Machine Learning Hands-on

KhanhPhamDinh, VinAI Applied, Ha Noi

my book: Machine Learning Algorithms to Practices
https://phamdinhkhanh.github.io/deepai-book/intro.html

Feature Engineering

References:

https://phamdinhkhanh.github.io/deepai-book/ch_ml/index_FeatureEngineering.html

http://kti.tugraz.at/staff/denis/courses/kddm1/featureengineering.pdf

https://people.eecs.berkeley.edu/~jordan/courses/294-fall09/lectures/feature/slides.pdf

https://www.cs.princeton.edu/courses/archive/spring10/cos424/slides/18-feat.pdf

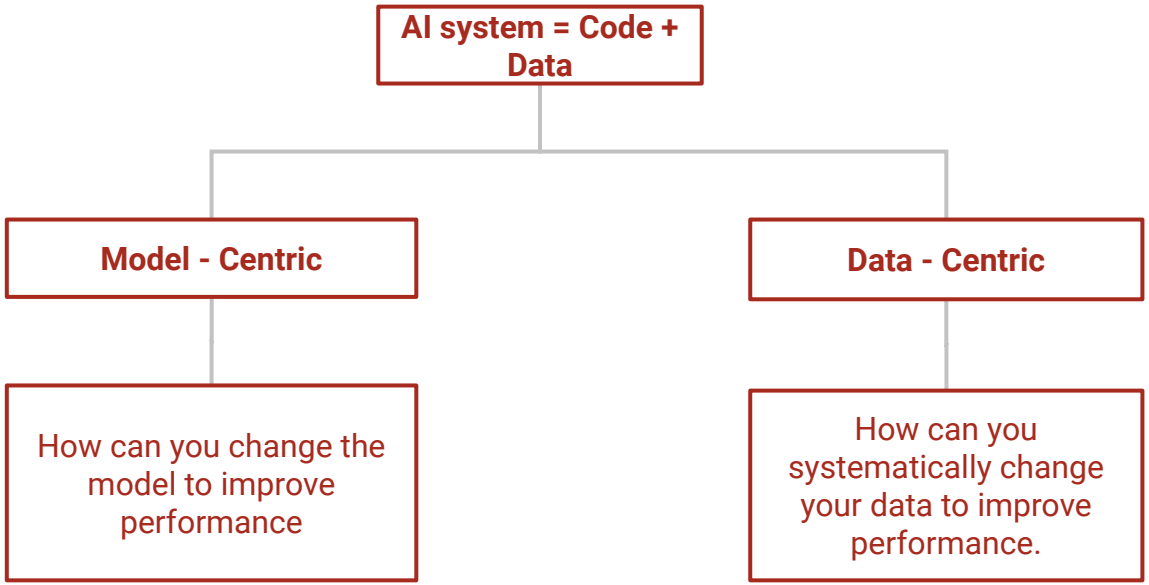
https://en.wikipedia.org/wiki/Feature_engineering

https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/

Content

- AI system improvement
- Importance of Data
- Importance of Feature Engineering
- Feature Extraction
- Feature Selection
- Feature Transformation

AI system improvement



Importance of Data

	Steel defect detection	Solar panel	Surface inspection
Baseline	76.2%	75.68%	85.05%
Model-centric	+0% (76.2%)	+0.04% (75.72%)	+0.00% (85.05%)
Data-centric	+16.9% (93.1%)	+3.06% (78.74%)	+0.4% (85.45%)

https://towardsdatascience.com/from-model-centric-to-data-centric-artificial-intelligence-77e423f3f593

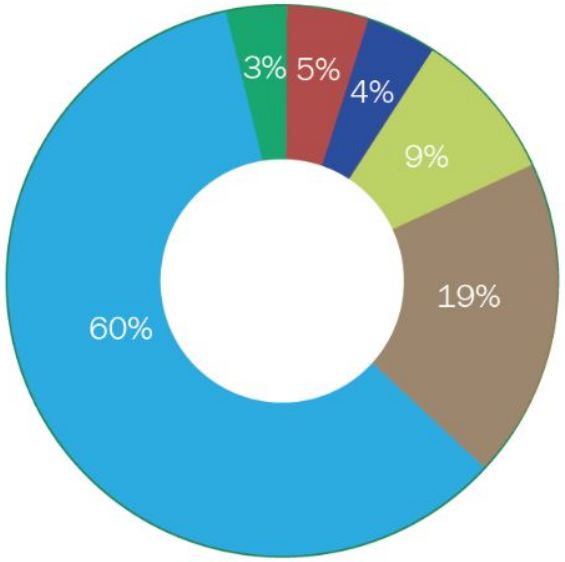
Importance of Feature Engineering

Time allocation:

- A survey conducted by data scientists revealed that over 80% of their time was spent capturing, cleaning, and organizing data.
- Less than 20% was spent creating these machine learning pipelines that end up dominating the conversation.

