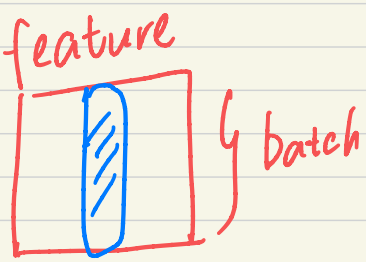
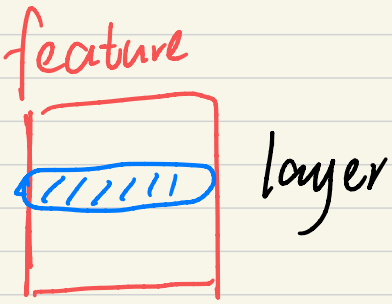
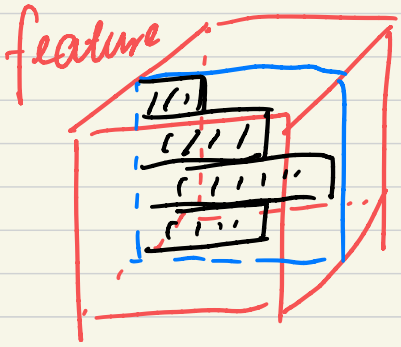
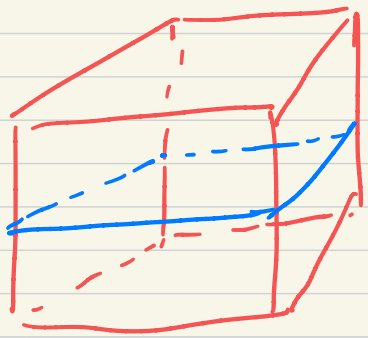
Batch normalization　Layer normalization





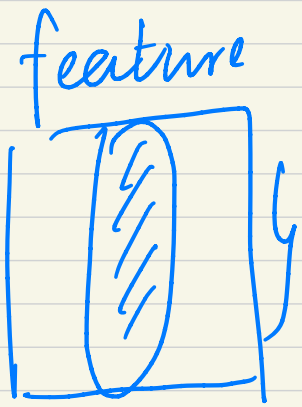
均值为0方差为1





seq(n)

seq长度不等



归一化的作用：将进入激活函数前

避免落在饱和区

$$BN(x_i)=\gamma\frac{x_i-\mu_i}{\sqrt{\sigma_i^2+\varepsilon}}+\beta$$ → 满足特定分布

↳ 分母小防止方差为0

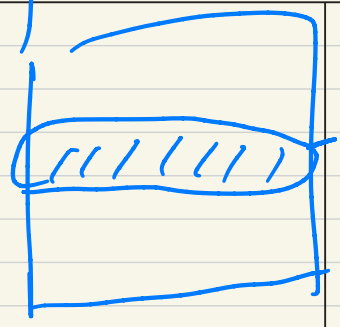
γ：控制归一化后的方差

β：控制归一化后的均值

$X\sim N(0,1)$

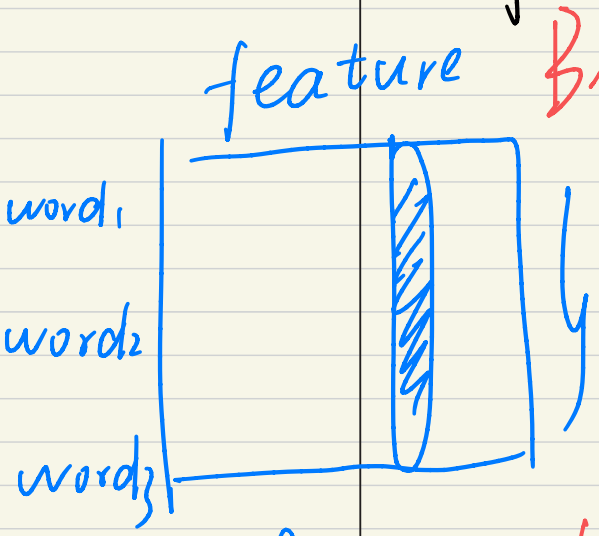
$Y=\gamma X+\beta\sim N(\beta,\alpha^2)$

