n} \sum_{i=1}^{n} (e_i)^2 = \frac{1}{2n} \sum_{i=1}^{n} [y_i - (\theta_0 + \theta_1 \cdot x_i)]^2 \tag{19}$$

To minimize $J(\theta_0, \theta_1)$, the **gradient** of J must be **minimized** or, ideally, zeroed.
This leads to the following formulas for θ_0 and θ_1

$$\theta_1 = \frac{S_{xy}}{S_{xx}} \tag{20}$$

$$\theta_0 = \bar{y} - \theta_1 \cdot \bar{x} \tag{21}$$

Where $\bar{x}$ and $\bar{y}$ are the **mean** of the x and y values respectively

S_{xx} and S_{xy} are called **regression coefficients** and are calculated as follows:

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad (22) \qquad\qquad S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (23)$$

Example

$X = [4, 6, 8, 10]$
$Y = [2.3, 4.1, 5.7, 6.9]$
$\bar{X} = \frac{1}{4}(4 \cdot 6 \cdot 8 \cdot 10) = 7$
$\bar{Y} = \frac{1}{4}(2.3 \cdot 4.1 \cdot 5.7 \cdot 6.9) = 7.25$
$S_{xx} = (4-7)^2 + (6-7)^2 + (8-7)^2 + (10-7)^2 = 20$
$S_{xy} = (4-7)(2.3-7.25) + (6-7)(4.1-7.25) + (8-7)(5.7-7.25) + (10-7)(6.9-7.25) = 15.4$
$\theta_1 = \frac{20}{15.4} = 0.77$
$\theta_0 = 4.75 - (0.77 - 7) = -0.64$

From this follows that the regression line is:

$$h_\theta(x) = -0.64 \cdot 0.77 \cdot x$$

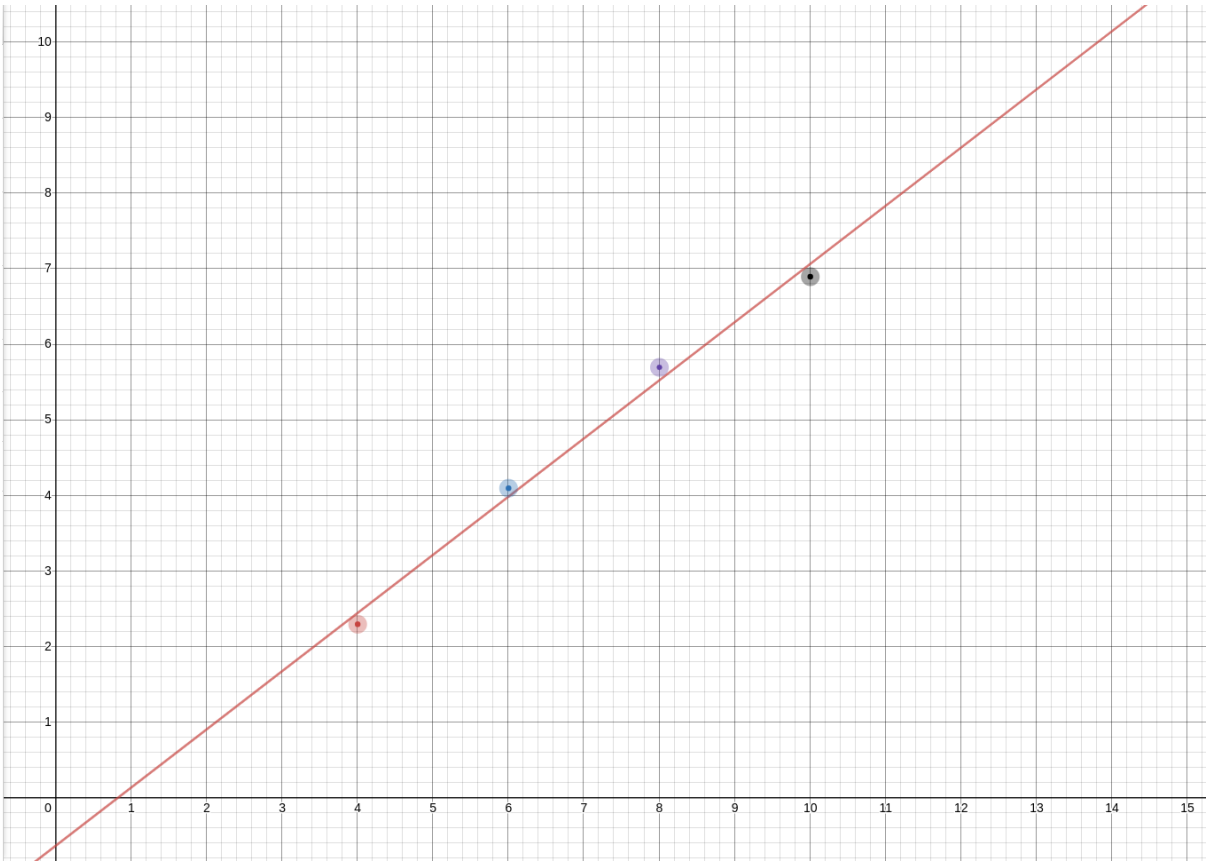


Figure 5.2: Regression line of the example

From figure 5.2 we can now see that for $x_i = 12$, y would most likely be at 8.6.

5.1 Coefficient of Determination

This coefficients was already mentioned in section 4.6.1 as 'way to measure how good a regression is'. But how does it even do that?

As mentioned earlier, the goal of a good regression is to **minimize the sum of squared errors**. However, there are three kinds of sum of squared Errors:

$\bar{y}$**:** Mean of all Y

$\bar{x}$**:** Mean of all X

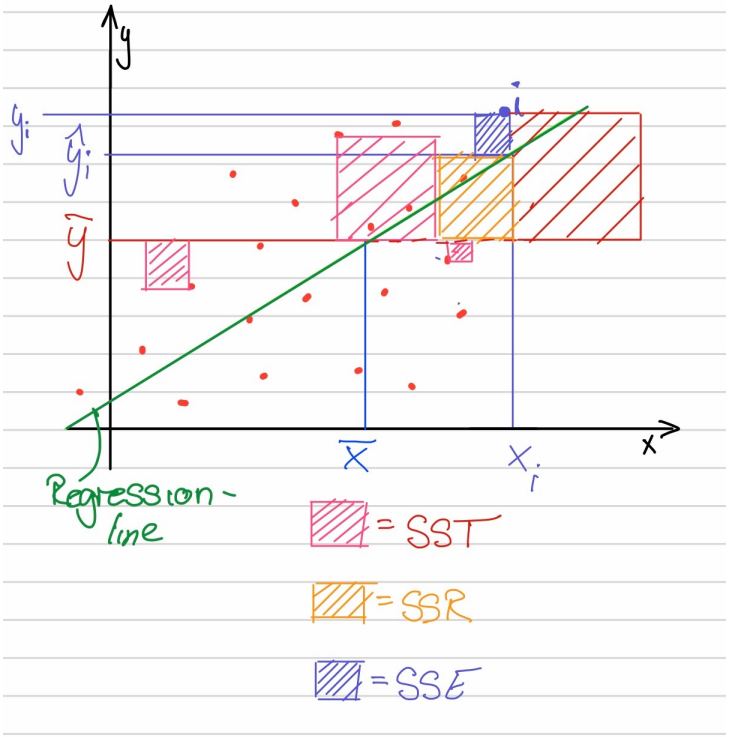
$\hat{y}_i$**:** Expected value of y based on the regression line

y_i**:** Actual value of y

SST : *Total sum of squares*
Sum of SSE and SSR

SSE : *Sum of squared Errors*
The sum of the deviations from the predicted points to the regression line

SSR : *Sum of squares explained by regression*
The sum of all deviations from the mean ($\bar{y}$) to the regression line



As illustrated in the figure above, $SST = SSR + SSE$.

