二分类模型精确度的一种指标, 用于测量不均衡数据的精度。它 同时兼顾了分类模型的精确率和 召回率。F1-score可以看作是模 型精确率和召回率的一种加权平 均,它的最大值是1,最小值是 F值=PrecisionxRecallx2 / 0。 (Precision+Recall) 在多分类问题中, 如果要计算模 micro-F1: 适用多分类不平衡数 型的F1-score,则有两种计算方 据,计算总的Recall和精确率,然 式,分别为micro-F1和macro-后计算出的F1值就是micro-F1, F1, 这两种计算方式在二分类> MiF 全部一起计算时叫mic 中与F1-score的计算方式一样, ro、当样本不均衡的时候用这种 所以在二分类问题中, 计算 方法较好 micro-F1=macro-F1=F1-score, macro-F1:是对每个类别的F1micro-F1和macro-F1都是多分类 score直接求平均。所以不适合 F1-score的两种计算方式; 不平衡数据,容易受到高recall、 高precision的类别影响,缩 写名字MaF • 0.5 < AUC < 1, 优于随机猜 测。这个分类器(模型)妥善设定 阈值的话,能有预测价值。 • AUC = 0.5, 跟随机猜测─样 (例:丟铜板),模型没有预测价 值。 • AUC < 0.5, 比随机猜测还差;但 Area under the ROC Curve (AUC) AUC 是ROC曲线下的面 只要总是反预测而行, 就优于随 机猜测。 AUC和ROC 例如Recall@1,计算Recall@1,说 明所选候选有多少次排名第一。 k是容忍度,前k个中预测正确答 案除以k Recall@K tp@k/(tp@k+fn@k) tp@k/(tp@k+fp@k) Precision@k 的一个限制是它不 考虑相关元素的位置。 Precision@k 该指标量化了 top-K 结果中有多 少项是相关的 排名评估指标 (0,1)之间, 越接近1越好 hit@k hit@k描述了出现在排序列表前k 个实体中的真实实体的比例 MR∈[1,∞), 越低越好 计算所有单独排名的算术平均值 是一个国际上通用的对 搜索算法 metric 进行评价的机制,即第一个结果 匹配,分数为1,第二个匹配分 数为0.5, 第n个匹配分数为1/n, 如果没有匹配的句子分数为0,最 终的分数为所有得分之和 mean rank (MR): 平均排名 优点: 对任何模型的性能变化都 很敏感, 而不仅仅是在某个截止 点his@k下发生的变化,因此它 反映了平均性能 缺点: 平均排名取决于候选者的 数量, 假如平均排名分数为10, 可能表明在100万个大的候选集 的情况下性能很强,但在20个小 的候选集的情况下, 性能却很 差。 当我们希望我们的系统返回最佳 相关项目并希望该项目位于更高 位置时, 此指标很有用。 MRR∈ (0,1]越接近1越好 mean reciprocal rank (MRR): 倒数排名的算术平均值 平均倒数排名  $MRR=\frac{1}{IQI} \sum_{i=1}$ IIQII表明总的查询次数 ^{|Q|} \frac{1} {\operatorname{ran} k\_{i}}\$ ranki表明第一个相关结果的排名 反算术平均排名(IAMR)可以被 Inverse Arithmetic Mean Rank 定义为平均排名,即MR的倒数 (IAMR) 反算术平均数排名 IAMR∈ (0,1] Harmonic Mean Rank (HMR) 调和平均排名是MRR的倒数 调和平均数排名 HMR∈[1,∞) geometric mean rank (GMR) 几 何平均排名 GMR∈[1,∞) inverse geometric mean rank GMR的倒数 (IGMR) 反几何平均排名 IGMR∈(0,1] 平均精度是评估模型选择的所有 真实相关项目是否排名较高的指 rel(k)是一个指示函数,当rank k Average Precision(AP) 平均精度  $A P=\frac{k=1}^{n}(P(k))$ 的元素相关时为1 \* \operatorname{rel}(k))}\\text { number of relevant items }}\$ P(k) 是 Precision@k BLEU分数 ROUGE 文本生成 **BLEURT** 

F1-score: 是统计学中用来衡量

NLP方向总结-Metric

BERTScore