$$LN(x_i) = \frac{x_i - \mu_i}{\sqrt{\sigma_i^2 + \rho}} + \beta$$ 一个所有特征

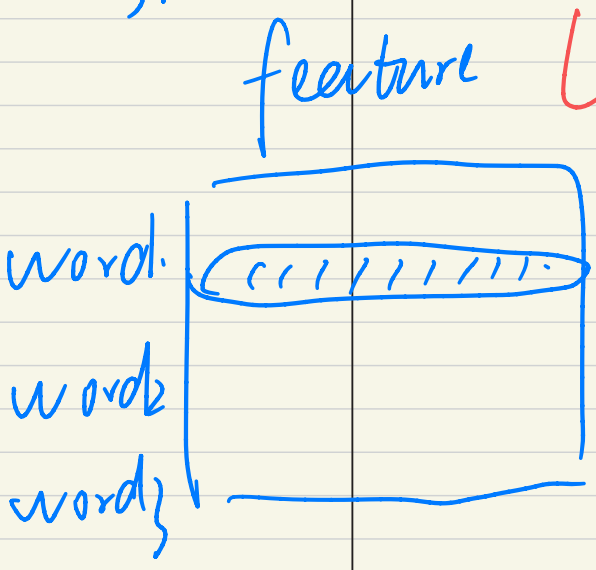


(适用变长数据)

二维: 字的情绪划分

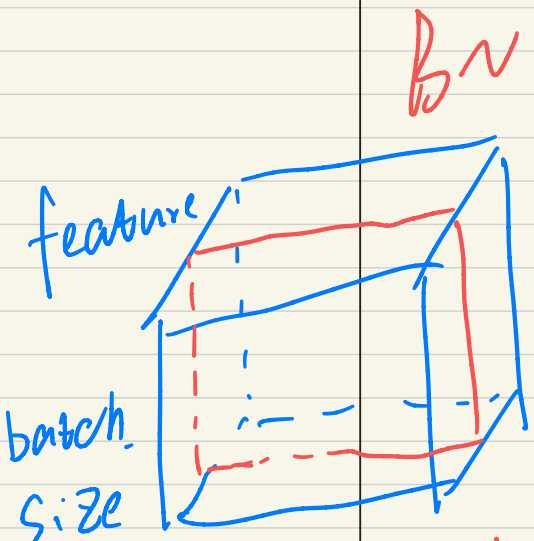


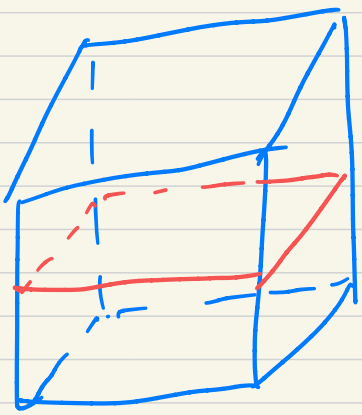
计算每个特征每个批量下的均值和方差

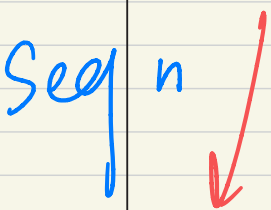


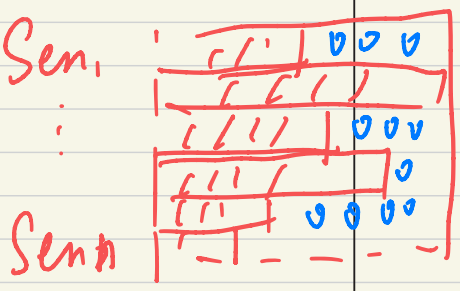
计算每个样本所有特征的均值和方差

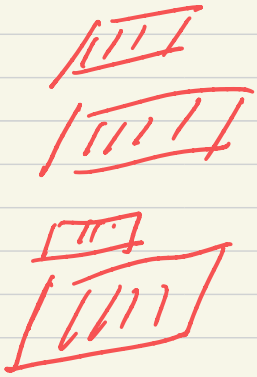
三维下：机器学习











不同句子的词在某个特征上的取值不一定有

按一批batch数据后针对同一特征的均值和方差变动很大

新句子很长的话不适用原来的均值和方差

不需要存全局的均值和方差

而且只针对该样本(更稳定)

切片后可以指定归一化的shape

可以选择是对最后一个维度归一化还是对整片数据归一化

nn.LayerNorm(embedding_dim)

/(5.10)