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2 \qquad SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

But how does the aforementioned Coefficient of Determination relate to all of this?

R^2 is the **fraction of the total error that can be explained by the regression**. If all errors can be explained by the regression ($R^2 = 1$), then the regression is perfect

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} \tag{24}$$

There's yet another error, the *Mean Squared Error* (MSE). MSE is basically the mean of SSE. It is calculated as

$$MSE = \frac{SSE}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{25}$$

5.2 Correlation Analysis

So now thath you have the regression parameters θ_0 and θ_1. which make out the regression line. However, that regression line is only as good as your parameters. So how do you test how well you calculated the parameters?

Easy, you calculate the **standard deviation of these parameters**.

$$s_{\theta_0} = \sqrt{MSE} \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{26}$$

$$s_{\theta_1} = \sqrt{MSE} \cdot \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{27}$$

Based on these standard deviations, you can calculate a **confidence interval**. This interval is a **degree of uncertainty**. Say we conducted a study and we publish our results with a **confidence level** of 95%.

This means that if we used the same sampling method to select different samples and computed an interval estimate for each sample, we would expect the true population parameter to fall within the interval estimates 95% of the time.

These intervals can be calculated with the aforementioned standard deviation of θ_0 and θ_1 and some *critical values*. These critical values depend on how high your confidence level and degrees of freedom are.

Kritische Grenzen der t-Verteilung

Kritische Grenzen $t_{1-\frac{\alpha}{2}}(\nu)$ der t-Verteilung mit ν Freiheitsgraden.

ν	1	2	3	4	5	6	7	8	9	10	11	12
$\alpha = 0.05$	12.71	4.30	3.18	2.78	2.57	2.45	2.36	2.31	2.26	2.23	2.20	2.18
$\alpha = 0.01$	63.66	9.92	5.84	4.60	4.03	3.71	3.50	3.36	3.25	3.17	3.11	3.05
ν	13	14	15	16	18	20	25	30	40	60	100	200
$\alpha = 0.05$	2.16	2.14	2.13	2.12	2.10	2.09	2.06	2.04	2.02	2.00	1.98	1.97
$\alpha = 0.01$	3.01	2.98	2.95	2.92	2.88	2.85	2.79	2.75	2.70	2.66	2.63	2.60

Für $\nu \to \infty$ konvergieren die Grenzen gegen die kritischen Grenzen 1.96 bzw. 2.58 der Normalverteilung.

Figure 5.3: Critical Values for Confidence Intervals

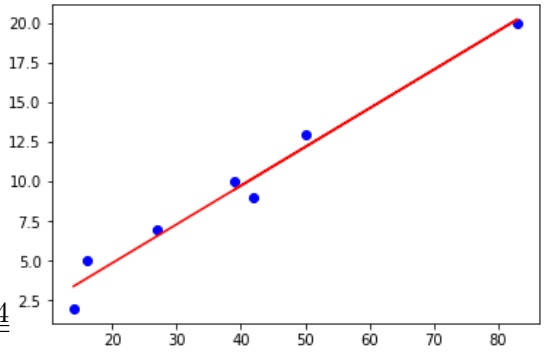
Degrees of freedom tell you how many regression parameters you already had to calculate. Take for example SSE,where both θ_0 and θ_1 are needed. Therefore, SSE has a degree of freedom of $n - 2$, given that you had to calculate 2 parameters.

Confidence Intervals can be calculated with $[\theta_0 - (critical\ value \cdot \theta_0); \theta_0 + (critical\ value \cdot \theta_0)]$

5.3 Linear Regression Example

$X = [14, 16, 27, 42, 83, 50, 39]$

$Y = [2, 5, 7, 9, 20, 13, 10]$



$$\bar{X} = \frac{1}{7}\sum_{i=1}^{7}(x_i) = \frac{14+16+27+42+83+50+39}{7}$$

$$= \frac{271}{7} = \underline{\underline{38.7}}$$

$$\bar{Y} = \frac{1}{7}\sum_{i=1}^{7}(y_i) = \frac{2+5+7+9+28+13+10}{7} = \frac{66}{7} = \underline{\underline{9.4}}$$

$$S_{xy} = \sum_{i=1}^{7}(x_i - \bar{x})(y_i - \bar{y}) = (14-38.7)(2-9.4) + (16-38.7)(5-9.4) + (27-38.7)(7-9.4)$$
$$+ (42-38.7)(9-9.4) + (83-38.7)(20-9.4) + (50-38.7)(13-9.4)$$
$$+ (39-38.7)(10-9.4) = \underline{\underline{819}}$$

$$S_{xx} = \sum_{i=1}^{7}(x_i - \bar{x})^2 = (14-38.7)^2 + (16-38.7)^2 + (27-38.7)^2 + (42-38.7)^2 + (83-38.7)^2$$
$$+ (50-38.7)^2 + (39-38.7)^2 = \underline{\underline{3363.4}}$$

$$\theta_1 = \frac{S_{xy}}{S_{xx}} = \frac{819}{3363.4} = \underline{\underline{0.24}}$$

$$\theta_0 = \bar{y} - \theta_1 \cdot \bar{x} = 9.4 - 0.24 \cdot 38.7 = \underline{\underline{-0.008}}$$

$$SSE = \sum_{i=1}^{7}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{7}(y_i - (\theta_0 + \theta_1 \cdot x))^2 = \sum_{i=1}^{7}(y_i - (-0.008 + 0.24 \cdot x))^2$$

$$= \sum_{i=1}^{7}(y_i - (-0.232 \cdot x))^2 = \underline{\underline{5.87}}$$

$$SSR = \sum_{i}^{7}(\hat{y}_i - \bar{y})^2 = \sum_{i}^{7}((\theta_0 + \theta_1 \cdot x) - 38.7)^2 = \sum_{i}^{7}((-0.008 + 0.24 \cdot x) - 38.7)^2$$

$$= \sum_{i}^{7}((-0.232 \cdot x) - 38.7)^2 = \underline{\underline{199.84}}$$

$$SST = SSE + SSR = 5.87 + 199.84 = \underline{\underline{205.71}}$$

$$R^2 = \frac{SSR}{SST} = \frac{199.84}{205.71} = \underline{\underline{0.97}}$$

$$MSE = \frac{SSE}{n-2} = \frac{5.87}{7-2} = \underline{\underline{1.17}}$$

$$s_{\theta_0} = \sqrt{MSE} \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{1.17} \cdot \sqrt{\frac{1}{7} + \frac{38.7^2}{\sum_{i=1}^{n}(x_i - 38.7)^2}} = \underline{\underline{-0.008}}$$

$$s_{\theta_1} = \sqrt{MSE} \cdot \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{1.17} \cdot \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - 38.7)^2}} = \underline{\underline{0.019}}$$

$$Interval\ \theta_0 = [\theta_0 - 2.57 \cdot \theta_0; \theta_0 + 2.57 \cdot \theta_0] = \underline{\underline{[-2.14; 2.12]}}$$

$$Interval\ \theta_1 [\theta_1 - 2.57 \cdot \theta_1; \theta_1 + 2.57 \cdot \theta_1] = \underline{\underline{[0.20; 0.29]}}$$

