Classification in the Presence of Noisy Label

School of Intelligence Science and Technology, Peking University

Lecturer: Youcheng LI
Date: 9/26/2023

1. What is Label Noise

- Data for supervised learning consists of $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
- Some output labels y are incorect.
- Example: dogs-vs-cats

2. Label Noise is common

- Many datasets have errors in labelling, e.g., ImageNet, MNIST, e.t.c. [1]

ImageNet given label:
patas monkey

ImageNet given label:
shopping basket

ImageNet given label:
dough

ImageNet given label:
dugong

ImageNet given label:
wool

ImageNet given label:
terrapin

MNIST given label:
8

MNIST given label:
0

MNIST given label:
5

MNIST given label:
6

MNIST given label:
6

MNIST given label:
6

[1] https://labelerrors.com/

3. Label Quality is VITAL

北京大学 PEKING UNIVERSITY

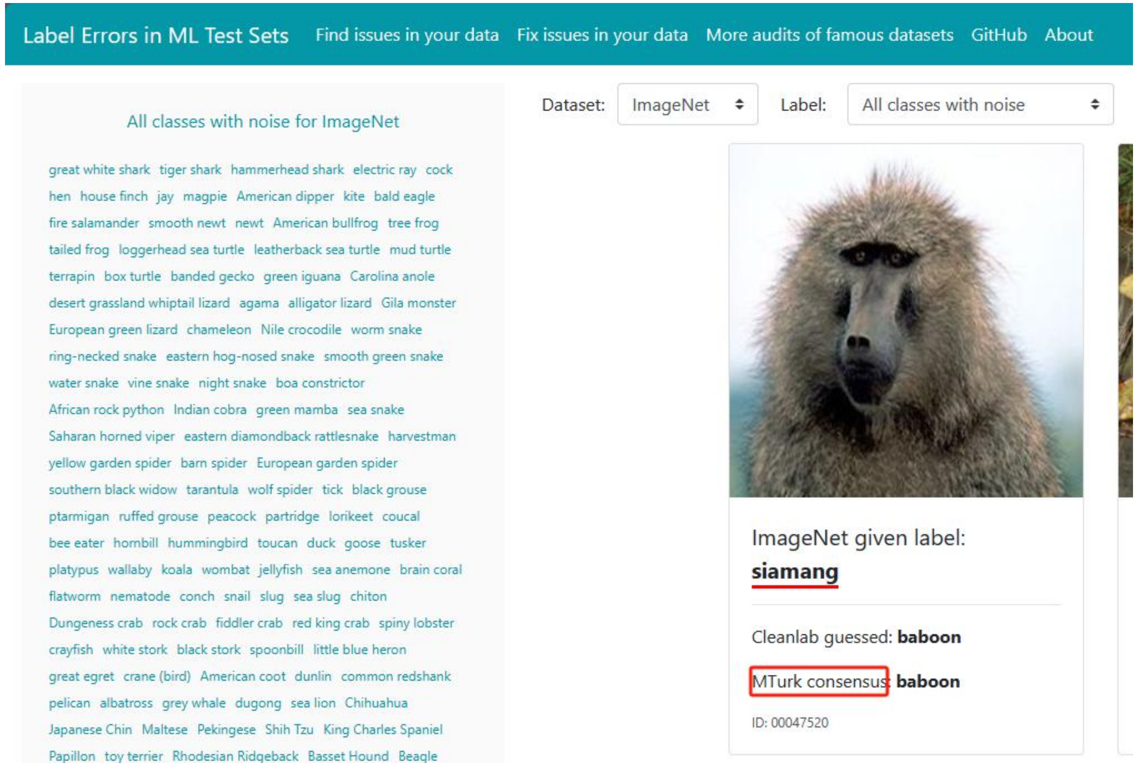
- LLAMA2: Quality Is All You Need. [1]

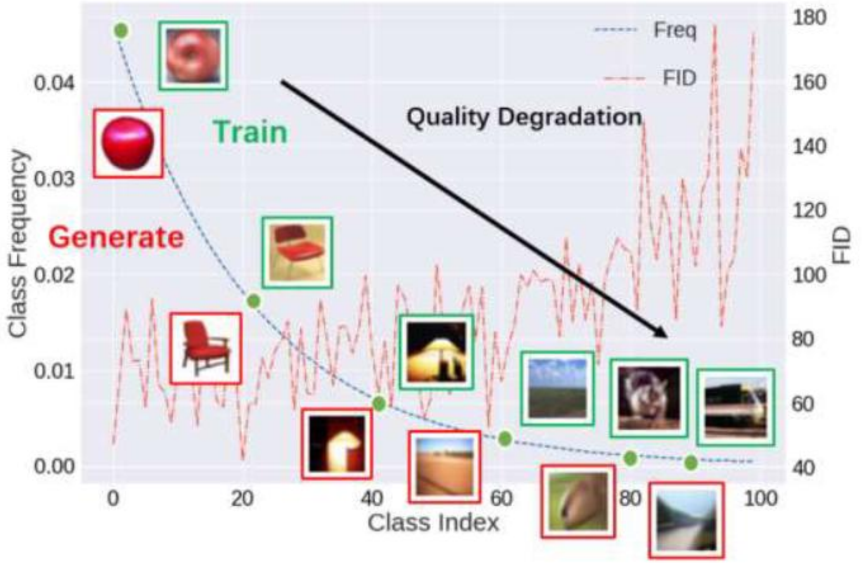
Quality Is All You Need. Third-party SFT data is available from many different sources, but we found that many of these have insufficient diversity and quality — in particular for aligning LLMs towards dialogue-style instructions. As a result, we focused first on collecting several thousand examples of high-quality SFT data, as illustrated in Table 5. By setting aside millions of examples from third-party datasets and using fewer but higher-quality examples from our own vendor-based annotation efforts, our results notably improved. These findings are similar in spirit to Zhou et al. (2023), which also finds that a limited set of clean instruction-tuning data can be sufficient to reach a high level of quality. We found that SFT annotations in the order of tens of thousands was enough to achieve a high-quality result. We stopped annotating SFT after collecting a total of 27,540 annotations. Note that we do not include any Meta user data.

[1] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models[J]. arXiv preprint arXiv:2307.09288, 2023.

