AI Engineer Training: III

In the Era of Deep Learning

IT21 Learning
Alvin Jin

AI News

- Impact AI summit aims to establish Ottawa as another AI hub in Canada
-
- Big Data & AI Toronto Expo, 12-13 Jun, MTCC
- SenseTime become the most funded AI startup valued over $3B.
- GDPR was effect in May 25, how will it impact AI industry?

Agenda

- Machine Learning Practices

- Case Studies:
  - Sentiment Analysis II: word embeddings
  - Regression Algorithm: house price prediction

Features

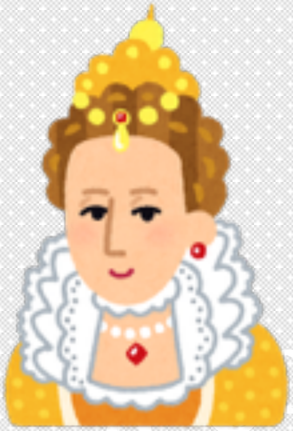
- Base Features
  - Directly observed from raw data
- New Features
  - Extract from raw data by simple transform and calculation.
- Latent/hidden features
  - Not obvious, no idea exactly what attributes these factors are describing

Feature Engineering

- The process of using **domain knowledge** of the data by expressing it in a simpler way.
- It is hard, time-consuming, arts rather than science.
- Good feature engineering
  - Solve problems elegantly using fewer resources.
  - Solve problems with far less data.

Quiz

- **King – Man + Woman ≈**



Word	*Power*	*Gender*	*Wealth*
King	1.0	1.0	1.0
Man	0.2	1.0	0.2
Woman	0.1	0.0	0.1
Queen	0.9	0.0	0.9

Word Embeddings

- Dense representations of word sequences in a low-dimensional vector space.
- Vector elements describe an as yet unknown feature
- Items with similar distributions have similar meanings

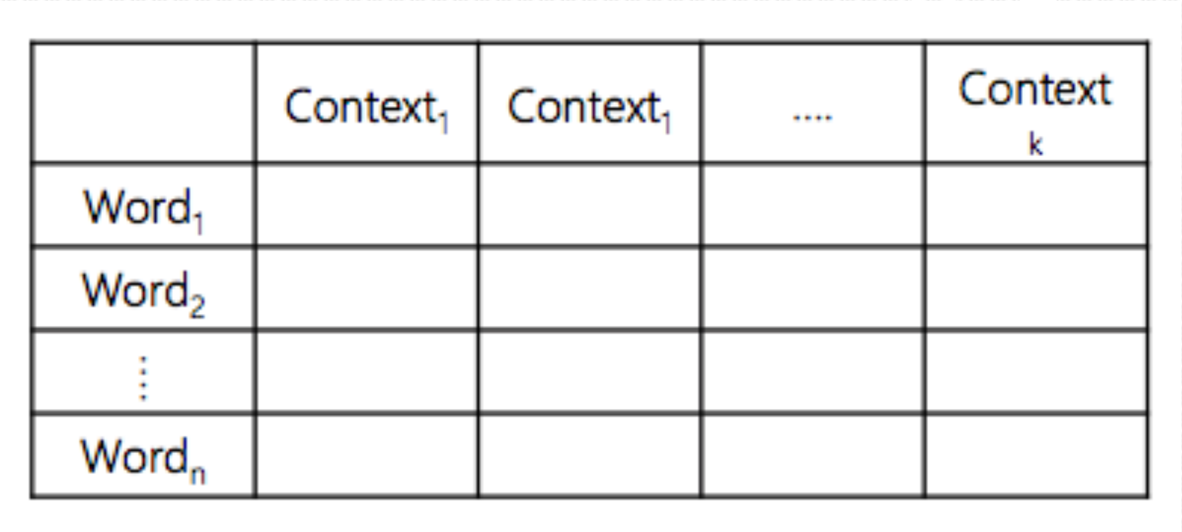
One-hot-encode vs. Embeddings

- One-hot-encode
  - binary, sparse, very high-dimensional vectors
  - built in data preprocess phase

- Word Embeddings
  - low-dimensional floating-point vectors
  - learned from data set via neural network

Embedding Matrix

- Initialize all word vectors randomly to form a matrix
- Weights are learned via gradient descent in neural networks



	Context_1	Context_1	....	Context_k
Word_1				
Word_2				
⋮				
Word_n				

Case Study: Regression Problem

- Predict a continuous value instead of a discrete label values for forecasting, time series modelling

- Understanding the causal effect relationship between variables.

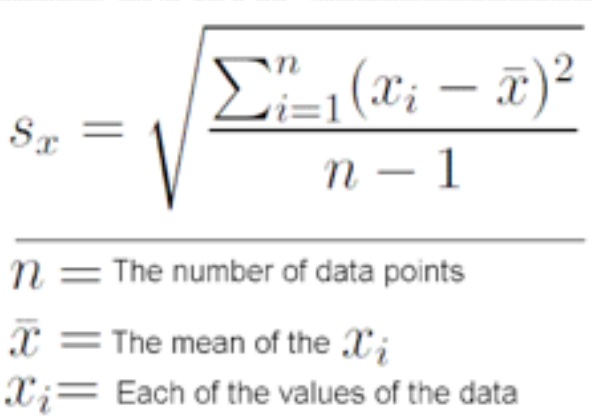
- However, Logistic Regression is a binary classification algorithm to regress for the probability of a binary categories.

Normalization

- Take small values
  - Most values are in 0-1 range.
- Homogenous
  - All features should take values in roughly the same range
  - Each feature has a standard deviation of 1 and a mean of 0

Standard Deviation

- A measure to quantify the amount of variation of a set of values.
- When low: the data points tend to be close to the mean
- When high: the data points are spread out over a wider range of values.



$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

n = The number of data points

$\bar{x}$ = The mean of the x_i

x_i = Each of the values of the data

Regression Functions & Metrics

- Mean Squared Error
  - more sensitive to extreme values

$$MSE_S(h) = \frac{1}{n}\sum_{x \in S}(f(x) - h(x))^2$$

- Mean Absolute Error
  - the same magnitude of the actual values

$$MAE_S(h) = \frac{1}{n}\sum_{x \in S}\left|f(x) - h(x)\right|$$

Adaptive Moment Estimation

- Adam(2015) combines the benefits of RMSProp and Momentum, normally is your first choice.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t .$$

- Adam multiplies the learning rate by the momentum, also divides by a factor related to the variance.

Q & A