The results of the last page can also be achieved with the following python-code:

```
import matplotlib.pyplot as plt
import scipy.stats as st
import seaborn as sns
import pandas as pd
import numpy as np

%precision 10
%matplotlib inline

x=np.array([14, 16, 27, 42, 83, 50, 39])
y=np.array([2, 5, 7, 9, 20, 13, 10])

mean_x = np.mean(x)
mean_y = np.mean(y)

Sxx = np.sum((x-mean_x)**2)
Syy = np.sum((y-mean_y)**2)
Sxy = np.sum((x-mean_x)*(y-mean_y))

thet1 = Sxy/Sxx
thet0 = mean_y-thet1*mean_x

# To show the regression line
plt.plot(x,y, 'bo')
plt.plot(x,thet0+thet1*x, 'r')
plt.show()

hat_y = thet0+thet1*x

SSE=np.sum(((y-hat_y))**2)

SSR=np.sum((hat_y-mean_y)**2)

SST = SSE + SSR

R_sq = SSR/SST

MSE = SSE/(len(x)-2)

Sthet0 = np.sqrt(MSE)*np.sqrt((1/len(x))+(mean_x**2)/(np.sum((x-mean_x)**2)))

Sthet1 = (np.sqrt(MSE))*np.sqrt(1/(np.sum((x-mean_x)**2)))

print("theta_1: " + str(thet1))

print("Interval theta0: [" + str(thet0 - (2.57*Sthet0)) + " ; " + str(thet0 +
    (2.57*Sthet0)) + "]")

print("Interval theta1: [" + str(thet1 - (2.57*Sthet1)) + " ; " + str(thet1 +
    (2.57*Sthet1)) + "]")
```

5.4 Multple Linear Regression

Linear Regeression is great and all, but what if you wanted to predict an independent value based on **multiple** other values? In that case we would need **multilinear regression**. In Muliple Linear Regression we try to fit a plane instead of a line. This plane is defined by

$$y = h_\theta(x_1, x_2) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \tag{28}$$

And mapped to the sample with index i

$$y_i = \theta_0 + \theta_1 x_{1,i} + \theta_2 x_{2,i} + e_i$$

This can be transformed to matrices

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ & \vdots & \\ 1 & x_{1,n} & x_{2,n} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

This can be solved as follows

$$\boldsymbol{\theta} = (\boldsymbol{X}^T \boldsymbol{X})^- 1 \boldsymbol{X}^T \boldsymbol{y}$$

In more than three dimensions our plane becomes a hyperplane, and the model is

$$y_i = \theta_0 + \theta_1 x_{1,i} + \theta_2 x_{2,i} + \cdots + \theta_m x_{m,i} + e_i = \sum_{k=0}^{n} \theta_k x_{k,i} + e_i$$

5.4.1 Example multilinear regression

Suppose we want to predict the weight of a weightlifter based on the training hours per week and the delivery of protein.

i	**y**	x_1	x_2
1	93	2	1.1
2	106	2	1.9
3	146	4	2
4	140	5	1.5
5	151	6	1.3
6	158	7	2.1
7	130	4	1.8
8	159	5	2.5

i**:** No. of observation

y**:** Weight in kg

x_1**:** Training h/week

x_2**:** Protein intake g/kg/day

We assume that the function to calculate the weight from training-hours and protein intake is as follows:

$$y = h_\theta(x_1, x_2) = \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2$$

The following solution minimizes the sum of squared errors in this model (and therefore get a hella good regression).

$$\theta = (X^T X)^{-1} X^T y = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} \rightarrow \begin{bmatrix} 55.7 \\ 11.1 \\ 17.5 \end{bmatrix}$$

According to this, the regression plane would be:

$$y = h_\theta(x_1, x_2) = 55.7 + 11.1 \cdot x_1 + 17.5 \cdot x_2$$

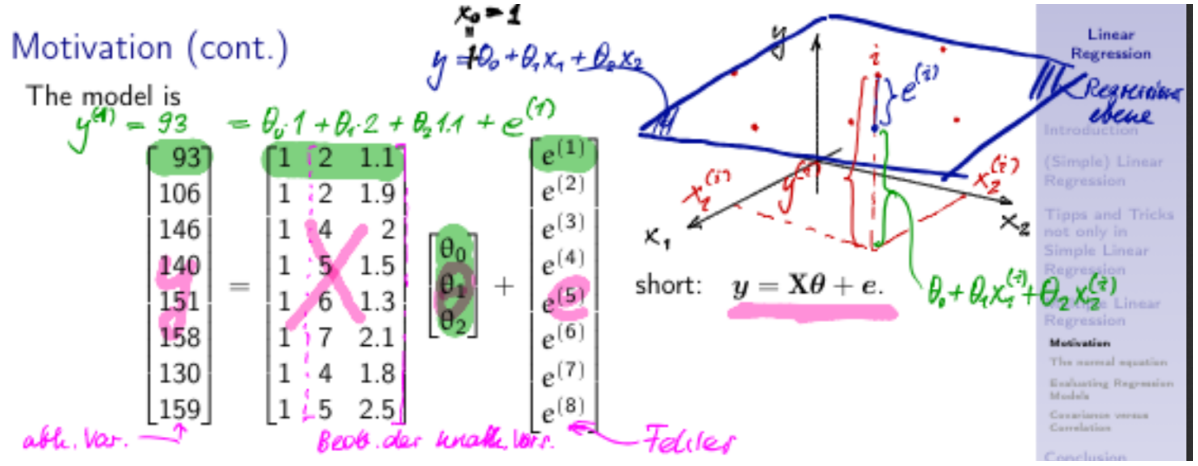


Figure 5.4: Model of the multilinear regression (©J. Bürgler, 2019)

Of course, this can also be done rather easily in Python:

```
import numpy as np

X = np.matrix([[2,1.1],[2,1.9],[4,2.0],[5,1.5],[6,1.3],[7,2.1],[4,1.8],[5,2.5]])

# Anzahl Reihen von X anhand der ersten Zeile (= Anzahl Beobachtungen)
n = X.shape[0]

y = np.matrix([93,106,146,140,151,158,130,159])

# Append a column of '1' infront of the X-matrix (to get the 'X'-matrix from the
    figure)
X_ext = np.c_[np.ones((n,1)),X]

theta = np.linalg.inv(X_ext.T.dot(X_est)).dot(X_ext.T).dot(y)

print(theta)
```

6 Gradient Descent

The problem with Linear Regression is, that it does not scale well. For problems like these we can use Gradient Descent. The gradient (denoted with the Nabla Operator ∇) uses the properties of partial derivatives, which compute the slope in regards to an axis in a point a, b.

The gradient ∇f of the scalar function $f : \mathbb{R}^2 \to \mathbb{R}, (x, y) \mapsto f(x, y) = z$ is a vector, whose components are partial derivatives of f.

$$\nabla f(x, y) = \begin{bmatrix} \frac{\delta f}{\delta x}(x, y) \\ \frac{\delta f}{\delta z}(x, y) \end{bmatrix} = \begin{bmatrix} f_x(x, y) \\ f_y(x, y) \end{bmatrix} \tag{29}$$

6.0.1 Example of a gradient

Calculate the gradient of the function $f(x, y) = (3x + 2y)^2$ in general and in the point $(1, 2)$.

