Machine Learning and Real-world Data

Example Sheet 2 - Zejia Yang

February 12, 2024

Statistical testing

1

Let the accuracies of system 1 and system 2 be A_1 and A_2, and the proportion of overlapping accuracies is B. Let **POS** be the proportion of 1 wins 2 and **NEG** be the proportion of 2 wins 1.

$$\begin{aligned} ties &= B + (1 - A_1 - A_2 + B) = 2B + 1 - A_1 - A_2 && (1)\\ POS &= A_1 - B && (2)\\ NEG &= A_2 - B && (3) \end{aligned}$$

Hence we have:

$$\begin{aligned} k &= ties/2 + min(POS, NEG) && (4)\\ &= 100(\frac{1 - A_1 - A_2}{2} + min(A_1, A_2)) = 50(1 - |A_1 - A_2|) && (5) \end{aligned}$$

2

To be honest, I don't quite get this question. I can't come up with how the number of ties can be used in designing an improved model. However, I think the number of ties can be used to further distinguish the differences between different models. For example, if two systems (B and C) both pass the sign test with the same k when compared to A, we should choose the system with the least number of ties (for example, B). This means B and C are both better than A, while B may be slightly better than C since B improves more.

Overtraining and cross-validation

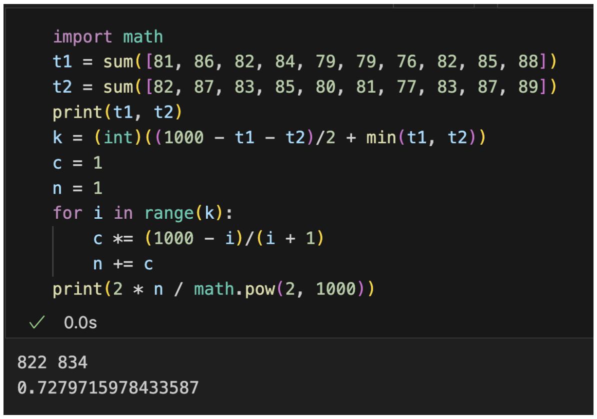
1

mean accuracy: 82 (82.2)%; variance: 0.0012 (0.001196)
I'm not sure about the precision. If it is the same with the amount of significant digits of the value with the least amount. Then it should be rounded to 1 significant number (same with 100)
mean accuracy : 80%; variance: 10. However, here I will stick with the principle that precision should be a soft indication of variance and round it to 2 significant numbers.

2

The mean accuracy of the alternative system: 0.83. The variance : 0.0012.
Although it is not explicitly stated, I assume the *result* in the question is that the alternative system is better than the previous system.

Null Hypothesis: the alternative system(**2**) has the same performance with the original system(**1**). The data implies that the results from the second system are just marginally better than 1 for each fold. This may be due to the fact that test sets were chosen with bias. Hence, my pre-assumption is that we can't reject the null hypothesis and it is not statistically significantly at 5% level.
There are three ways of explanation.
Statistics The variance of the first system is around 0.82 and the standard deviation is around 0.034. Clearly, 0.83 lie within 0.82 +/- 0.034. It can't be significant at 5% level.
Sign Test Using the formula k in the first question (assuming no correction is needed for the number of ties being odd), view the 10-fold test sets as a total test set of 1000. calculate the two-tailed p-value: 0.728 > 0.025. Thus, it fails the sign-test.



```
import math
t1 = sum([81, 86, 82, 84, 79, 79, 76, 82, 85, 88])
t2 = sum([82, 87, 83, 85, 80, 81, 77, 83, 87, 89])
print(t1, t2)
k = (int)((1000 - t1 - t2)/2 + min(t1, t2))
c = 1
n = 1
for i in range(k):
    c *= (1000 - i)/(i + 1)
    n += c
print(2 * n / math.pow(2, 1000))
```

✓ 0.0s

```
822 834
0.7279715978433587
```

T-test to verify my prior results.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}$$

The t-value for these two samples is $0.7303 < t_{0.975}(18) = 2.101$. We can't reject our null hypothesis. Hence, we can't conclude that the alternative system is significantly better than the original ones.

3

The "Wayne Rooney" Effect. Time changes public opinions on particular people or effects. For an old model trained using bag-of-words, it will be more obvious since the bag-of-words model simply ignores the position and relation between tokens and only focuses on the frequency of their type. As the meaning and polarity of the words change with time, the accuracy of your model will decline dramatically based on that.

Statistical properties of language

1

Adding a third category may not necessarily improve Cohen's kappa. There are the qualitative reasons?

My Question: how to explain it quantitatively?
What factors influence the kappa scores?[prevalence, bias, and nonindependent ratings, ...]

prevalence (are the classes equiprobable or do their probabilities vary) and bias (are the marginal probabilities for the two observers similar or different).

1. Increased Ambiguity: The interpretation of neutral sentiment can vary widely among annotators and may depend on individual perspectives or contextual factors. For instance, a review containing both positive and negative aspects may be classified as neutral by some annotators, while others may classify it as positive or negative based on their subjective judgment. Hence, adding another category will increase uncertainty, which undermines agreement. My assumption is that people tend to choose neutral classes when they are indecisive, or the review doesn't explicitly express the obvious feelings. This causes an imbalanced distribution among classes, and since it applies to all samples, it may result in a low kappa.

2. Increase Annotation Noise: Annotation noise refers to inconsistencies or errors in the annotations made by human annotators. As the number of categories increases, annotators may struggle to make accurate and consistent judgments, leading to a higher likelihood of noise in the annotations.

Above all, I think the key problem in adding another category is that the difficulty in deciding across multiple classes will increase in the meantime, despite some samples being more accurately described as neutral (since other samples will need to be re-judged as well).

2

The advantages of having human annotations.

1. Human agreement is the only empirically available source of truth in decisions that are influenced by subjective judgement. For example, for tasks like sentiment analysis, subjective judgements on a review may vary greatly among different reviewers. And crowd-sourcing will give you a more impartial and objective label, which will improve data quality and facilitate the training process.

2. Data verification can be done with human annotations. To provide a predicted rating, for instance, we might first employ an automated algorithm in the movie rating process. Following that, multiple human evaluations can be used to support the accuracy of the model labelling.

3. Additionally, even though model labelling works well in most cases, there may be some exceptional or edge cases where the automated process is unable to recognise a particular pattern or perform well. In these cases, human annotations are required to handle the difficult cases.