4. Data Cleaning is HARD

- Noisy labels are common: Entry error; Inadequate information, e.t.c.
- Removing noisy labels is costly even impossible: money and time [1] (synthetic data [2]).

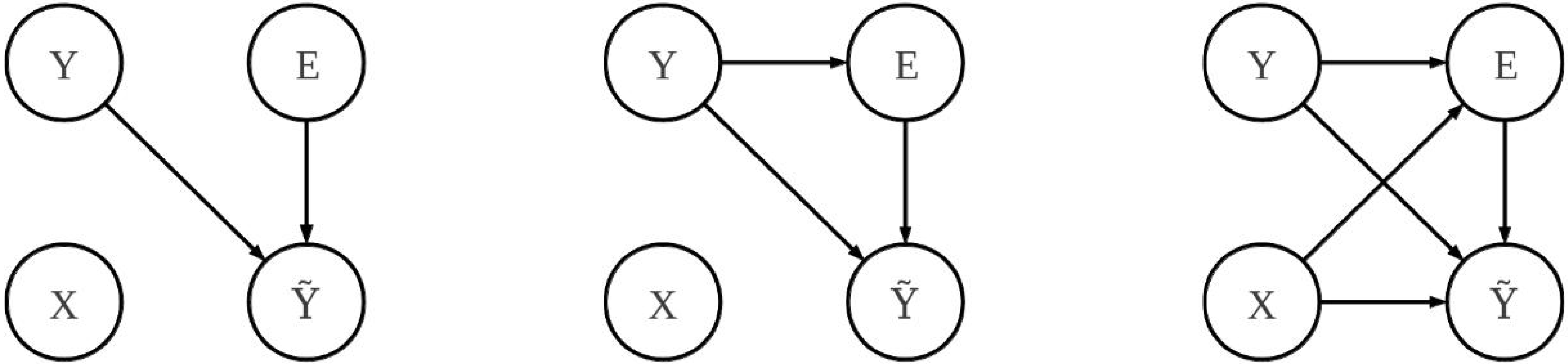




[1] https://labelerrors.com/
[2] Qin Y, Zheng H, Yao J, et al. Class-Balancing Diffusion Models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 18434-18443.

5. What is Label Noise: A Stochastic Process Perspective

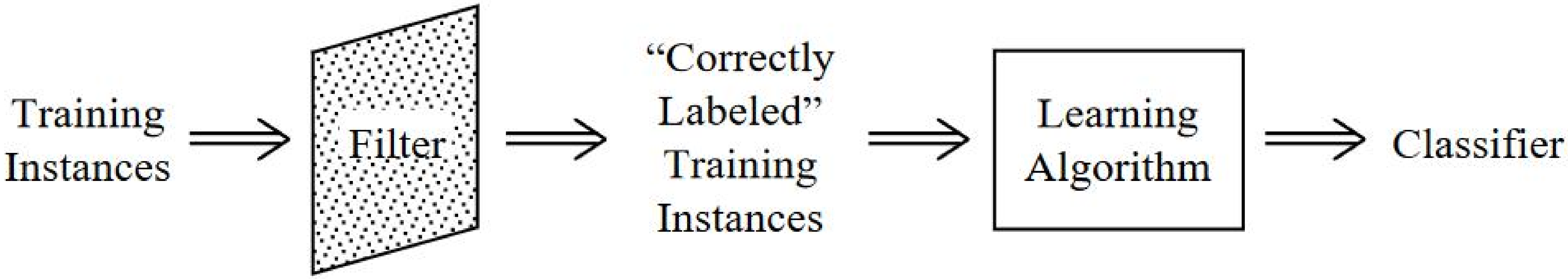
- Notation: X is instance, Y is the true label, $\tilde{Y}$ is the label with noise and E stands for error [1].
- Goal: Estimate the transition matrix.



[1] Frénay B, Verleysen M. Classification in the presence of label noise: a survey[J]. IEEE transactions on neural networks and learning systems, 2013, 25(5): 845-869.

5. What is Label Noise: A Stochastic Process Perspective

- Trivial Method: Train some classifiers on training set and infer on validation set [1].



[1] Brodley C E, Friedl M A. Identifying mislabeled training data[J]. Journal of artificial intelligence research, 1999, 11: 131-167.

北京大学 PEKING UNIVERSITY

Confident Learning: Estimating Uncertainty in Dataset Labels

Curtis G. Northcutt CGN@MIT.EDU
Massachusetts Institute of Technology, Department of EECS, Cambridge, MA, USA

Lu Jiang LUJIANG@GOOGLE.COM
Google Research, Mountain View, CA, USA

Isaac L. Chuang ICHUANG@MIT.EDU
Massachusetts Institute of Technology, Department of EECS, Department of Physics, Cambridge, MA, USA

[1] Northcutt C, Jiang L, Chuang I. Confident learning: Estimating uncertainty in dataset labels[J]. Journal of Artificial Intelligence Research, 2021, 70: 1373-1411.