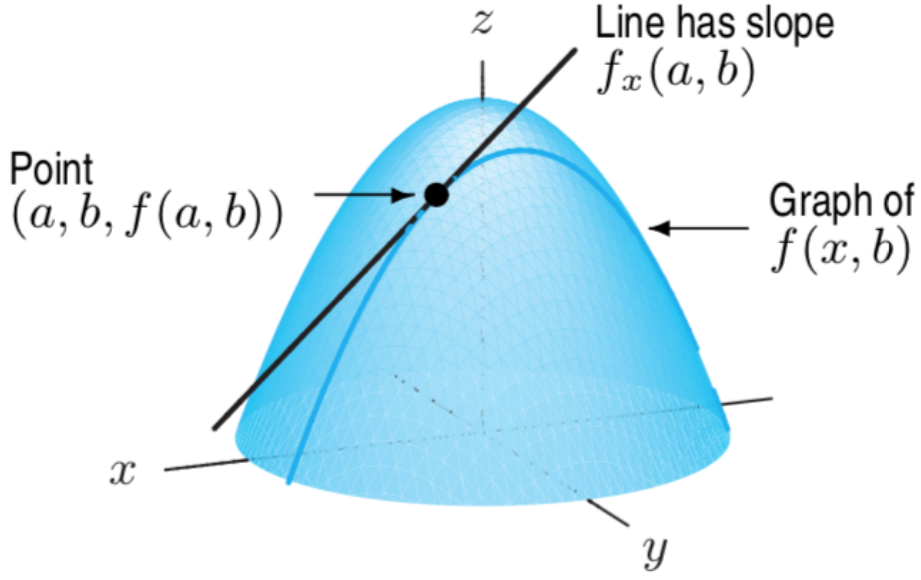


Figure 6.1: Partial derivative in regards to the x-axis in the point a,b (©J. Bürgler, 2019)

$$\nabla f(x,y) = \begin{bmatrix} 6(3x+2y) \\ 4(3x+2y) \end{bmatrix}$$
$$\nabla f(1,2) = \begin{bmatrix} 42 \\ 28 \end{bmatrix}$$

6.1 Properties of the gradient

- The Gradient of a function f is a vector consisting of the partial derivatives of f with respect to all of its variables
- It points to the direction of maximal increase
- The negative Gradient points to the direction of maximal decrease
- The Gradient being a null-vector indicates a local extrema
- The Gradient is perpendicular to the contour lines of that function

6.2 Gradient descent

To find a local minima of a function f we follow the direction of the steepest descent, and thus the negative Gradient. As the Gradient is constantly changing, we have to take small steps.

6.2.1 Batch Gradient Descent

Elevating the strengths of vectors we can Gradient Descent along all columns of our data. For this we want to minimize the cost function

$$J(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (h(bm\theta, \mathbf{x}^{(i)}) - y^{(i)})^2 \tag{30}$$

where $h(\boldsymbol{\theta}, \mathbf{x}^{(i)})$ is

$$h(\boldsymbol{\theta}, \boldsymbol{x}^{(i)}) = \boldsymbol{x}^{(i)}\boldsymbol{\theta} = \theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_m x_m^{(i)} \tag{31}$$

And calculate the next iteration with the learning rate α (usually $0 \leq \alpha \leq 0.1$)

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \nabla J(\boldsymbol{\theta}_k) \tag{32}$$

It can be shown that this yields

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \frac{1}{n}(h(\boldsymbol{\theta}, \boldsymbol{x}^{(i)}) - y^{(i)}) \cdot \boldsymbol{x}^{(i)} \tag{33}$$

6.3 Stochastic Gradient Descent

In Batch Gradient Descent the update to the parameter vector $\boldsymbol{\theta}$ is done all at once using the whole set of n training samples. If n is in the order of several 10^6, this update process can become slow. One possible alternative is to update $\boldsymbol{\theta}$ for each training sample separately. This is done by randomly shuffling the training examples first and then updating $\boldsymbol{\theta}$ for each training sample.

6.3.1 Instructions

- Choose an initial parameter vector $\boldsymbol{\theta}$ and a learning rate α
- Repeat until an approximate miminum is obtained
  - Randomly shuffle the samples in the training set
  - For each sample $i \in 1, 2, ..., n$ do

    $$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \nabla J_i(\boldsymbol{\theta}_k) = \boldsymbol{\theta}_k - \alpha \frac{1}{n}(h(\boldsymbol{\theta}_k, \boldsymbol{x}^{(i)}) - y^{(i)}) \cdot \boldsymbol{x}^{(i)}$$

    where $h(\boldsymbol{\theta}_k, \boldsymbol{x}^{(i)}) = \boldsymbol{\theta}_k^T \boldsymbol{x}^{(i)}$ and

    $$J_i(\boldsymbol{\theta}) = \frac{1}{2n}(h(\boldsymbol{\theta}_k, \boldsymbol{x}^{(i)}) - y^{(i)})^2$$

    is the cost for one data point $(x^{(i)}, y^{(i)})$

The difference to Batch Gradient Descent is that only one piece of data from the dataset is used to calculate the updated parameter and this piece of data is choosen randomly.

6.3.2 Stochastic Gradient Descent - Example

```
def parallel_cost(X,Y,x_data,y_data):
    m = X.shape[0]; n = X.shape[1]
    tot = np.zeros((m,n))
    for i in range(1,len(x_data)):
        tot += (X + Y * x_data[i] -
            y_data[i]) ** 2
    return tot/(2 * len(x_data))

alpha = 0.03; epochs = 2000

for epoch in np.arange(0,epochs):
    arr = np.arange(X.shape[0])
    np.random.shuffle(arr)
    for sample in np.arange(0,X.shape[0]):
        i = arr[sample]
        preds_i = X[i].dot(theta)
        error_i = preds_i - y[i]
        grad_i =
            X[i].dot(error_i)/X.shape[0]
        theta = theta - alpha * grad_i
        theta0_path.append(theta[0])
        theta1_path.append(theta[1])
```

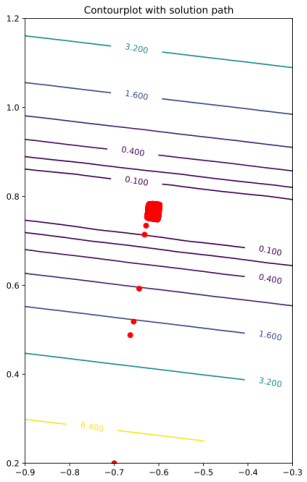


Figure 6.2: Stochastic Gradient Descent

6.4 Polynomial Regression, Feature Scaling and Checking Convergence

If we do not have a linear relationship, but a polynomial one between the features and estimated outcome, our equation is in the general form of

$$h(\boldsymbol{\theta}, \boldsymbol{x}) = \theta_0 + \theta_1 x^1 + \theta_2 x^2 + \cdots + \theta_n x^n \tag{34}$$

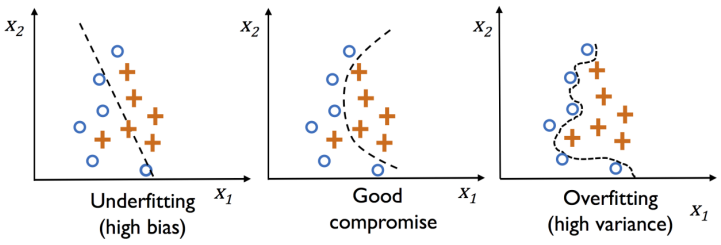
We could then use the features $x_1 = x$, $x_2 = x^2$ and $x_3 = x^3$, but if the features are not in the same range, we have a problem. To solve this we must use feature scaling:

