

思考题汇总

思考题需要在报告中进行回答

步骤2：根据相似度加权

计算test1与每个train的距离，选取TopK个训练数据
把该距离的倒数作为权重，计算test1属于该标签的概率：

$$P(\text{test1 is happy}) = \frac{\text{train1 probability}}{d(\text{train1}, \text{test1})} + \frac{\text{train2 probability}}{d(\text{train2}, \text{test1})} + \frac{\text{train3 probability}}{d(\text{train3}, \text{test1})}$$

思考：为什么是倒数呢？

思考：同一测试样本的各个情感概率总和应该为1 如何处理？

不同距离度量方式

- 距离公式：

L_p 距离(所有距离的总公式)：

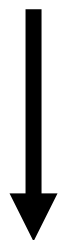
- $$L_p(x_i, x_j) = \left\{ \sum_{l=1}^n \left| x_i^{(l)} - x_j^{(l)} \right|^p \right\}^{\frac{1}{p}}$$

- $p = 1$ ：曼哈顿距离；
- $p = 2$ ：欧式距离，最常见。

思考：在矩阵稀疏程度不同的时候，这两者表现有什么区别，为什么？

Example

ID	text	class label
1	good,thanks	joy
2	No impressive, thanks	sad
3	Impressive good	joy
4	No, thanks	?



ID	goods	thanks	no	impressive	class label
1	1	1	0	0	joy
2	0	1	1	1	sad
3	1	0	0	1	joy
4	0	1	1	0	?

Bernoulli Model (伯努利模型) :

$$P_{(\text{thanks}|\text{joy})} = 1/2$$

Multinomial Model (多项式模型) :

$$P_{(\text{thanks}|\text{joy})} = 1/4$$

思考题：这两个模型分别有什么优缺点

Task

- (1) 分类 (使用**准确率**衡量结果)
分类只要求实现**多项式模型**
- (2) 回归 (使用**相关系数**衡量结果)
 - 归一化最后的情感概率, 使得六中情感概率相加为 1
 - 本次实验同样提供了 validation 数据集
- (3) 推荐实现拉普拉斯平滑

思考题: 如果测试集中出现了一个之前全词典中没有出现过的词该如何解决