6. Confident Lerning: Estimating Uncertainty in Dataset Labels

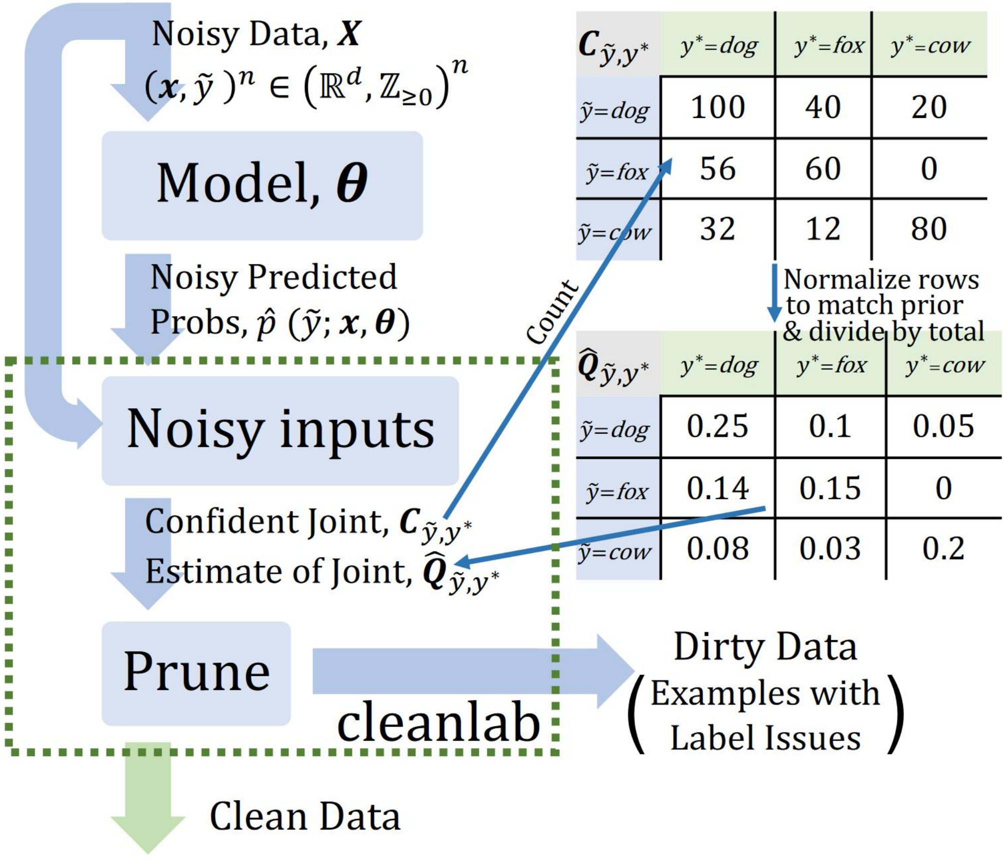
北京大学 PEKING UNIVERSITY

- Assumptions: There exists a latent true label y^* for every example. Prior to observing $\tilde{y}$, a class-conditional classification noise process mapping $y^* \rightarrow \tilde{y}$ (data-independent).
- Notation:
  - The subset of examples in X with noisy class label i is denoted $X_{\tilde{y}=i}$.
  - The discrete joint probability of noisy and latent labels as $p(\tilde{y}, y^*)$, where conditions $p(\tilde{y}|y^*)$ and $p(y^*|\tilde{y})$ denotes probabilities of label flipping.
  - The prior of latent labels is $Q_{y^*} := p(y^* = i)$.
  - The $m \times m$ joint distribution matrix is $Q_{\tilde{y},y^*} := p(\tilde{y} = i, y^* = j)$ (The Goal is to estimate it).
  - The $m \times m$ noise transition matrix (noisy channel) of flipping rates is $Q_{\tilde{y}|y^*} := p(\tilde{y} = i|y^* = j)$.
  - The $m \times m$ mixing matrix is $Q_{y^*|\tilde{y}} := p(y^* = j|\tilde{y} = i)$.
- Goal: Estimate $Q_{\tilde{y},y^*}$ and use it to find all mislabeled examples x in dataset X where $y^* \neq \tilde{y}$ **(HARD)**.

6. Confident Lerning: Method

Algorithm 1 (**Confident Joint**) for class-conditional label noise characterization.

input $\hat{\boldsymbol{P}}$ an $n \times m$ matrix of out-of-sample predicted probabilities $\hat{\boldsymbol{P}}[i][j] := \hat{p}(\tilde{y} = j; x, \boldsymbol{\theta})$
input $\tilde{\boldsymbol{y}} \in \mathbb{N}_{\geq 0}{}^n$, an $n \times 1$ array of noisy labels
procedure ConfidentJoint($\hat{\boldsymbol{P}}$, $\tilde{\boldsymbol{y}}$):
PART 1 (Compute thresholds)
for $j \leftarrow 1, m$ **do**
  for $i \leftarrow 1, n$ **do**
    $l \leftarrow$ new empty list []
    if $\tilde{\boldsymbol{y}}[i] = j$ **then**
      append $\hat{\boldsymbol{P}}[i][j]$ to l
  $\boldsymbol{t}[j] \leftarrow \text{average}(l)$ ▷ May use percentile instead of average for more confidence
PART 2 (Compute confident joint)
$\boldsymbol{C} \leftarrow m \times m$ matrix of zeros
for $i \leftarrow 1, n$ **do**
  $cnt \leftarrow 0$
  for $j \leftarrow 1, m$ **do**
    if $\hat{\boldsymbol{P}}[i][j] \geq \boldsymbol{t}[j]$ **then**
      $cnt \leftarrow cnt + 1$
      $y^* \leftarrow j$ ▷ guess of true label
  $\tilde{y} \leftarrow \tilde{\boldsymbol{y}}[i]$
  if $cnt > 1$ **then** ▷ if label collision
    $y^* \leftarrow \arg\max \hat{\boldsymbol{P}}[i]$
  if $cnt > 0$ **then**
    $\boldsymbol{C}[\tilde{y}][y^*] \leftarrow \boldsymbol{C}[\tilde{y}][y^*] + 1$
output $\boldsymbol{C}$, the $m \times m$ unnormalized counts matrix



