# 语言智能测试建议标准

高思集团技术中心
AI Lab 语言智能组
2019.06.25

教育领域中语言智能的评测通常包括算法评测和业务评测。在本标准中，算法评测是指从人工智能研究的角度对语言智能算法的性能进行定量评测和定性评测。业务评测是指从教育应用的角度对语言智能系统的实际效果进行的定量评测和定性评测。

## 1 算法评测标准

在人工智能研究中，通常采用可重复实验对算法性能（泛化误差）进行评估。为此，需要使用一个测试集来测试算法对于未见样本的判别能力，然后以测试集上的测试误差作为泛化误差的近似。这种测试一般称为体外测试（ex vivo）。需要注意，测试集应该尽可能与训练集互斥。

对于训练集 $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)\}$ ，预测任务希望在训练集上建立从输入空间 $\mathbf{X}$ 到输出空间 $\mathbf{Y}$ 的映射 $f : \mathbf{X} \to \mathbf{Y}$ 。如果 $\mathbf{Y}$ 是离散空间，则称预测任务为分类问题。如果输出空间 $\mathbf{Y}$ 是连续空间，则称预测任务为回归问题。

### 1.1 分类问题

#### 1.1.1 错误率与精度

错误率（error rate）是分类错误的样本数占样本总数的比例。记为 $\frac{1}{m}\sum_{i=1}^{m} I\left(f\left(\mathbf{x}_i\right) \neq y_i\right)$ ，其中 $I(\bullet)$ 为示性函数， $m$ 为测试集样本个数。

精度（accuracy）是分类正确的样本数占样本总数的比例。记为 $\frac{1}{m}\sum_{i=1}^{m} I\left(f\left(\mathbf{x}_i\right) = y_i\right)$ 。

#### 1.1.2 查准率、查全率与 F1

一般用于与信息检索相关的分类任务。在二分类的情形下，TP 指真正例（true positive）样本数， FP 指假正例（false positive）样本数，TN 指真反例（true negative）样本数，FN 指假反例（false negative）样本数。

查准率（precision）定义为 $P = \frac{TP}{TP + FP}$ 。

查全率（recall）定义为 $R = \frac{TP}{TP + FN}$ 。

F1 值定义为 $F1 = \frac{2 \times P \times R}{P + R}$ 。

查准率和查全率是一对矛盾的度量。一般来说，查准率高时，查全率往往偏低。查全率

高时，查准率往往偏低。F1 值对两者进行了权衡。

### 1.2 回归问题

#### 1.2.1 均方误差（mean squared error，MSE）

均方误差的定义为 $\frac{1}{m}\sum_{i=1}^{m}\left(f\left(\mathbf{x}_i\right)-y_i\right)^2$ 。

#### 1.2.2 平均绝对值误差（mean absolute error，MAE）

平均绝对值误差的定义为 $\frac{1}{m}\sum_{i=1}^{m}\left|f\left(\mathbf{x}_i\right)-y_i\right|$ 。

#### 1.2.3 平均绝对值百分比误差 $R^2$ （mean absolute percentage error，MAPE）

平均绝对值百分比误差的定义为 $\frac{1}{m}\sum_{i=1}^{m}\left|\frac{f\left(\mathbf{x}_i\right)-y_i}{y_i}\right|$ 。

#### 1.2.4 $R^2$ （coefficient of determination）

$R^2$ 的定义为 $R^2=1-\frac{\sum_i\left(y_i-f\left(\mathbf{x}_i\right)\right)^2}{\sum_i\left(y_i-\bar{y}\right)^2}$ 。取值越大越好。

## 2 业务评测标准

在教育应用中，语言智能算法一般内置于某个系统中，通常对系统的整体性能进行测试。这种测试称为体内测试（in vivo）。体内测试一般根据从特定教育应用效果的角度出发设定评测指标，可以反映语言智能算法对教育应用的作用。

### 2.1 题目知识点预测

知识点预测通常是一个多分类问题。由于每个学科包括数百至上千个知识点，学科老师通常关注预测结果是否包含了标准答案。目前，我们采用五选精度对算法进行评估。五选精度指的是，对于每个待预测的题目，系统给出 5 个知识点，若标准答案在这个 5 个知识点中，则认为预测正确，然后采用精度（accuracy）的公式进行计算。

### 2.2 题目难度预测

题目难度预测通常是一个回归问题。题目难度归一化至[0, 1]之间，归一化后采用平均绝对值误差（mean absolute error，MAE）为测试指标。