1. Make sure the features are on a similar scale and get every feature into the range of $-1 \leq x_i \leq 1$
   (a) Replace x_i with $x_i - \bar{x}_i$ to have features with an approximately zero mean. Does not apply to $x_0 = 1$
   (b) Divide $x_i - \bar{x}_i$ by the range of $x_i = max(x_i) - min(x_i)$ or the standard deviation of x_i
2. Test if Gradient Descent is working correctly by plotting the cost function $J(\boldsymbol{\theta})$ versus the number of iterations
3. Declare convergence if $J(\boldsymbol{\theta})$ decreases by less than a given threshold (for example 10^-3) in one iteration

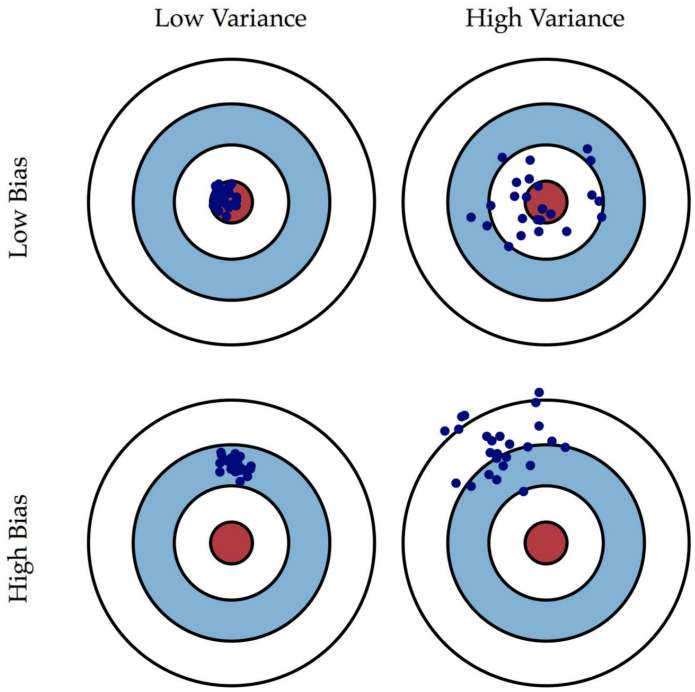
For sufficiently small α, $J(\boldsymbol{\theta})$ should decrease in every iteration, but a too small α leads to a slow convergence. Start with $\alpha = 0.001$, proceed with $0.003, 0.01, 0.03, 0.1, \ldots$.

7 Regularisation

Under- and overfitting are common problems in machine learning. Underfitting occurs when the model is to general and can't aptly represent the data. Overfitting occurs, when the model is too precisely fitted to the training data.



(a) Example of different fittings



(b) Examples of different biases and variances

Figure 7.1: Under- and Overfitting, Biases and Variances

To address overfitting we can try these approaches

1. Reduce the number of features
    - Manually select which features to keep
    - Model selection algorithm
2. Regularisation
    - Keep all features, but reduce magnitude of parameters θ_i
    - Works well with lots of features which each contribute just a bit to predict y

7.1 Ridge Regularisation

$$J(\theta) = \frac{1}{2n}\left[\sum_{i=1}^{n}(h(\theta, x_i) - y_i)^2 + \lambda\sum_{j=1}^{m}\theta_j^2\right] \tag{35}$$

The second sum goes from 1 to the number of features m and it does not contain θ_0. The regularisation hyperparameter λ controls the amount of regularisation. If $\lambda = 0$ there will be no regularization, and if $\lambda = \infty$ there will be underfitting because only θ_0 will be different from zero. In Ridge regularised regression, we choose θ to minimize $J(\theta)$:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} J(\theta) \tag{36}$$

large λ	High Bias / Underfit
intermediate λ	Correct
small λ	High Variance / Overfit

7.2 How to choose the right model

7.2.1 Hold-out / Simple Cross Validation

1. Randomly split the Sample S into S_{train} and S_{cv} called the **hold-out cross validation set**. A standard split is 70% S_{train} and 30% S_{cv}.
2. For each degree of the polynomial k, train the model M_k on S_{train} to get the parameter vector $\boldsymbol{\theta}_k$
3. Select $\boldsymbol{\theta}_k$ with the smallest error $\hat{\varepsilon}_{S_{cv}}(\boldsymbol{\theta}_k)$

   $$k = \underset{k\in 1,2,...,10}{\operatorname{argmin}}\ \hat{\varepsilon}_{S_{cv}}(\boldsymbol{\theta}_k)$$

   where $\hat{\varepsilon}_{S_{cv}}(\boldsymbol{\theta}_k)$ is the empirical error, if we use $\boldsymbol{\theta}_k$ as the parameter vector on the cross validation set S_{cv}.

The issue with **hold-out cross validation** is, that it wastes the amount of data used for the cross validation.

7.2.2 k-fold Cross Validation

To reduce the waste of data, we could hold out less data used per run: we use k-fold cross validation **k-fold cross validation**.

1. Randomly split S into l disjoint subsets $S_1, S_2, ..., S_l$ of $\frac{n}{l}$ training examples each
2. Evaluate each model $M_k, k \in 1, 2, ..., 10$ as follows
   - For $j = 1, 2, ..., l$ train M_k on all data, except the subset S_j to get $\boldsymbol{\theta}_{kj}$ and test $\boldsymbol{\theta}_{kj}$ on S_j to get the empirical error $\hat{\varepsilon}_{S_j}(\boldsymbol{\theta}_{kj})$
   - The estimated generalisation of the model M_k is then calculated as the average of the $\hat{\varepsilon}_{S_j}(\boldsymbol{\theta}_{kj})$

   $$\frac{1}{l}\sum_{j=1}^{l}\hat{\varepsilon}_{S_j}(\boldsymbol{\theta}_{kj}) \text{ where } \hat{\varepsilon}_{S_j}(\boldsymbol{\theta}_{kj}) = \frac{1}{2n}\sum_{i=1}^{n}(h(\boldsymbol{\theta}_k, \boldsymbol{x}^{(i)}) - y^{(i)})^2$$

3. Pick the model $M - K$ with the lowest estimated generalisation error and retrain that model on the entirety of the training set S.

8 Recommender Systems

8.1 Definition

> "A recommender system is any system that produces indiviudalized recommendations as output or has the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options" (Burke, 2002)

Or easier said:

> "A recommender system is a software application capable of suggesting interesing things to its users after learning their preferences over time" (Jennach et. al., 2010)

So basically recommender systems are the algorithms that show up under your amazon searches as 'customers who bought [xxx] also bought [yyy]'.

As opposed to Support or Lift, recommender systems are **indiviudalized**.

The goal of recommender systems is to give customers a better overview, to help them cope with the information overload problem (Amazon currently has over 20'000 results listed of 'Laptop Stand', but I only need one).

As a nice side effect, companies make more sales, because their customers can actually find what they're looking for.

8.2 History

The history of recommender systems can be split into three waves:

1st wave (1991 - 2000)
5-star ratings

2nd wave (2001 - 2010)
Integration of personal web pages/facebook/myspace etc.