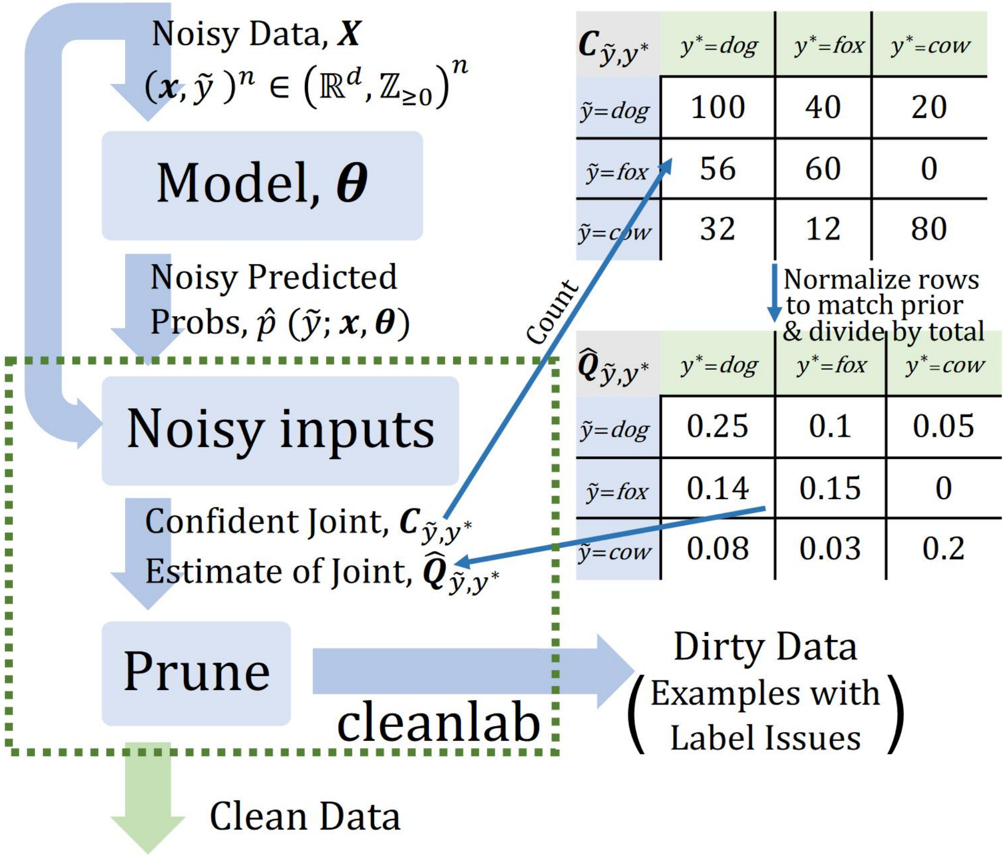
6. Confident Lerning: Method

	y*=0	y*=1	y~
0	0.9	0.1	0
1	0.9	0.1	0
2	0.5	0.5	0
3	0.3	0.7	0
4	0.3	0.7	0
5	0.2	0.9	1
6	0.2	0.8	1
7	0.4	0.7	1
8	0.5	0.5	1
9	0.6	0.4	1
10	0.9	0.1	0
11	0.9	0.1	0
12	0.5	0.5	0
13	0.3	0.7	0
14	0.3	0.7	0
15	0.2	0.9	1
16	0.2	0.8	1
17	0.4	0.7	1
18	0.5	0.5	1
19	0.6	0.4	1

$t_0 = 0.58, t_1 = 0.66$

$C_{\tilde{y},y^*}$

	true_0	true_1
pred_0	4	4
pred_1	2	6



6. Confident Lerning: Method

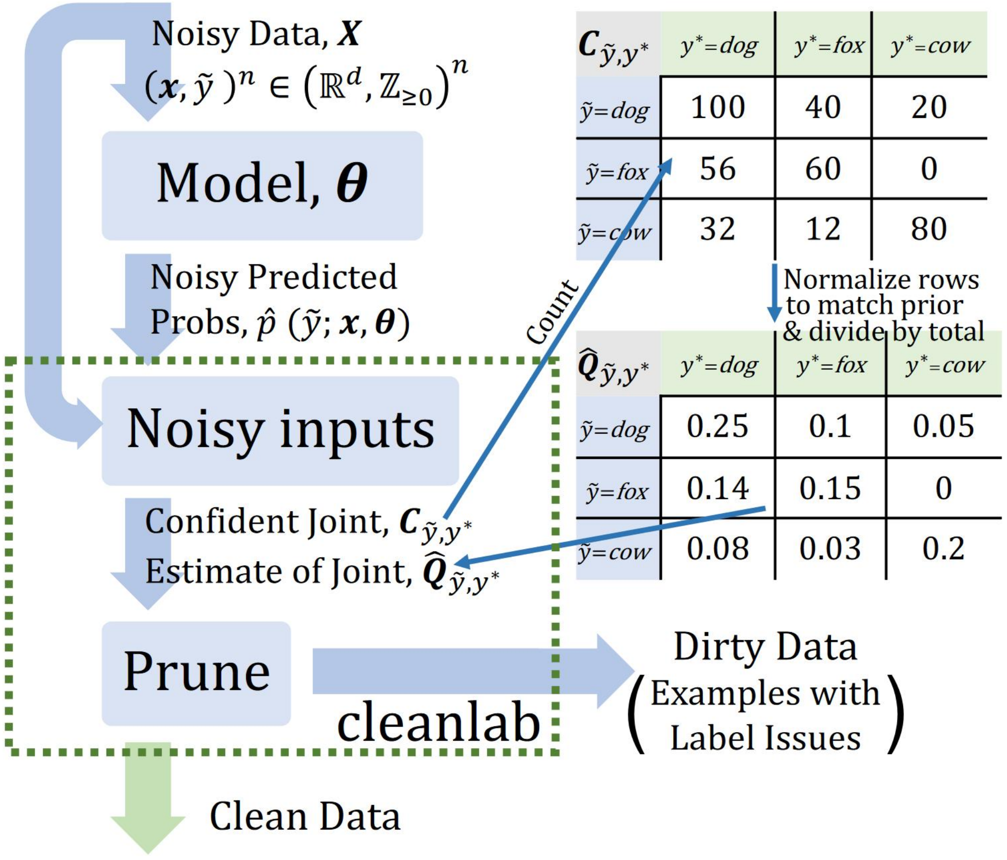
Algorithm 2 (**Joint**) calibrates the confident joint to estimate the latent, true distribution of class-conditional label noise

input $C_{\tilde{y},y^*}[i][j]$, $m\times m$ unnormalized counts
input $\tilde{\boldsymbol{y}}$ an $n \times 1$ array of noisy integer labels
procedure JointEstimation($\boldsymbol{C}$, $\tilde{\boldsymbol{y}}$):

$$\tilde{C}_{\tilde{y}=i,y^*=j} \leftarrow \frac{C_{\tilde{y}=i,y^*=j}}{\sum_{j\in[m]} C_{\tilde{y}=i,y^*=j}} \cdot |\boldsymbol{X}_{\tilde{y}=i}| \qquad \triangleright \text{ calibrate marginals}$$

$$\hat{Q}_{\tilde{y}=i,y^*=j} \leftarrow \frac{\tilde{C}_{\tilde{y}=i,y^*=j}}{\sum\limits_{i\in[m],j\in[m]} \tilde{C}_{\tilde{y}=i,y^*=j}} \qquad \triangleright \text{ joint sums to 1}$$

output $\hat{Q}_{\tilde{y},y^*}$ joint dist. matrix $\sim p(\tilde{y}, y^*)$



$\boldsymbol{C}_{\tilde{y},y^*}$	y^*=dog	y^*=fox	y^*=cow
$\tilde{y}$=dog	100	40	20
$\tilde{y}$=fox	56	60	0
$\tilde{y}$=cow	32	12	80