Source: https://www.packtpub.com/product/feature-engineering-made-easy/9781787287600

Following is the graph of the what Data Scientist spend the most time doing:



As seen from the preceding graph, we breakup the Data Scientists's task in the following percentage :

- **Building training sets:** 3%
- **Cleaning and organizing data:** 60%
- **Collecting data for sets:** 19%
- **Mining data for patterns:** 9%
- **Refining algorithms:** 5%

Importance of Feature Engineering

Expert's statement:

- "Feature engineering is the art part of data science."

  Sergey Yurgenson, former #1 ranked global competitive data scientist on Kaggle

- "Coming up with features is difficult, time-consuming, requires expert knowledge. "Applied machine learning" is basically feature engineering."

  Andrew Ng, chief scientist of Baidu, co-chairman and co-founder of Coursera, and adjunct professor at Stanford University

  Source: *https://en.wikipedia.org/wiki/Feature_engineering*

Role of feature engineering:

- Determine which features are the most important
- Invent the new features in real-world problem domains.
- Core-method improves model accuracy.
- Vital to machine learning success.

Feature Engineering Technique

DS
boosting your data exploration

Feature Engineering

Feature Extraction

- Derive features to be informative and non-redundant.
- Facilitate subsequent learning and generalization steps.
- Lead to better interpretation

Feature Transformation

- Apply a function on data points.
- Closely meet assumption of statistical inference procedure.
- Improve interpretability or appearance of graphs.

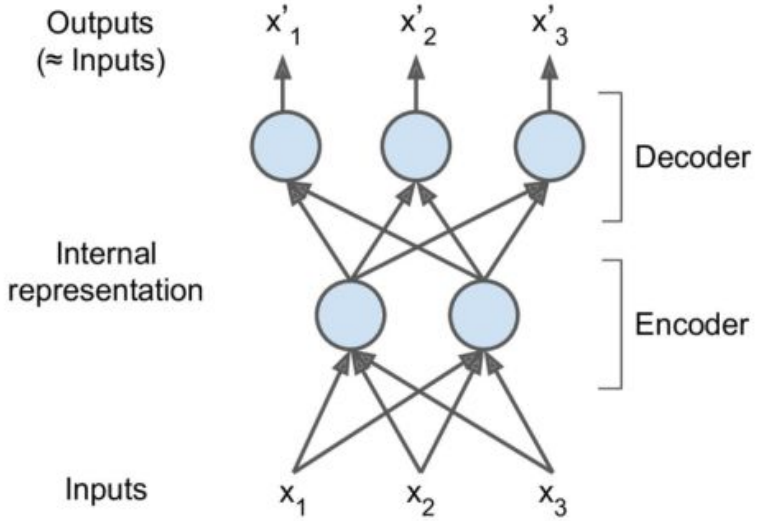
Feature Selection

- Ranking importance of features
- Filter out a set of the most important features.
- Save computational cost

Feature Extraction

Auto-Encoder

- Encoder: Transform from high-dimensional data into low-dimensional data.
- Decoder: Transform from low-dimensional data into high dimensional data.
- Inputs and Outputs is nearly similar.



practice: *Auto-Encoder phamdinhkhanh*

Feature Extraction

Bag-of-word

- Define a bag-of-word.
- Encoding sentence into a frequency-vector.

I	2
AI	1
a	1
about	1
book	1
deep	0
greate	1
is	0
this	0
machine	0
learning	0
have	2
to	1
read	1
twice	1
times	1

I have a greate AI book,
I have to read twice times

practice: https://phamdinhkhanh.github.io/deepai-book/ch_ml/FeatureEngineering.html#phuong-phap-bag-of-words

Feature Extraction

Bag-of-ngrams

- Define many sequential words into a gram.
- Two sequential words is bigram, three sequential words is trigram
- Example: ‘The fox jumps over the grass’ → is tokenized by bigram: [‘the fox’, ‘fox jumps’, ‘jumps over’, ‘over the’, ‘the grass’].
- Increase size of dictionary than bag-of-word.
- Embedding sequence is usually a sparse vector.

