MLRD Supervision 2

Neelu Saraswatibhatla (srns2)

Statistical testing

1. Of the two systems being tested, let a be the number of documents one system correctly classifies but the second system doesn't, and let b be the number of documents the second system correctly classifies but the first system doesn't, with $a \leq b$.Let c be the number of ties, i.e. $c = 100 - a - b$.

   $$k = a + c = a + 100 - a - b = 100 - b \implies b = 100 - k$$

   The accuracy of the first system $\text{acc}_a = \frac{a+c}{100}$ and the accuracy of the second system $\text{acc}_b = \frac{b+c}{100}$.

   $$\text{acc}_a = \frac{a+c}{100} = \frac{a+100-a-b}{100} = \frac{100-b}{100} = 0.01k$$

   $$\text{acc}_b = \frac{b+c}{100} = \frac{100-k+c}{100} = 1 - \text{acc}_a + 0.01c$$

2. By looking at what the ties are, we can directly focus on those documents which both systems did equally badly on to improve the systems, and keep the traits which caused both systems to do equally well on.

Overtraining and cross-validation

1.

$$\begin{aligned}
\text{mean} = \bar{x} &= \frac{\sum x_i}{n} \\
&= \frac{81+86+82+84+79+79+76+82+85+88}{10} \\
&= 82.2 \\
\text{variance} &= \frac{\sum (x_i - \bar{x})^2}{n} \\
&= \frac{1.2^2 + 3.8^2 + 0.2^2 + 1.8^2 + 3.2^2 + 3.2^2 + 6.2^2 + 0.2^2 + 2.8^2 + 5.8^2}{10} \\
&= 11.96
\end{aligned}$$

2. Each result in this second system is 1 greater than the first system (with the exception of 79 $\rightarrow$ 81 which is 2 greater). Therefore the second system correctly classified 11 more documents than the first system did.

The most favourable situation for these results to be statistically significant is therefore for system 2 to be better than system 1, and for it to have the maximum difference in sign test results, i.e. classifying as many documents as possible correctly that system 1 didn't and correctly classifying all documents that system 1 correctly classified. If documents 0 to 821 inclusive were correctly classified by the first system, and documents 0 to 832 inclusive were correctly classified by the second system, then the second system correctly classified 11 documents that the first one didn't, and the first one correctly classified no documents that the second one didn't. The others were all ties, so we add 0.5 * 989 to each side, getting 494.5 for the first system and 494.5 + 11 = 505.5 for the second system. We round up both at the end, getting 495 for System 1 and 506 for System 2.

To get the result of the sign test we therefore find $\sum_{i=0}^{495} \binom{495+506}{i} 0.5^i (1-0.5)^{495+506-i} \approx 0.38$, which is far greater than the 0.025 required to pass a two-tailed sign test, or even the 0.05 required to pass a one-tailed sign test.

3. When a system is trained on older data, it may not work as effectively on newer data as public perceptions change, such as in the case of the 'Wayne Rooney effect' where the general perception of Wayne Rooney was good in the past before a series of scandals turned that public image around.

   A system trained on older data may learn that certain actors are good actors and be more likely to classify reviews they appear in as positive, but if in the future they get negative publicity then a future review is more likely to be negative but the system trained on old data is more likely to incorrectly classify it as positive.

Uncertainty and human agreement

1. There could be several reasons for this:

   (a) Kappa doesn't work well with small datasets, so even with a third category may not be much higher with just four documents.

   (b) People would still likely disagree about the sentiment of each review, and especially with neutral reviews, people may be more likely to vote positive or negative because of small individually perceived subtleties, which may actually decrease kappa.

2. Different people may interpret the star rating system differently, so some people may be more inclined to vote higher while others lower, and it would be useful to have human annotation on a sample so that this variation can be normalised for future samples without human annotation by looking at the difference between estimated ground truth and the human annotation and extrapolating.