$\hat{\boldsymbol{Q}}_{\tilde{y},y^*}$	y^*=dog	y^*=fox	y^*=cow
$\tilde{y}$=dog	0.25	0.1	0.05
$\tilde{y}$=fox	0.14	0.15	0
$\tilde{y}$=cow	0.08	0.03	0.2

6. Confident Lerning: Method

	y*=0	y*=1	y~
0	0.9	0.1	0
1	0.9	0.1	0
2	0.5	0.5	0
3	0.3	0.7	0
4	0.3	0.7	0
5	0.2	0.9	1
6	0.2	0.8	1
7	0.4	0.7	1
8	0.5	0.5	1
9	0.6	0.4	1
10	0.9	0.1	0
11	0.9	0.1	0
12	0.5	0.5	0
13	0.3	0.7	0
14	0.3	0.7	0
15	0.2	0.9	1
16	0.2	0.8	1
17	0.4	0.7	1
18	0.5	0.5	1
19	0.6	0.4	1

$t_0 = 0.58, t_1 = 0.66$

$C_{\tilde{y},y^*}$

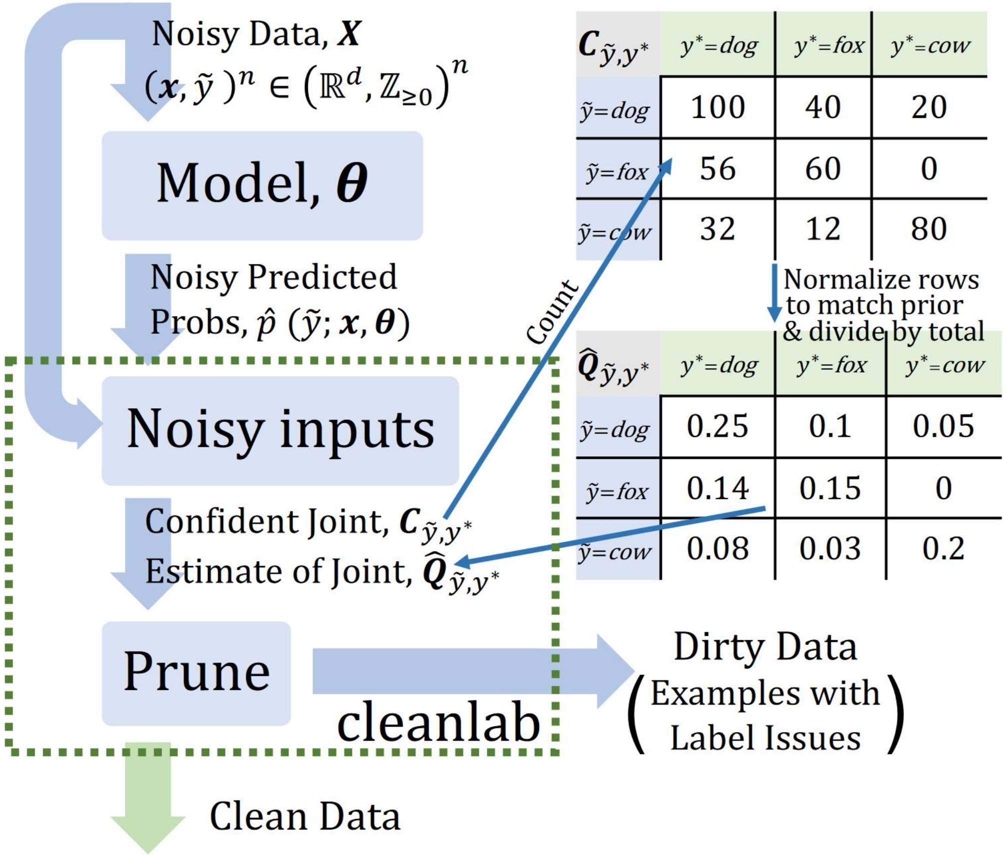
	true_0	true_1
pred_0	4	4
pred_1	2	6

$\widetilde{C}_{\tilde{y},y^*}$

	true_0	true_1
pred_0	5.0	5.0
pred_1	2.5	7.5

$\widehat{Q}_{\tilde{y},y^*}$

	true_0	true_1
pred_0	0.250	0.250
pred_1	0.125	0.375



6. Confident Lerning: Method

- Approach 1: Use off-diagnoals of $C_{\tilde{y},y^*}$ to estimate $\hat{X}_{\tilde{y}=i,y^*=j}$
  1. $C_{confusion}$. Estimate label errors as the Boolean vector $\tilde{y}_k \neq \arg\max_{j\in[m]} \hat{p}(\tilde{y}=j; x_k, \theta)$, for all $x_k \in X$, where true imples label error and flase implies clean data.
  2. $C_{\tilde{y},y^*}$. Estimate label errors as $\{x \in \hat{X}_{\hat{y}=i,y^*=j}: i \neq j\}$ from the diagnoals of $C_{\tilde{y},y^*}$.
- Approach 2: Use $n \cdot \hat{Q}_{\tilde{y},y^*}$ to estimate $|\hat{X}_{\hat{y}=i,y^*=j}|$, prune by probability ranking.
  1. Prune by Class. For each class $i \in [m]$, select the $n \cdot \sum_{j\in[m]:j\neq i} (\hat{Q}_{\hat{y}=i,y^*=j})$ examples with lowest self-confidence $\hat{p}(\tilde{y}=i; x \in X_i)$.
  2. Prune by Noise Rate. For each off-diagnoal entry in $\hat{Q}_{\hat{y}=i,y^*=j}$, $i \neq j$, select $n \cdot \hat{Q}_{\tilde{y}=i,y^*=j}$ examples $x \in X_{\tilde{y}=i}$ with max margin $\hat{p}_{x,\tilde{y}=j} - \hat{p}_{x,\tilde{y}=i}$.
  3. PBC+PBNR. Prune an example if both methods PBS and PBNR prune that example.

