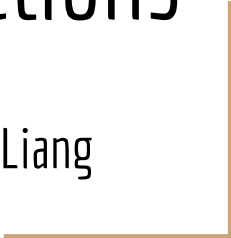


Understanding Black-box Predictions via Influence Functions



Pang Wei Koh & Percy Liang

*Wait, this isn't a
black box!*

Limitations from Dataset

- Assumes access to training data for 'black-box' models
- Need to go through every training sample for each target test sample
- How to effectively find target test samples in large datasets?
- How much explainability does a particular training sample offer?

Application to More Complex Problems

Authors used simple datasets and/or simple problems:

10 / 2 class MNIST, dog vs. fish, hospital readmission, email spam

How well would influence functions work on multiclass problems on more complex data?

Eg. Fixing mislabeled examples in multiclass problems where labels are not blatantly wrong



Label: Guitar



Label: Guitar



Label: Guitar



Label: Guitar

Scalability / Runtime

50,000 HVPs to evaluate the influence of *one* training point on *one* test point

With many training / test points...

They don't include runtime

Ideas for Extension

Training point selection:

If influence of train point is small on every test point...

... we know similar types of training points don't influence model

Perform this upweighting for features instead of training samples (?)

→ Similar to asking "How does a subset of training points affect the model?"