practice: https://phamdinhkhanh.github.io/deepai-book/ch_ml/FeatureEngineering.html#phuong-phap-bag-of-words

Feature Extraction

TF-IDF

- Term frequency - inverse document frequency
- Calculated by term frequency x inverse document frequency.

$$\mathrm{idf}(t, D) = \log \frac{|D|}{|\{d \in D; t \in d\}| + 1} = \log \frac{|D|}{\mathrm{df}(d, t) + 1}$$
$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D)$$

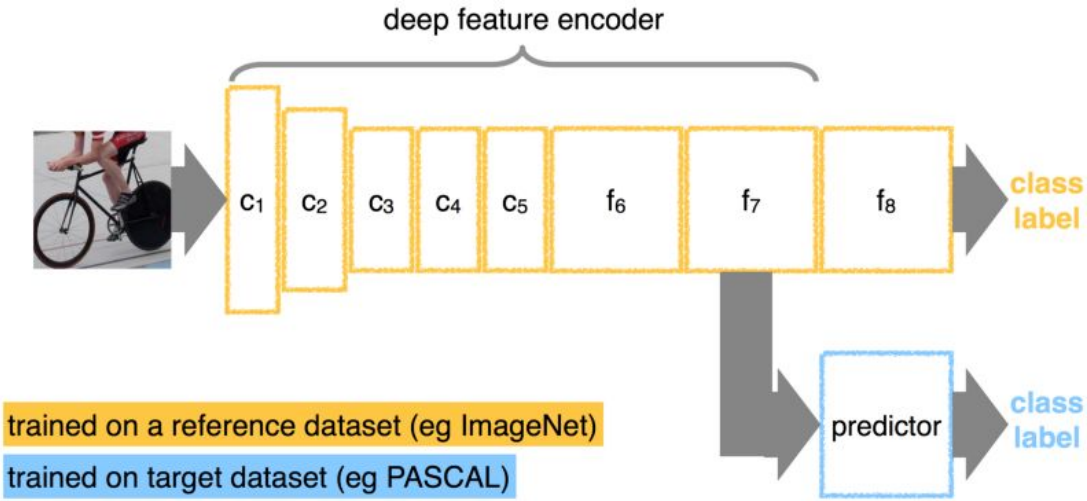
- Question: What is the meaning of high tf-idf and low tf-idf?

practice: https://phamdinhkhanh.github.io/deepai-book/ch_ml/FeatureEngineering.html#phuong-phap-tf-idf

Feature Extraction

Image processing

- CNN backbone as a deep feature extractor: ResNet, ResNeXt, MobileNet, VGG, AlexNet, EfficientNet
- Can be transfer learning between dataset similar domains



practice: https://phamdinhkhanh.github.io/2020/04/15/TransferLearning.html

Feature Extraction

Datetime

- From datetime → extract: day_in_week, date_in_moth, quater_of_year,...
- using package datetime, calendar.

```
from datetime import datetime
import pandas as pd

dataset = pd.DataFrame({'created': ['2021-08-13 00:00:00', '2021-08-12 00:00:00', '2021-08-11 00:00:
                                    '2021-08-10 00:00:00', '2021-08-09 00:00:00', '2021-08-08 00:00:

def parser(x):
    # Để biết được định dạng strftime của một chuỗi kí tự ta phải tra trong bàng string format
    return datetime.strptime(x, '%Y-%m-%d %H:%M:%S')

dataset['created'] = dataset['created'].map(lambda x: parser(x))
print(dataset['created'].dtypes)
```

```
datetime64[ns]
```

practice: https://www.dataquest.io/blog/python-datetime-tutorial/

Feature Extraction

Web log data

- From web log data → extract information about user: os, device, browser,...
- package: user_agents.

```
import user_agents
# Giả định có một user agent như bên dưới
ua = 'Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Ubuntu Chromium/56.0.29
# Parser thông tin user agent
ua = user_agents.parse(ua)
# Khai thác các thuộc tính của user
print('Is a bot? ', ua.is_bot)
print('Is mobile? ', ua.is_mobile)
print('Is PC? ',ua.is_pc)
print('OS Family: ',ua.os.family)
print('OS Version: ',ua.os.version)
print('Browser Family: ',ua.browser.family)
print('Browser Version: ',ua.browser.version)
```

```
Is a bot?  False
Is mobile?  False
Is PC?  True
OS Family:  Ubuntu
OS Version:  ()
Browser Family:  Chromium
Browser Version:  (56, 0, 2924)
```

Feature Extraction

Extract Text from Image

- Information in text is also useful.
- package: general_ocr , tessaract.