6. Confident Lerning: $C_{confusion}$

北京大学 PEKING UNIVERSITY

	y*=0	y*=1	y~
0	0.9	0.1	0
1	0.9	0.1	0
2	0.5	0.5	0
3	0.3	0.7	0
4	0.3	0.7	0
5	0.2	0.9	1
6	0.2	0.8	1
7	0.4	0.7	1
8	0.5	0.5	1
9	0.6	0.4	1
10	0.9	0.1	0
11	0.9	0.1	0
12	0.5	0.5	0
13	0.3	0.7	0
14	0.3	0.7	0
15	0.2	0.9	1
16	0.2	0.8	1
17	0.4	0.7	1
18	0.5	0.5	1
19	0.6	0.4	1

$t_0 = 0.58, t_1 = 0.66$

Estimate label errors as the Boolean vector $\tilde{y}_k \neq \arg\max_{j \in [m]} \hat{p}(\tilde{y} = j; x_k, \theta)$, for all $x_k \in X$, where true imples label error and flase implies clean data.

6. Confident Lerning: $C_{\tilde{y},y^*}$

	y*=0	y*=1	y~
0	0.9	0.1	0
1	0.9	0.1	0
2	0.5	0.5	0
3	0.3	0.7	0
4	0.3	0.7	0
5	0.2	0.9	1
6	0.2	0.8	1
7	0.4	0.7	1
8	0.5	0.5	1
9	0.6	0.4	1
10	0.9	0.1	0
11	0.9	0.1	0
12	0.5	0.5	0
13	0.3	0.7	0
14	0.3	0.7	0
15	0.2	0.9	1
16	0.2	0.8	1
17	0.4	0.7	1
18	0.5	0.5	1
19	0.6	0.4	1

$t_0 = 0.58, t_1 = 0.66$

Estimate label errors as $\{x \in \hat{X}_{\hat{y}=i,y^*=j}: i \neq j\}$ from the diagnoals of $C_{\tilde{y},y^*}$.

Keep the hard examples (near the threshold).

$C_{\tilde{y},y^*}$

	true_0	true_1
pred_0	4	4
pred_1	2	6

6. Confident Lerning: Prune by Class

PEKING UNIVERSITY

	y*=0	y*=1	y~
0	0.9	0.1	0
1	0.9	0.1	0
2	0.5	0.5	0
3	0.3	0.7	0
4	0.3	0.7	0
5	0.2	0.9	1
6	0.2	0.8	1
7	0.4	0.7	1
8	0.5	0.5	1
9	0.6	0.4	1
10	0.9	0.1	0
11	0.9	0.1	0
12	0.5	0.5	0
13	0.3	0.7	0
14	0.3	0.7	0
15	0.2	0.9	1
16	0.2	0.8	1
17	0.4	0.7	1
18	0.5	0.5	1
19	0.6	0.4	1

$t_0 = 0.58, t_1 = 0.66$

For each class $i \in [m]$, select the $n \cdot \sum_{j\in[m]:j\neq i} \left(\hat{Q}_{\tilde{y}=i,y^*=j}\right)$ examples with lowest self-confidence $\hat{p}(\tilde{y} = i; x \in X_i)$.

$\hat{Q}_{\tilde{y},y^*}$

	true_0	true_1
pred_0	0.250	0.250
pred_1	0.125	0.375

$$n \cdot \sum_{j\in[m]:j\neq i} \left(\hat{Q}_{\tilde{y}=i,y^*=j}\right) = 20 * 0.25 = 5, i = 0$$

$$n \cdot \sum_{j\in[m]:j\neq i} \left(\hat{Q}_{\tilde{y}=i,y^*=j}\right) = 20 * 0.125 = 2.5, i = 1$$

6. Confident Lerning: Prune by Noise Rate

北京大学 PEKING UNIVERSITY

	y*=0	y*=1	y~
0	0.9	0.1	0
1	0.9	0.1	0
2	0.5	0.5	0
3	0.3	0.7	0
4	0.3	0.7	0
5	0.2	0.9	1
6	0.2	0.8	1
7	0.4	0.7	1
8	0.5	0.5	1
9	0.6	0.4	1
10	0.9	0.1	0
11	0.9	0.1	0
12	0.5	0.5	0
13	0.3	0.7	0
14	0.3	0.7	0
15	0.2	0.9	1
16	0.2	0.8	1
17	0.4	0.7	1
18	0.5	0.5	1
19	0.6	0.4	1

$t_0 = 0.58, t_1 = 0.66$

For each off-diagnoal entry in $\hat{Q}_{\tilde{y}=i,y^*=j}$, $i \neq j$, select $n \cdot \hat{Q}_{\tilde{y}=i,y^*=j}$ examples $x \in X_{\tilde{y}=i}$ with max margin $\hat{p}_{x,\tilde{y}=j} - \hat{p}_{x,\tilde{y}=i}$.

$\hat{Q}_{\tilde{y},y^*}$

	true_0	true_1
pred_0	0.250	0.250
pred_1	0.125	0.375

$$n \cdot \sum_{j\in[m]:j\neq i} \left(\hat{Q}_{\tilde{y}=i,y^*=j}\right) = 20 * 0.25 = 5, i = 0$$

$$n \cdot \sum_{j\in[m]:j\neq i} \left(\hat{Q}_{\tilde{y}=i,y^*=j}\right) = 20 * 0.125 = 2.5, i = 1$$

Which CL method to use? Five methods are presented to clean data. By default we use CL: $C_{\tilde{y},y^*}$ because it matches the conditions of Thm. 2 exactly and is experimentally performant (see Table 4). Once label errors are found, we observe ordering label errors by the normalized margin: $\hat{p}(\tilde{y}=i; \boldsymbol{x}, \boldsymbol{\theta}) - \max_{j\neq i} \hat{p}(\tilde{y}=j; \boldsymbol{x}, \boldsymbol{\theta})$ (Wei et al., 2018) works well.

7. What is Label Noise: A Neuron Network Complexity Perspective

北京大學
PEKING UNIVERSITY

Understanding Black-box Predictions via Influence Functions

Pang Wei Koh [1] Percy Liang [1]

[1] Koh P W, Liang P. Understanding black-box predictions via influence functions[C]//International conference on machine learning. PMLR, 2017: 1885-1894.