3rd wave (2011 -)
Smart technologies, big data, voice activated assistants

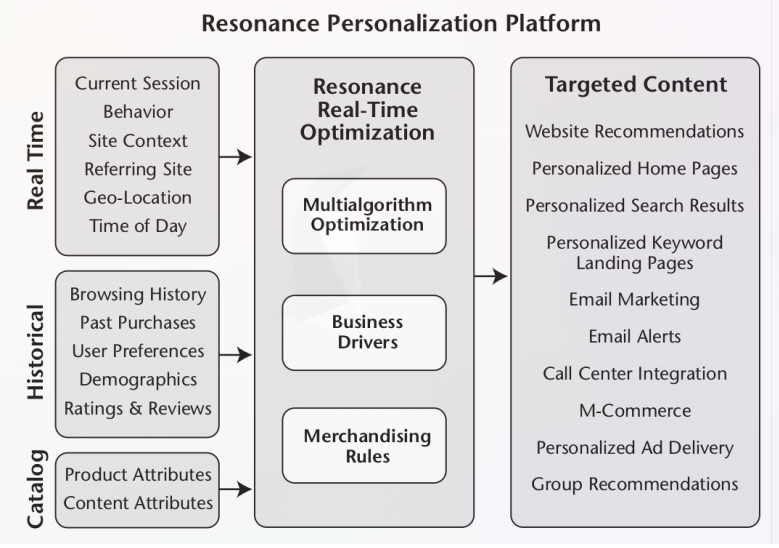


Figure 8.1: Generic Picture of a Commercial Recommender System

8.3 Business Model

- Recommender Systems are only proviced as SaaS (Software as a Service), you only rent the system
- All data is managed by the vendors. The clients only get the recommendations produced by this data and can display them at their will
- Subscription prices for these systems are related to their success, usually clients pay a percentage of their revenue
- Due to the recent privacy-boom, some customers and clients do not want to provide all shopping data to these system-vendors. They'd rather have an in-house system.
- Having loads of data is hardly bad, the *real* success in these systems lies in the configuration of the system and the evaluation of the data. This requires a deep understanding of the recommender algorithm

8.4 Recommendations by Associations

Most people think alike. People who buy cereal mostly also buy milk.

To calculate the percentage of buyers who bought Y after buying X:

$$\frac{X \text{ and } Y}{X} \tag{37}$$

This method is simple enough, but does not really work. A good example of why this won't work is Anchovies Paste and Banana. Basically everybody who buys Anchovies Paste also buys Banana. Does that mean they should be sold together as a package? No. Because these two products do not have anything in common. You could pair up basically every 'exotic' product with Banana according to this calculation. This is the case because *Everybody needs bananas*. The fact that they also bought Anchovies Paste does not have anything to do with the fact that they needed bananas

A better way to calculate basically the exact same thing, without having the Anchovies/Banana problem would be to check the percentage of buyers who bought the Anchovies Paste and also bought bananas and compare that number to the percentage of buyers who has not bought Anchovies Pate but bought bananas. That way, the popularity of bananas is of no consequences.

$$\frac{\frac{X \text{ and } Y}{X}}{\frac{X \text{ and } Y}{X}} \tag{38}$$

Another important thing to check before making a recommendation: What does the buyer already have in their cart? Let's take the example of a fast food restaurant: I am ordering ice cream and get the recommendation that I might also like ketchup. While I *do* like ketchup (although Mayonnaise, especially garlic one is still superior), I do not necessarily want it as a side to my ice cream.

Another example: If I buy ski boots, you could assume that I already own a pair of skis, so why would you recommend me to buy yet another pair of skis? Most people who bought skis and ski boots probably bought them at the same time.

To summarize:

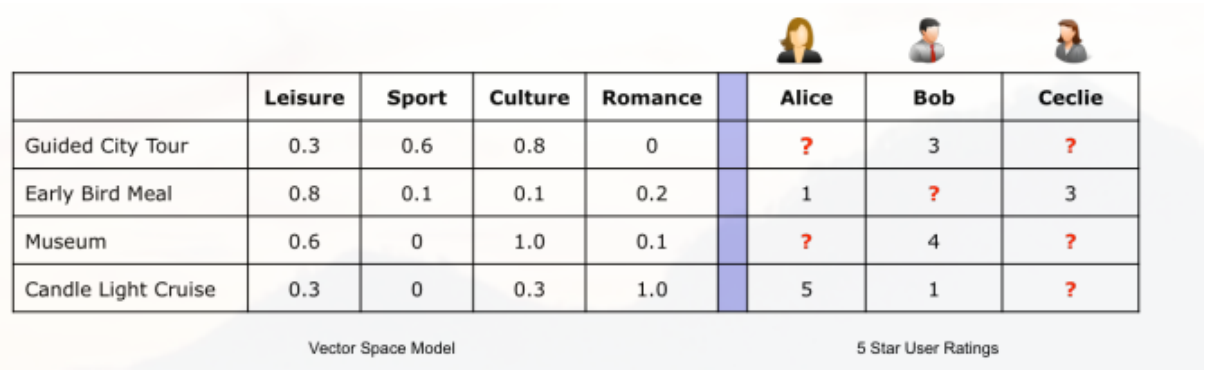
- Recommendations by Associations are *fast*. No data collection and evaluation necessa4ry

- Associations are *context aware*. Only pairs of products that match together will be recommended together
- Associations can be specifically *tailored* to certain businesses
- Associations look at *shopping carts* instead of buyers and do not consider the customers personal preferences

8.5 Context-Based Recommendations

Instead of checking what customers actually bought, why not check and leverage what customers *want* to buy?

Thanks to user ratings, wish lists, watchlists, viewed items etc, online vendors can create a very detailed shopping profile of their customers.

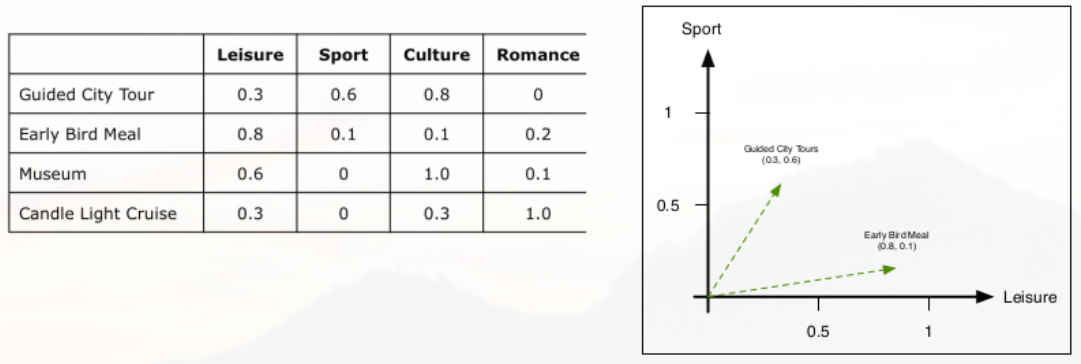


	Leisure	Sport	Culture	Romance		Alice	Bob	Ceclie
Guided City Tour	0.3	0.6	0.8	0		?	3	?
Early Bird Meal	0.8	0.1	0.1	0.2		1	?	3
Museum	0.6	0	1.0	0.1		?	4	?
Candle Light Cruise	0.3	0	0.3	1.0		5	1	?

Vector Space Model

5 Star User Ratings

Figure 8.2: Context Based Recommendations



	Leisure	Sport	Culture	Romance
Guided City Tour	0.3	0.6	0.8	0
Early Bird Meal	0.8	0.1	0.1	0.2
Museum	0.6	0	1.0	0.1
Candle Light Cruise	0.3	0	0.3	1.0

Figure 8.3: Item Similarity

The idea of Context-Based Recommendations is to recommend the user similar products they have not rated/bought yet.

