Performance measures used in the Protein Matching Problem

The results we would submitted can be an ASCII file, each line of which contains a block ID number, an example ID number and the predicted probability that the case is in class 1, separated by any whitespaces (spaces, tabs, commas etc.).

The input file of the Perf program can also be in ASCII format. Typically, each lines consists of a block ID number, a target value (class) and the predicted probability separated by any white space.

Here is a toy example of an input file "tmp.txt".

```
1 1 .9
1 1 .8
2 0 .9
2 1 .5
1 0 .7
```

The usage of the Perf program:

```
PATH/perf [options] -file InputFileName
```

For our problem, four performance measures are used and the command may be:

```
PATH/perf -top1 -rms -rkl -apr -blocks -file tmp.txt
MEAN_BLOCK_APR     0.25000
MEAN_BLOCK_RKL     2.00000
MEAN_BLOCK_RMS     0.57614
MEAN_BLOCK_TOP1    0.50000
```

For each measure, e.g., RMS, the Perf program calculates the score for each block and then averages them up over all blocks to get the "MEAN_BLOCK" measure.

Within each block, the four measures work this way:

- **TOP1**: Sort the data within block according to the predicted probabilities in descending order. TOP1= 1 if the top ranked case is in class 1; TOP1= 0 otherwise. (If ties occur among predicted probabilities, the corresponding target class values

would be changed by their average. TOP1= 1 if the average = 1, TOP1= 0 otherwise.)

- **RKL**: RKL= rank of the last class 1 case in the above sorted data.

- **RMS**: The usual root of mean squared error.

- **APR**: Basically, this is the area under the Precision/Recall curve.

  (Assume no ties occur in predicted probabilities for the moment. In case of ties, just replace target values by their average and the program takes ties into consideration already.)

  Suppose there are N cases altogether in a block and n cases of them are in class 1. First we sort cases according to the predicted probabilities in descending order (prioritize cases). Let t_i be the target value of the ith ranked case. The precision and recall at the ith case are

$$p_i = \frac{1}{i}\sum_{j=1}^{i} t_i,$$

$$r_i = \frac{1}{n}\sum_{j=1}^{i} t_i,$$

  respectively. Let $i_0 = \min\{i : t_i > 0\}$. The version of average precision in the program is

$$\text{APR} = \sum_{i=i_0+1}^{N} \frac{1}{2}(p_i + p_{i-1})(r_i - r_{i-1}).$$