7. What is Label Noise: A Neuron Network Complexity Perspective

北京大学 PEKING UNIVERSITY

1. **What is the result of adding or removing an instance from training dataset?** Given n training examples $z_1 \cdots z_n$, where $z_i = (x_i, y_i)$. Let $L(z, \theta)$ is the loss function. Then empirical risk is $\frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta)$. By ERM, $\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta)$.
2. **Change the weight of one instance.**
   1. $\hat{\theta}_{\epsilon,z} = \underset{\theta}{argmax} \frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta) + \epsilon L(z, \theta)$
   2. **Influence function:** $\mathcal{I}_{up,params}(z) = \frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon}\Big|_{\epsilon=0} = -H_{\hat{\theta}}^{-1} \nabla_\theta L(z, \hat{\theta})$, where $H_{\hat{\theta}} = \frac{1}{n}\Sigma_{i=1}^{n} \nabla_\theta^2 L(z_i, \hat{\theta})$ is the Hessien matrix for empirical risk and is assumed to be positive definite.
   3. **Chain Rule:** The effect of changing the weights of a particular training sample on the test sample loss

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

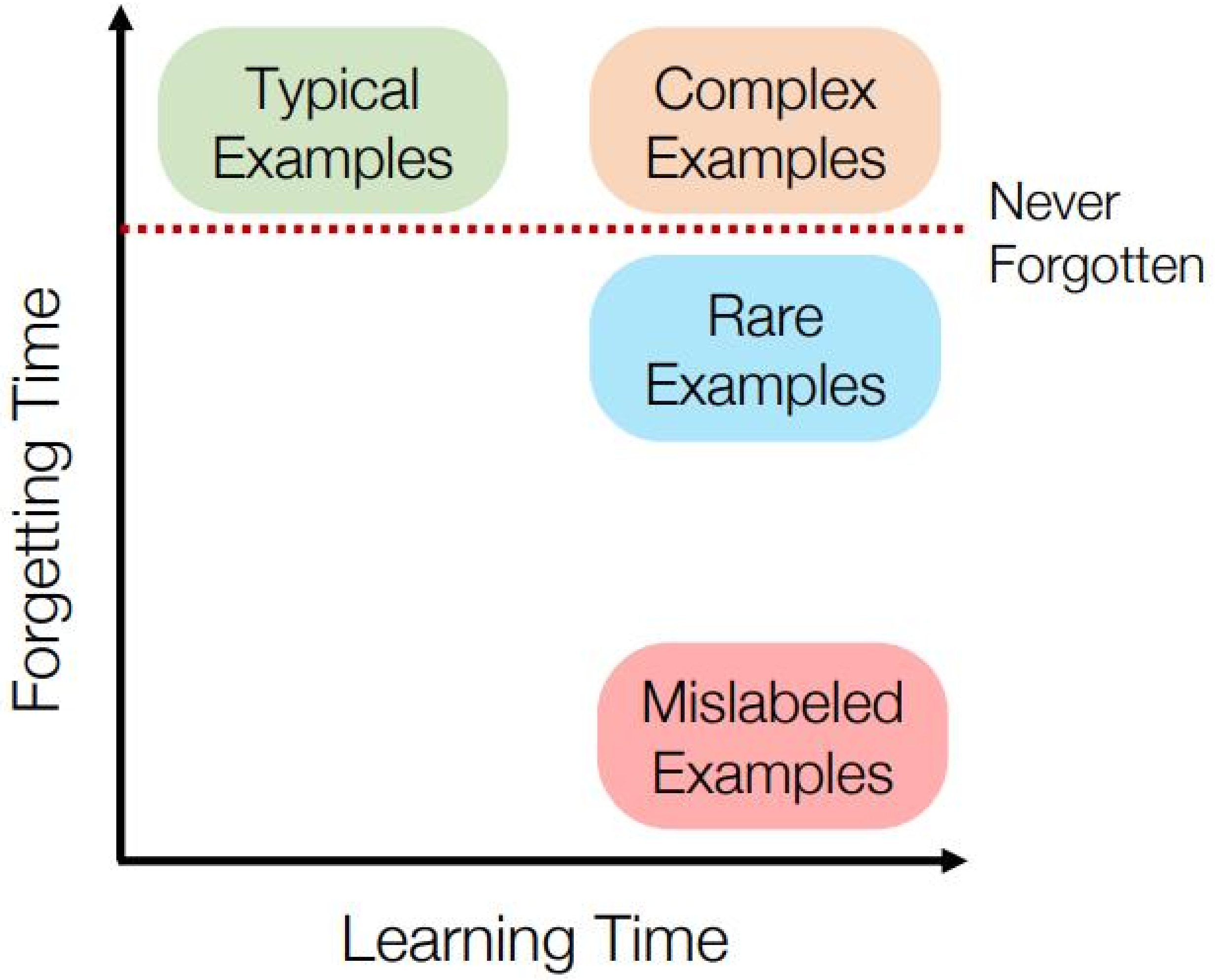
北京大学 PEKING UNIVERSITY

1. **Relation to the Euclidean distance.** To calculate the closeness of the relationship between a test sample and a training sample, one way is to directly find the Euclidean distance between the samples, the smaller the distance, the closer the relationship. But now, the **influence function** can be used instead of the Euclidean distance.
   - **Logistic Regression.** Let $p(y|x) = \sigma(y\theta^T x)$, where σ is sigmoid function.

$$\mathcal{I}_{up,loss}(z, z_{test}) = -y_{test} y \cdot \sigma(-y_{test}\theta^T x_{test}) \cdot \sigma(-y\theta^T x) \cdot x_{test}^T H_{\hat{\theta}}^{-1} x$$

