Paper:
Understanding Black-box Predictions via Influence Functions
https://arxiv.org/pdf/1703.04730.pdf

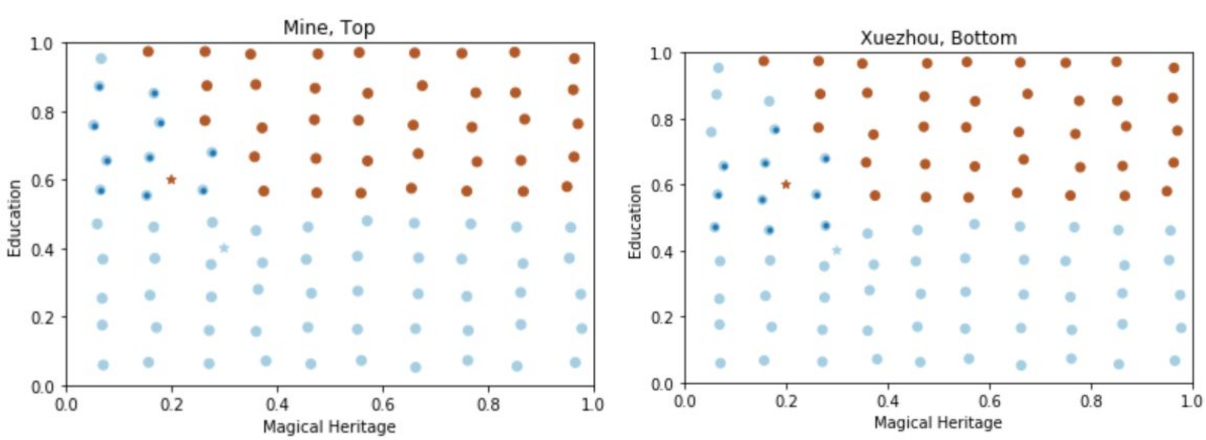
Calculation of Influence:

My approach:
From formula mentioned in original paper:

$$\begin{aligned}\mathcal{I}_{\text{up,loss}}(z, z_{\text{test}}) &\stackrel{\text{def}}{=} \frac{dL(z_{\text{test}}, \hat{\theta}_{\epsilon,z})}{d\epsilon}\Big|_{\epsilon=0} \\ &= \nabla_\theta L(z_{\text{test}}, \hat{\theta})^\top \frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon}\Big|_{\epsilon=0} \\ &= -\nabla_\theta L(z_{\text{test}}, \hat{\theta})^\top H_{\hat{\theta}}^{-1} \nabla_\theta L(z, \hat{\theta}).\end{aligned}$$

Xuezhou's approach:
Consider a weighted training set. Weights w for original training dataset, (1-w) for dataset with flipped labels. Influence = derivative of loss on trusted items wrt w.

Flagged points by the two methods:

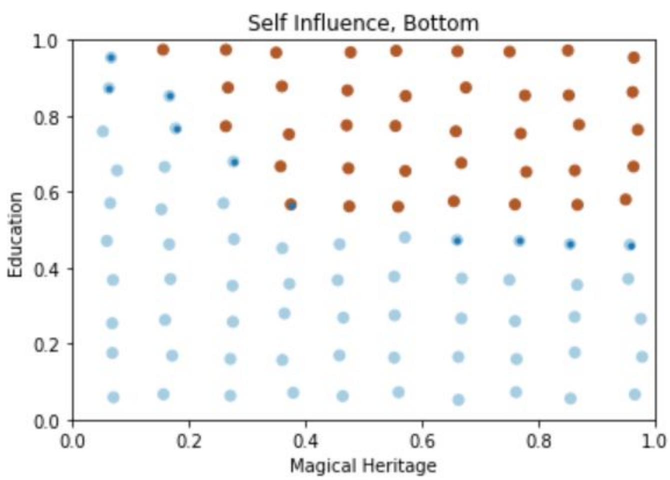


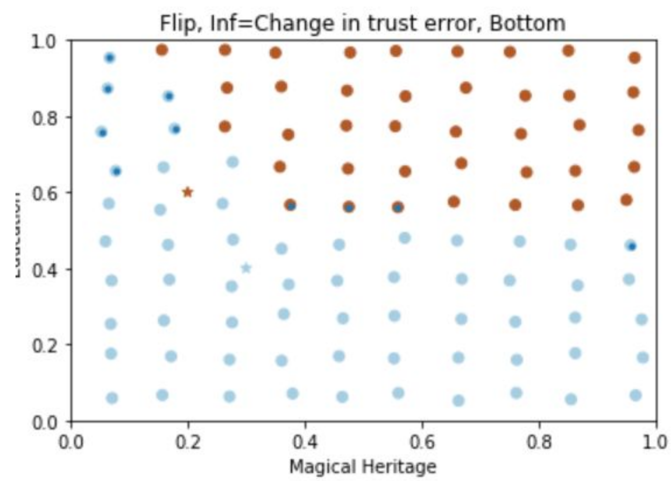
First principle estimation of influence in various different ways:

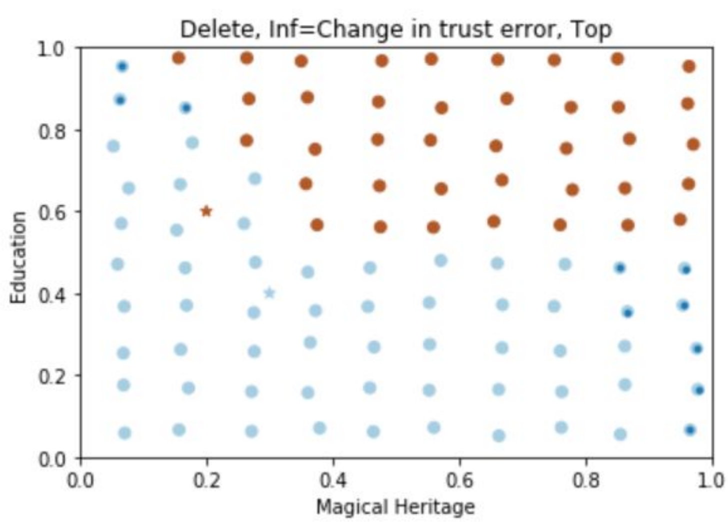
1. Calculating influence of a point on itself
2. Influence = Change in training loss when point label is flipped
3. Influence = Change in training loss when point is deleted from dataset
4. Influence = Change in loss on misclassified trusted items when point label is flipped
5. Influence = Change in loss on misclassified trusted items when point is deleted

Methods 2 and 3 yield no reasonable plots:

Plots which seem to flag up (some) meaningful points:







Shapley Value:
Intuitively, it is the contribution of a feature to a prediction.

Shapley Sampling: For approximating shapley values

Succinct summary of applying it in interpretability:
https://christophm.github.io/interpretable-ml-book/shapley.html

A Unified Approach to Interpreting Model Predictions (NIPS 2017):
https://arxiv.org/pdf/1705.07874.pdf

Codebase:
https://github.com/slundberg/shap