1. Manually or automatically attribute your catalog data
2. Collect as much user preferences as you can (e.g. by ratings, wish lists, item viewed etc.)
3. Compute predictions for products labelled with ? based on item similarity (See Equation 8 on page 12)
4. Recommend products with high prediction value

Another approach of Context-Based Recommendations is to analyze similar products:

Neighbors N_j refer to the (at most) N rated items most similar to item j, N being a hyperparameter that can be tweaked for improved algorithm-performance Select only rated items with similarity > 0 (choose less than N if you cannot find enough candidates)

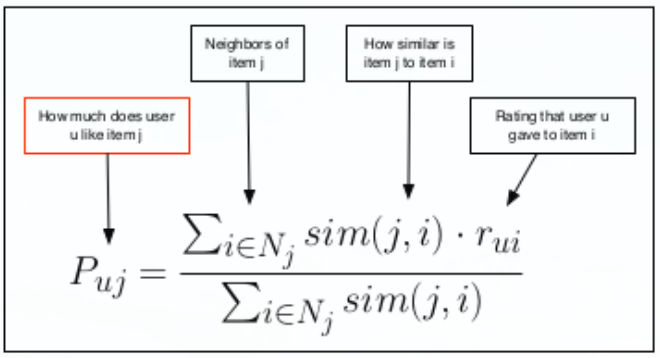


Figure 8.4: Context Based Recommendations with Neighbours

Figure 8.5: Similarity between books

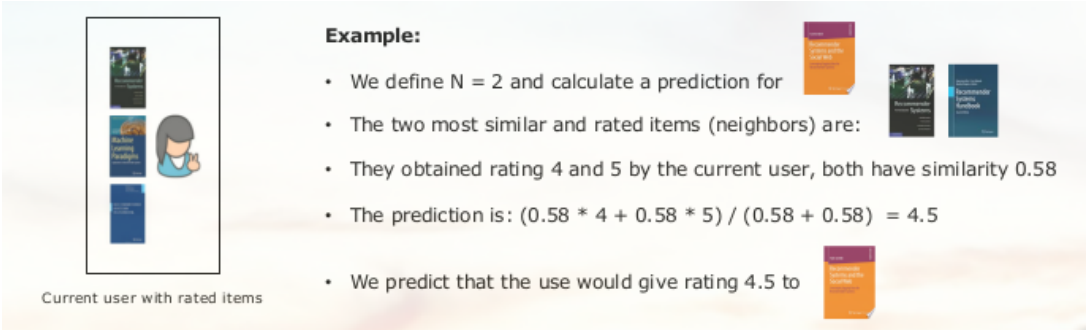


Figure 8.6: Example of N-Based Context-Based Recommendations

Advantages of Contex-Based Recommendations

- A users profile is built only from their own ratings and recommendations can therefore be better explained
- New items can be recommended without having to be rated first

Disadvantages

- The user will never find something unexpected they might like in their recommendations
- Do product attributes really reflect how users decide?
- What do you recommend new users?

8.6 User-to-User Collaborative Filtering

This approach is similar to the similar products context based recommendations, but instead of similar products, it compares the customer to *other similar customers*.

A set of users who liked the same items in the past, probably share the same preferences. Therefore, the user gets recommendations of products that many 'Neighbours' (similar users) liked, but the user has not purchased (yet).

There's a similar approach as to the Context-Based Recommendations, but there's no need for product attributes anymore. The formula to calculate how much user u will wile product j is also fairly similar.

