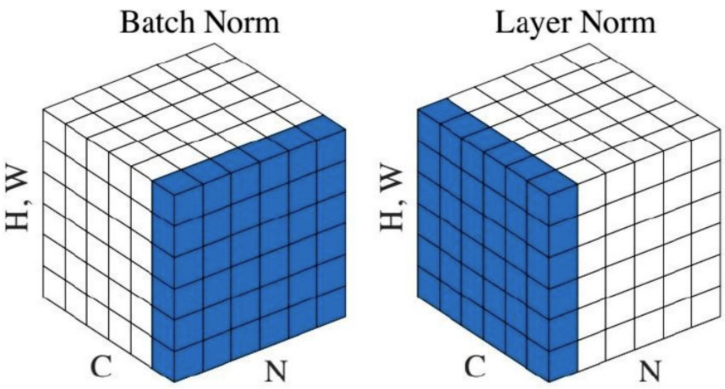
Normalization

- Motivation
  - a data pre-processing procedure that brings the numerical data to a common scale without distorting its shape, comparing features with every datapoint **having the same scale**, so each feature is equally important.
  - iid: independent and identically distributed is the wholly assumption.
  - **Internal Covariate Shift (ICS)**
    - A phenomenon that parameter initialization and changes in the distribution of the inputs of each layer affect the learning rate of the network.
    - the learning speed will decrease, because the upper layers need to adapt to new input data distribution updates
    - the changes of lower layers inputs will become too large or too small, leading to stop learning too early.
  - Solution
    - Whitening白化
      - transform data to have a covariance matrix that is the identity matrix : 1 in the diagonal, 0 for the other cells.
      - To remove correlation or dependencies between features in a dataset. To fix the input distribution of each layer in network.
      - After whitening, we gain the input distribution with the same mean and variance. Application: PCA and ZCA
        - PCA: $\mu = 0, \sigma = 1$
        - ZCA: $\mu = 0, \sigma same$
      - Steps:
        - Zero-center
        - Decorrelate
        - Rescale
      - Downside: too costy and the data expression ability is limited
    - Normalization
      - Simplified in calculation and keep the data expression ability, compared with whitening
      - Procedure
        - input: $x = (x_1, x_2, ..., x_d)$
        - output: $y = f(x)$
        - Transformation Structure: $h = f(g * \frac{x-\mu}{\sigma} + b)$
          - re-shift paramater: b

        - re-scale parameter: g
        - $y = g * x' + b \quad \sim dis(b, g^2)$
- **Batch Normalization**纵向规范化
    - function: $\mu_i = \frac{1}{M} * \sum x_i \ ; \ \ \sigma_i = \sqrt[2]{\frac{1}{M} * \sum (x_i - \mu_i)^2 + \epsilon}$
        - M: the size of mini-batch
    - Intuition:use mini-batch data to calculate this *single neuron* mean and variance
    - Downside:
        - if each original mini-batch distribution are different with each other, the different transformation will apply to different mini-batch data, the complexity of training is added.
        - do not fit for RNN and dynamic neural network structure (with time-series data)
        - largely depend on big size of mini-batch, if the size is small, the effect is not so good.
    - Application: this method is fitted for mini-batch with big size and the approximate same distribution. Applied more in Computer Vision area.
- **Layer Normalization**横向规范化
    - Intuition: wholly consider the inputs in one layer, calculate the average and variance of this layer
    - Function: $\mu = \sum_i x_i; \ \ \sigma = \sqrt[2]{\sum_i (x_i - \mu)^2 + \epsilon}$
        - i: list all the neurons in this layer
    - Also applied to small size of mini-batch, dynamic NN, and RNN, saving the memory without saving the mean and variance of mini-batch.
    - Have a more limited model expression ability than BN, if the inputs are different features(eg. color and size). Because LN ensures all the input stay in the same range.
    - Applied more in NLP area: Transformer, BERT因为Transformer堆叠了很多层很容易梯度消失
        - Different sentences have different lengths.
        - Small correlation between different sample batch
        - Decrease the loss of differential infromation among sample batches.
- BN对不同样本同一特征的缩放for batches；LN对单个样本所有不同的特征的缩放for hidden state



- NLP:
  - N : $batch\ size$
  - C : $sequence\ length$
  - H, W : $dimension$

- Other Methods in Feature Engineering
  - **Min-Max Normalization**
    - Function: $\frac{value-min}{max-min}$
    - Downside: do not handle outliers
    - change the original data into range [0,1] proportionally, depend more on maximum/minimum value
  - **Z-Score Normarlization**
    - Function: $\frac{value-\mu}{\sigma}$
    - Evaluation: avoid the outlier issue, but does not produce normalized data with the *exact* same scale.
  - **Linear Ratio Switch**
    - Function: $\frac{value}{max}$
    - change the original data into range [0,1] proportionally, depend more on maximum value
- References
  - ICS and Normalization. https://blog.csdn.net/sinat_33741547/article/details/87158830
  - PCA-whitening VS ZCA-whitening: a numpy 2d visual. https://towardsdatascience.com/pca-whitening-vs-zca-whitening-a-numpy-2d-visual-518b32033edf

以上内容整理于 幕布文档