Feature Transformation

DS boosting your data exploration

Feature Transformation

Standardization

- Using Gaussian distribution.
- X ~ N(0, 1).

Min-max Scaling

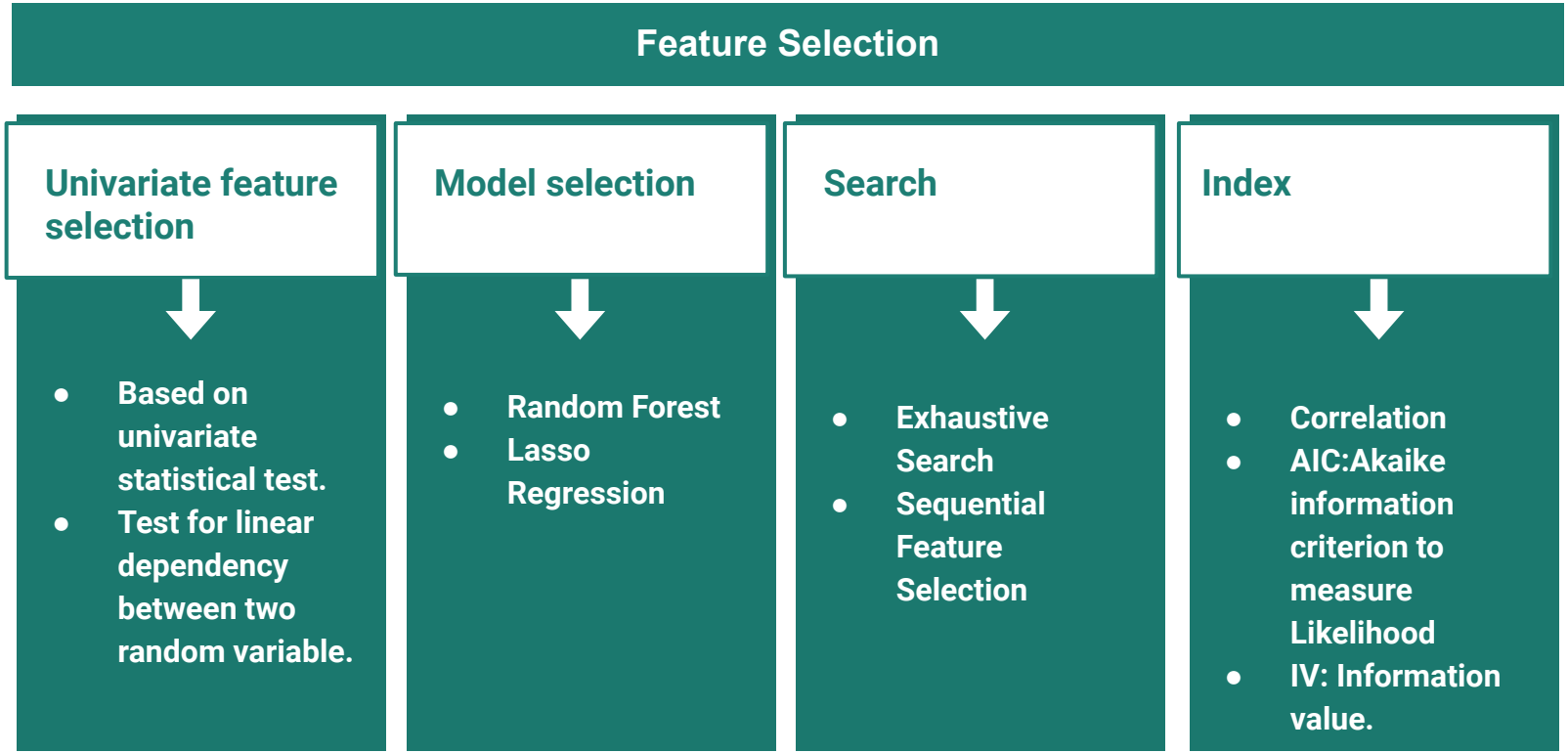
- X = (x-x_min)/(x_max - x_min)
- X in [0, 1]

Robust Scaling

- X = (X-Q2(X))/(Q(3)-Q(1))
- Better for data when it exists outliers.

https://phamdinhkhanh.github.io/deepai-book/ch_ml/FeatureEngineering.html#id2

Feature Selection



https://phamdinhkhanh.github.io/deepai-book/ch_ml/FeatureEngineering.html#id3