David Murray
for
Sadiq Jaffer

MLRD Supo 2 work

2. Both filters performed quite well. Filter 1 got 80 predictions correct and Filter 2 got 84.
I would definitely choose filter 2 in practice though since I wouldn't want non-spam to be flagged as spam as I could miss important emails. Filter 2 only had 1 false positive in comparison to filter 1 which had 16.

Statistical testing

1. Say system 1 has accuracy A_1 and system 2 has accuracy A_2.
For an event E the chance of system 1 beating system 2 is $A_1(1-A_2)$,
The chance of 2 beating 1 is $(1-A_1)\cdot A_2$ and a tie is $A_1 \cdot A_2 + (1-A_1)(1-A_2)$.
Assume we're trying to prove system 1 is significantly better than system 2.
Let's say we have N outcomes. Then k is the number of negative outcomes (2 beats 1).
That is,

$$k = \underbrace{(1-A_1)\cdot A_2 \cdot N}_{\text{neg events}} + \underbrace{(A_1 \cdot A_2) + (1-A_1)\cdot(1-A_2)}_{\text{ties} \div 2 \text{ since adding 0.5 for ties}} \cdot N \cdot \frac{1}{2}$$

2. ? Maybe subtract ties from number of outcomes and just compare positive and negative.

Overtraining and cross-validation

1. mean = 82.2
variance = 11.96

2. mean = 83.4
variance = 12.04

No since, ~~for example, in a~~ when tested on 100 items you'd only expect system 2 to ~~get~~ beat system 1 on 1 items and then get the rest as ties. Then we'd calculate the $P(X \leq 49)$ which will be greater than 0.5.

3. The "Wayne Rooney" effect - the opinion of the public on people or events can change over time. The words people use change over time as well.

Uncertainty and human agreement

1. It can decrease the number of pairwise agreements which would decrease $\bar{P}_a$ and in turn decrease kappa since $\kappa = \frac{\bar{P}_a - \bar{P}_e}{1 - P_e}$

2. People can have different ~~opinions~~ opinions so the reviewer may have been an outlier compared to the majority ~~opinion.~~ opinion.