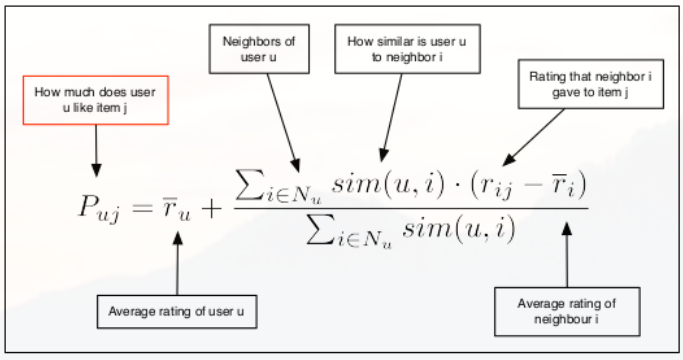


Figure 8.7: Context Based Recommendations with Neighbours

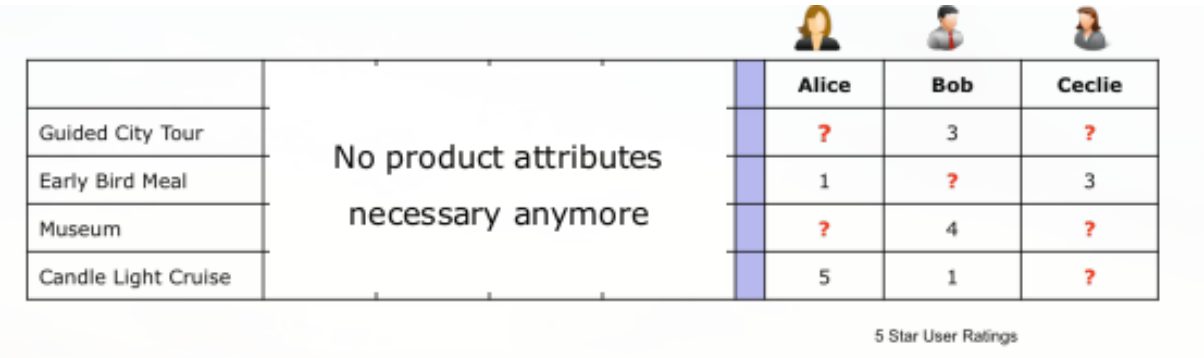


Figure 8.8: User to User Collaborative Filtering

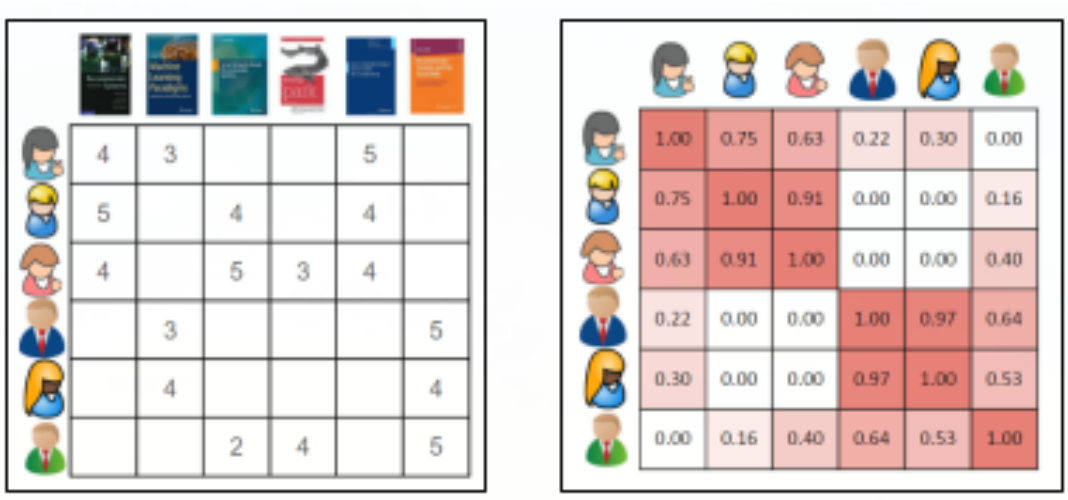


Figure 8.9: Similarity between users

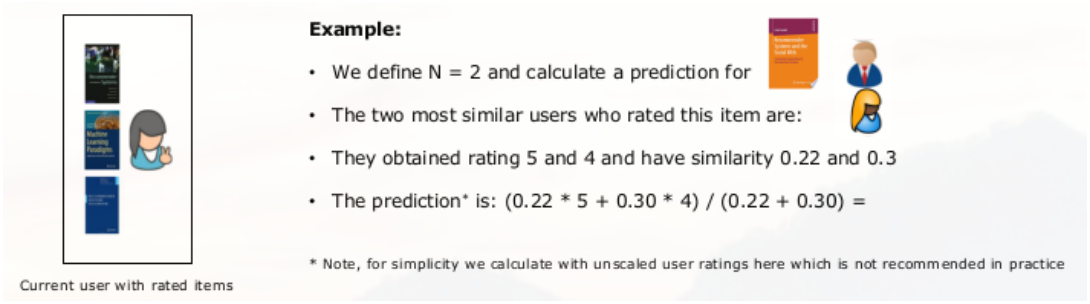


Figure 8.10: Example of N-Based User to User Filtering Recommendations

8.7 Item-to-Item Collaborative Filtering

Beause user proviles constantly change, pre-computation does not necessarily work. In contrast, for pairs fo items, similarity is stable and can be pre-computed.

The main difference is that we compare row vectors for neighbour search instead of column vectors.

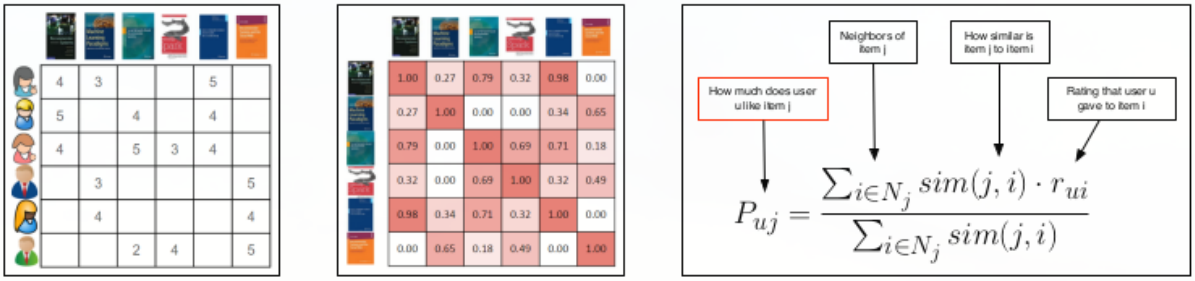


Figure 8.11: Item to Item Collaborative Filtering

Advantages of Collaborative Fiiltering Recommendations

- Products from different categories can be recommended

Disadvantages

- If the matrix is too sparsely filled, no neighbours will be found
- It needs loads of people to participate to make decent recommendations
- The system must first learn new users preferences
- Users with non-mainstream taste will not get decent recommendations

To improve the recommendations for sparsely filled matrices, there's something called **Low Rank Matrix Factorization**. The matrix we saw before (Fig. 8.11, 8.9, 8.5) can be split up in two matrices, which are more densely filled.

One matrix holds all users and their ratings for the products they rated and the second matrix holds all products.

The first matrix (U) holds all users and their ratings. Each row is a user and all the ratings on all products they have made.

The second matrix (V) holds all the products and the rating they've gotten. This matrix is **transposed**. o

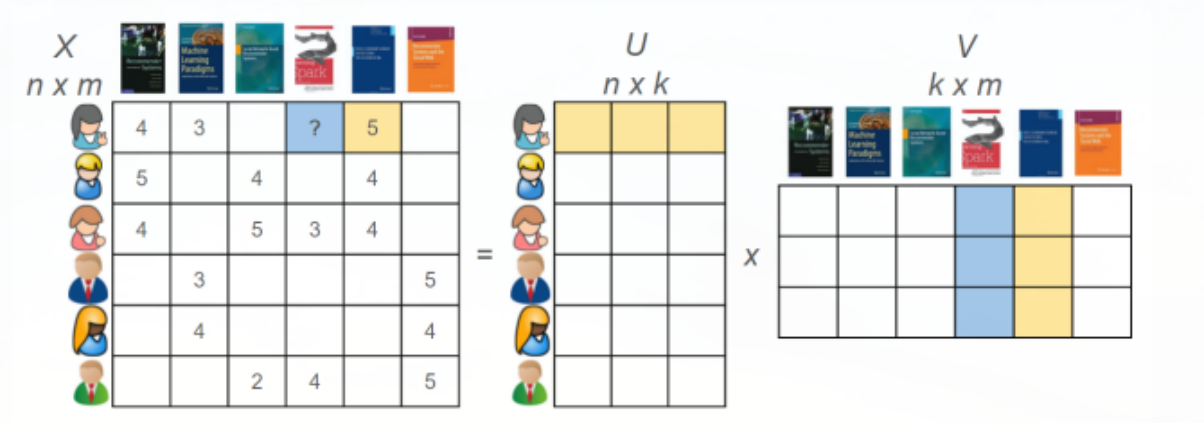


Figure 8.12: Low Rank Matrix Factorization

From this, we can see that the original matrix can be written as $U \times V^T$

$$U = \begin{bmatrix} U_{1,1} & U_{1,2} & \dots & U_{1,k} \\ U_{2,1} & U_{2,2} & \dots & U_{2,k} \\ U_{3,1} & U_{1,2} & \dots & U_{3,k} \\ U_{4,1} & U_{4,2} & \dots & U_{4,k} \\ \dots & & & \\ U_{n,1} & U_{n,2} & \dots & U_{n,k} \end{bmatrix} \qquad V = \begin{bmatrix} V_{1,1} & V_{1,2} & V_{1,3} & V_{1,4} \dots & V_{1,m} \\ V_{2,1} & V_{2,2} & V_{2,3} & V_{2,4} \dots & V_{2,m} \\ \vdots & & & & \\ V_{k,1} & V_{k,2} & V_{k,3} & V_{k,4} \dots & V_{k,m} \end{bmatrix}$$

$U_{1,1}$ is the rating User 1 gave the book 1, $U_{2,1}$ is the rating User 2 gave the book 1 etc.

v is the product matrix. The V_1 column stores all ratings book 1 has gotten, V_2 all ratings of book 2 etc.

So if we want to check what rating user 3 has given book 4, we just need to calculate $U \times V^T$ and get the entry $(3, 5)$.

That way, we can also 'guess' what rating a user would give a book they have not rated or bought yet. The rating of book 2 by user 2 is soterd in $U \times V, (2, 2)$.

8.8 Hybrid Recommender Systems

The best result can usually be achieved if you mash multiple recommender-algorithms together. That way, their different advantages and disadvantages might cancel each other out. If you mix

a content-based and a collaborative algorithm, you could for example eliminate the cold start problem.

But how could you interpret the different results from the different algorithms?

Weighted: Score an item as a weighted sum of scores from different algorithms

Mixed: Results of different algorithms presented to the user

Cascade: One algorithm refines the result of another one

Switching: Different methods/algorithms can be used for different usecases

8.9 Evaluation

How would you evaluate recommender systems? Typical train/test/validate splits of your data will not work.

Instead the data must be split over several users, because to assess how good the recommendations are, we need the users whole shopping history.

An often used performance measurement is $P@k$ (Precision @ k). How many of the k recommended items were actually relevant. In case many users have less than k recommended products, you could also go with $minP@k$.