Supervision questions: set 2.

Statistical testing

1. Assume that two systems are used for a binary classification task on 100 test items, and that accuracy is calculated for each system. Derive the relationship between the accuracies and k as used in the sign test under the assumption that no correction is needed for the number of ties being odd.
2. The number of ties found between two systems is calculated as part of the sign test. How might this information be used in designing an improved system?

Overtraining and cross-validation

1. Suppose you test a binary classification system using 10-fold cross-validation with 100 items in each fold. You obtain the following results for the folds: 81, 86, 82, 84, 79, 79, 76, 82, 85, 88. What is the mean accuracy and the variance?
2. An alternative system, tested using exactly the same folds, gives the following results: 82, 87, 83, 85, 80, 81, 77, 83, 87, 89. Could this result be statistically significant at the 5% level? Explain your answer. (Full significance testing is not required.)
3. What effects can cause the accuracy of a sentiment analysis system trained on old data using bag-of-words to decrease when applied to later data?

Uncertainty and human agreement

1. The experiment in Task 1 where you all had to choose between positive and negative for a movie review which was more accurately described as neutral sentiment demonstrates that kappa will be much higher on some items than others. However, adding a third category won't necessarily improve kappa. Why not?
2. Why might it be informative/useful to use human annotation on a sample of data even if you already have annotation which corresponds closely to ground truth, such as movie review stars?