Appendix Learning From Mistakes: Machine Learning Enhanced Human Expert Effort Estimates

Federica Sarro, Rebecca Moussa, Alessio Petrozziello and Mark Harman

In this appendix we explain the mathematical formulation of the Linear Programming model we used in RQ2 to predict the MisestimationMagnitude.

Linear Programming (LP) [**?**] aims to achieve the best outcome from a mathematical model with a linear objective function subject to linear equality and inequality constraints. The feasible region is given by the intersection of the constraints and the Simplex (linear programming algorithm) is able to find a point in the polyhedron where the function has the smallest value (minimisation) in polynomial time.

Here, we generalize the model proposed for the effort estimation by Sarro and Petrozziello [**?**]. In the original implementation, the model is subject to an inequality constraint imposing that the value estimated for each of the observations in the training set has to fall in R_0^+. Here, we remove the inequality constraints allowing the model to use both positive and negative feature values as well as to optimize for both positive and negative values, as follows:

$$\begin{aligned} \text{minimise} \quad & \sum_{i=1}^{n} |\sum_{j=1}^{m} a_{ij}x_j - \mathit{ActualValue}_i| \\ & x_j \text{free}, \qquad j = 1, ..., m \end{aligned} \tag{1}$$

where a_{ij} represents the coefficient of the j^{th} feature for the i^{th} project, x_j is the value of the j^{th} feature, and *ActualValue*$_i$ is the actual effort of the i^{th} project.

Due to the non-linearity of the absolute value function, the above model has been linearised as follows:

$$\begin{aligned} \text{minimise} \quad & \sum_{i=1}^{n} t_i \\ \text{subject to} \quad & \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}x_j - \mathit{ActualValue}_i - t_i \leq 0 \\ & \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}x_j - \mathit{ActualValue}_i + t_i \geq 0 \\ & x_j \text{ free}, \qquad j = 1, ..., m \\ & t_i \text{ free}, \qquad i = 1, ..., n \end{aligned} \tag{2}$$

Let $X_i, \forall i$ be the part of Eq. (1) wrapped in the absolute value. $\forall i$, the slack variable t_i and the following two constraints have been added to the model: $X_i \leq t_i$ and $-X_i \leq t_i$. Therefore we can have one of the following cases:

$X_i > 0$ The second constraint, $-X_i \leq t_i$, is always fulfilled as $-X_i$ is negative and t_i is implicitly ≥ 0. Since t_i is minimised by the objective function and $0 \leq X_i \leq t_i$, the first constraint, $X_i \leq t_i$, is satisfied and t_i is $abs(X)$.

$X_i < 0$ The first constraint, $X_i \leq t_i$, is always fulfilled as X_i is negative and t_i is implicitly ≥ 0. Since t_i is minimised by the objective function and $0 \leq -X_i \leq t_i$, the second constraint, $-X_i \leq t_i$, is satisfied and t_i is $abs(X)$.

$X_i = 0$ Both constraints are always fulfilled since t_i is implicitly ≥ 0. Since t_i is minimised by the objective function, $0 = X_i = t_i$. So t_i is $abs(X)$.

- *E-mail: f.sarro@ucl.ac.uk, rebecca.moussa.18@ucl.ac.uk, a.petrozziello@ucl.ac.uk, mark.harman@ucl.ac.uk*

Manuscript received 14 Mar. 2020; revised 16 Nov. 2020