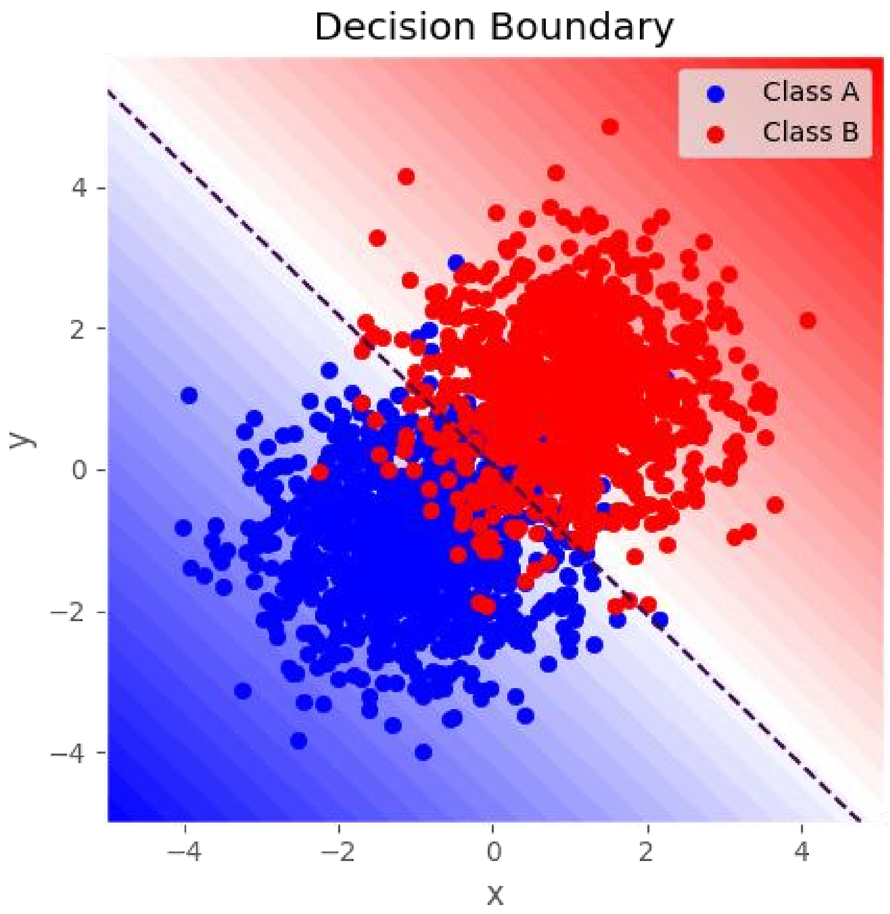
   while Euclidean distance is $x_{test}^T x$.
   - **Differences.**
     - The $\sigma(-y\theta^T x)$ is a weight that relates only to the training samples.
     - The $H_{\hat{\theta}}^{-1}$ reacts to the resistance of all other samples in the training set to $\mathcal{I}_{up,loss}(z, z_{test})$

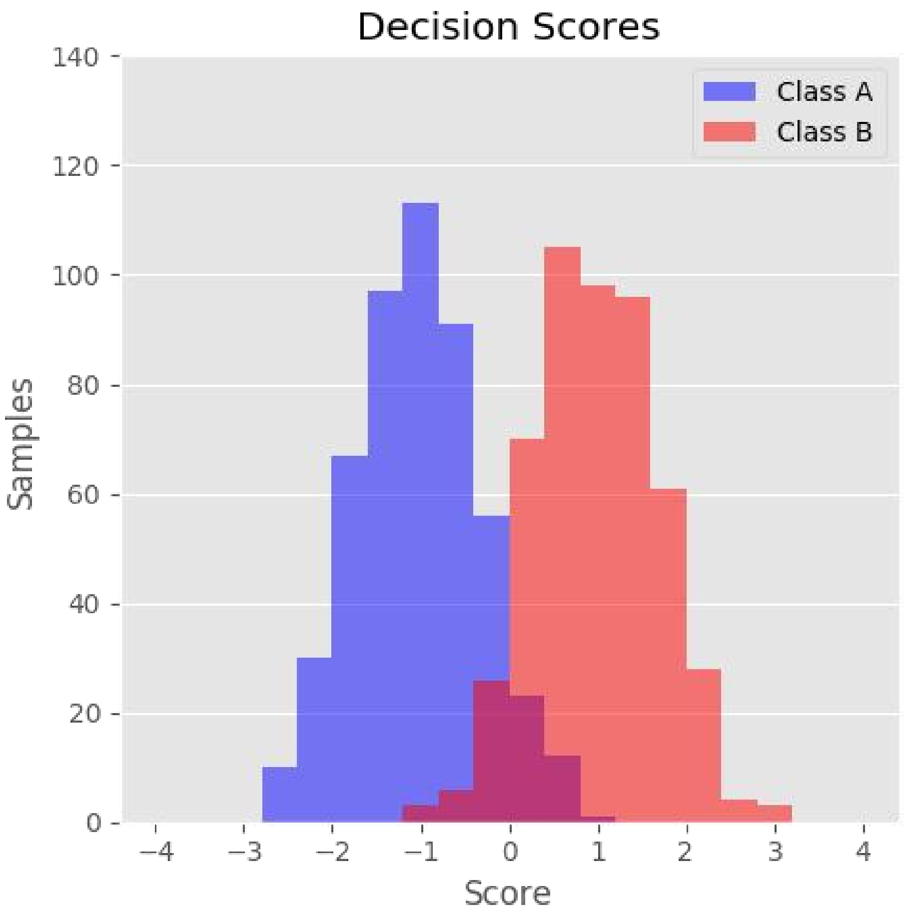
8. What is Label Noise: A Neuron Network Memory Perspective



[1] Maini P, Garg S, Lipton Z, et al. Characterizing datapoints via second-split forgetting[J]. Advances in Neural Information Processing Systems, 2022, 35: 30044-30057.

9. Discussion: Global or Local?





[1] https://www.google.com.hk/url?sa=i&url=http%3A%2F%2Fscikit-hep.org%2Froot_numpy%2Fauto_examples%2Ftmva%2Fplot_twoclass.html&psig=AOvVaw0Gf7tE1UXyPfMZSCd0gSxL&ust=1695757778360000&source=images&cd=vfe&opi=89978449&ved=0CBAQjhxqFwoTCPCu2aHExoEDFQAAAAAdAAAAABAD

9. Discussion: DL or non-DL?

1. **Large Scale Dataset.** Hard to train and inference on whole dataset. For natrual language dataset, measuring data quality using metadata, data sources, visits seems to be an economical choice.
2. **Synthetic Dataset.** Small-scale real data and large-scale uncertain synthetic data. Long tail or mislabeled data?
3. **Data cleansing strategies that vary with dataset size.**

Q&A