# Naive Bayes Practice / 朴素贝叶斯练习

## Part I: Questions / 第一部分：问题

Dataset 1 Overview (数据概览 1)

| | |
|---|---|
| Part 1: Training Data (Movie Reviews) | 第一部分：训练数据（电影评论） |
| Based on the provided slides, we have the following training corpus: | 基于提供的幻灯片，我们有以下训练语料库： |
| Total Docs: 7 (Pos: 3, Neg: 4). Vocabulary Size: 12. | 文档总数：7 (正面: 3, 负面: 4)。词汇表大小：12。 |

| | |
|---|---|
| Exercise 1: Standard Naive Bayes | 练习 1：基础朴素贝叶斯 |
| Classify the following sentence: 'This was a fantastic story, great, lovely' | 对以下句子进行分类： 'This was a fantastic story, great, lovely' |
| Exercise 2: Repeated Words | 练习 2：重复词语 |
| Classify the following sentence: 'Great great great' | 对以下句子进行分类： 'Great great great' |
| Exercise 3: Zero Probability | 练习 3：零概率问题 |
| Classify the following sentence: 'Excellent cast, unimaginative ending' | 对以下句子进行分类： 'Excellent cast, unimaginative ending' |
| Exercise 4: Laplace Smoothing | 练习 4：拉普拉斯平滑 |
| Recalculate the classification for Exercise 3 ('Excellent cast, unimaginative ending') using Laplace (Add-1) Smoothing. | 使用拉普拉斯（加一）平滑重新计算练习 3 ('Excellent cast, unimaginative ending') 的分类结果。 |

Dataset 2 Overview (数据概览 2)

| | |
|---|---|
| Part 2: New Training Data | 第二部分：新训练数据 |
| Docs 1-3: Positive (words: sensitive, moving, brilliant, edgy...) Docs 4-7: Negative (words: neither, revelatory, flamboyant, awkward...) | 文档 1-3：正面 (单词: sensitive, moving, brilliant, edgy...) 文档 4-7：负面 (单词: neither, revelatory, flamboyant, awkward...) |

| | |
|---|---|
| Exercise 5: Standard NB (No Smoothing) | 练习 5：基础朴素贝叶斯（无平滑） |
| Classify the text (Doc 8) using Standard NB without smoothing: 'A flamboyant romcom, sensitive but awkward at times.' | 使用无平滑的基础朴素贝叶斯对文本 (Doc 8) 进行分类： 'A flamboyant romcom, sensitive but awkward at times.' |
| Exercise 6: Binary Naive Bayes | 练习 6：二值朴素贝叶斯 |
| Classify the same text (Doc 8) using Binary Naive Bayes (No smoothing). Count presence (1) or absence (0) instead of frequency. | 使用二值朴素贝叶斯（无平滑）对同一文本 (Doc 8) 进行分类。统计是否存在 (1或0) 而非词频。 |

## Part 3: Reflection / 第三部分：思考

| Bonus Reflection: Frequency vs. Presence | 思考题：词频 vs 存在 |
| --- | --- |
| Imagine a document has the word 'good' 20 times and 'bad' 1 time. How would the Standard Naive Bayes model differ from the Binary Naive Bayes model in handling this? | 想象一个文档中单词 'good' 出现了 20 次，而 'bad' 出现了 1 次。 基础朴素贝叶斯模型和二值朴素贝叶斯模型在处理这种情况时会有什么不同？ |

## Part II: Answers / 第二部分：答案

| Answer 1 | 答案 1 |
|---|---|
| Result: Positive P(+) * P(fan\|+) * P(grt\|+) * P(lov\|+) = 0.00043 P(-) term becomes 0 because P(fantastic\|-) is 0. | 结果：正面 (Positive) P(+) * P(fan\|+) * P(grt\|+) * P(lov\|+) = 0.00043 P(-) 项变为 0，因为 P(fantastic\|-) 为 0。 |
| Answer 2 | 答案 2 |
| Result: Negative P(+) = 3/7 * (1/10)^3 = 0.00043 P(-) = 4/7 * (2/8)^3 = 0.00893 Negative probability is higher. | 结果：负面 (Negative) P(+) = 3/7 * (1/10)^3 = 0.00043 P(-) = 4/7 * (2/8)^3 = 0.00893 负面概率更高。 |
| Answer 3 | 答案 3 |
| Result: Undefined (Zero Probability) P(unimaginative\|+) is 0, making Positive prob 0. P(excellent\|-) is 0, making Negative prob 0. Model fails without smoothing. | 结果：无法判断（零概率） P(unimaginative\|+) 为 0，导致正面概率为 0。 P(excellent\|-) 为 0，导致负面概率为 0。 没有平滑处理，模型失效。 |
| Answer 4 | 答案 4 |
| Result: Negative Apply smoothing (add 1 to numerator, add 12 to denominator). Pos: 3/7 * (2/22) * (1/22) = 0.00176 Neg: 4/7 * (1/20) * (2/20) = 0.00286 | 结果：负面 (Negative) 应用平滑（分子加 1，分母加 12）。 正面: 3/7 * (2/22) * (1/22) = 0.00176 负面: 4/7 * (1/20) * (2/20) = 0.00286 |
| Answer 5 | 答案 5 |
| Result: Negative Word 'awkward' appears in Negative docs but NOT in Positive docs. P(awkward\|+) = 0 -> Pos prob = 0. P(awkward\|-) > 0 -> Neg prob > 0. | 结果：负面 (Negative) 单词 'awkward' 出现在负面文档中，但未出现在正面文档中。 P(awkward\|+) = 0 -> 正面概率 = 0。 P(awkward\|-) > 0 -> 负面概率 > 0。 |
| Answer 6 | 答案 6 |
| Result: Negative Binary NB checks if word exists in the class documents. 'awkward' exists in 0/3 Positive docs -> P(awkward\|+) = 0. 'awkward' exists in 1/4 Negative docs -> P(awkward\|-) = 0.25. Positive prob is 0. | 结果：负面 (Negative) 二值朴素贝叶斯检查单词是否存在于该类别的文档中。 'awkward' 出现在 0/3 个正面文档中 -> P(awkward\|+) = 0。 'awkward' 出现在 1/4 个负面文档中 -> P(awkward\|-) = 0.25。 正面概率为 0。 |
| Reflection Answer | 思考题答案 |
| Standard NB: Multiplies P('good'\|class) 20 times. It cares deeply about frequency/intensity. Binary NB: Counts 'good' only once (1). It only cares that the word appeared, ignoring intensity. | 基础模型：会将 P('good'\|类别) 连乘 20 次。它非常关注词频和情感强烈程度。 二值模型：只计算 'good' 一次 (1)。它只关注单词是否出现，忽略了强度。 |