THE UNIVERSITY OF SYDNEY

Advanced Machine Learning

(COMP 5328)

Learning with Noisy Data II: Label Noise

Tongliang Liu

Announcements

- Assignment 2 is online now
  - Assignment 2 due on 03/11/2022, 11:59pm
  - Group-based (2-3 students per group). Find you teammates by yourselves.

Assignment 2

THE UNIVERSITY OF SYDNEY

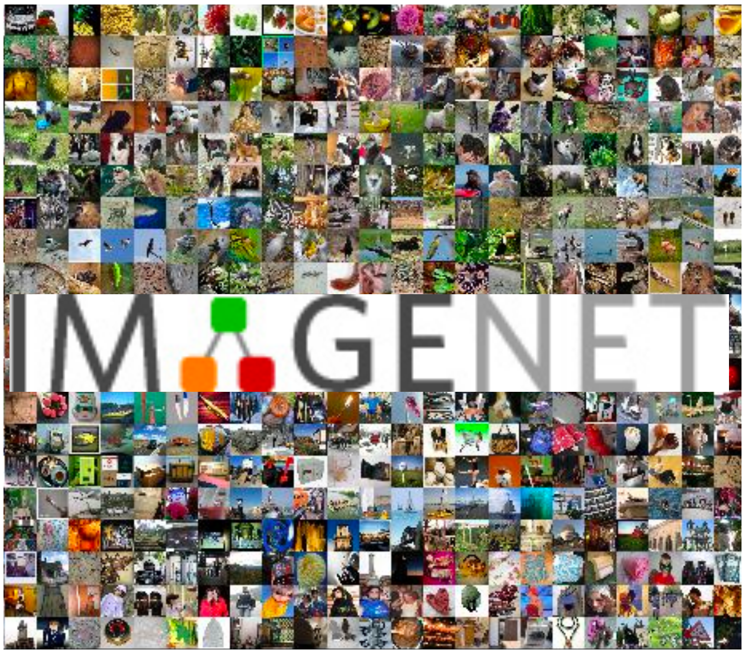
Summary

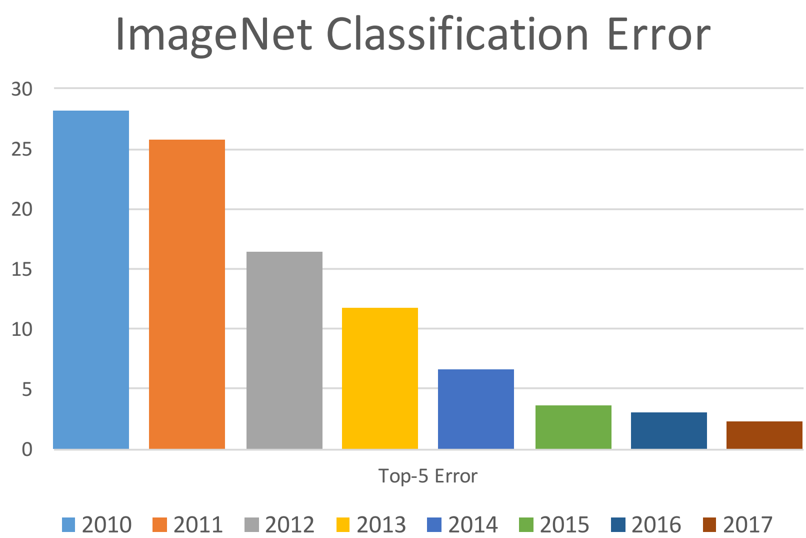
The objective of this assignment is to design algorithms that are robust to label noise. Three input datasets are given. For each dataset, the training and validation data contains class-conditional random label noise, whereas the test data is clean. You need to build at least two different classifiers trained and validated on the noisy data, that can have a good classification accuracy on the clean test data.

The big data era

https://www.zdnet.com/article/big-data-2018-cloud-storage-becomes-the-de-facto-data-lake/

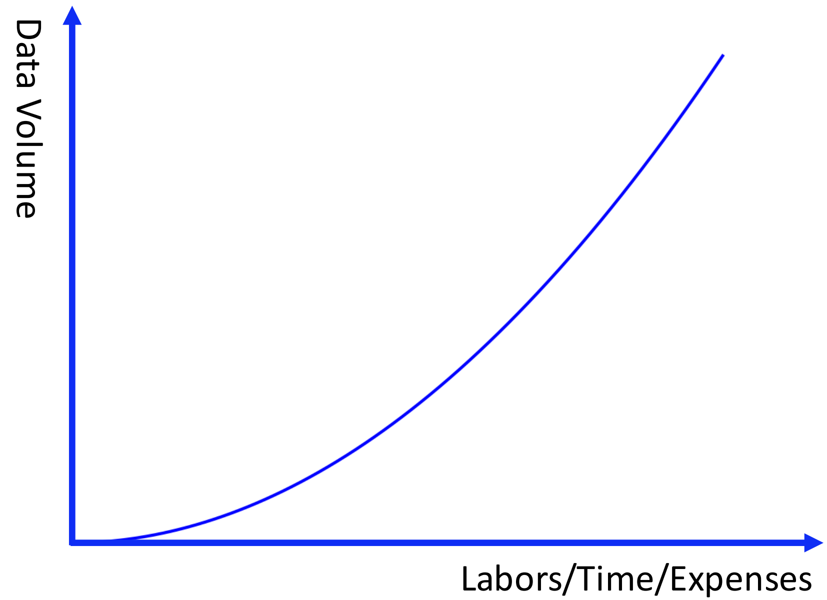
The big data era





Data credit: http://www.image-net.org/

The big data era



Data credit: http://www.image-net.org/

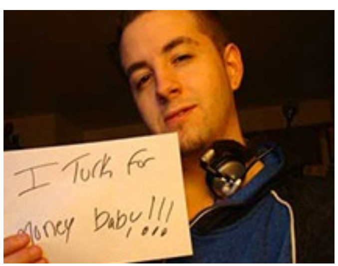
Labelling Costs

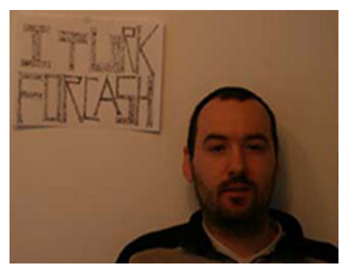
Labelling Costs

We pay for employees for labelling

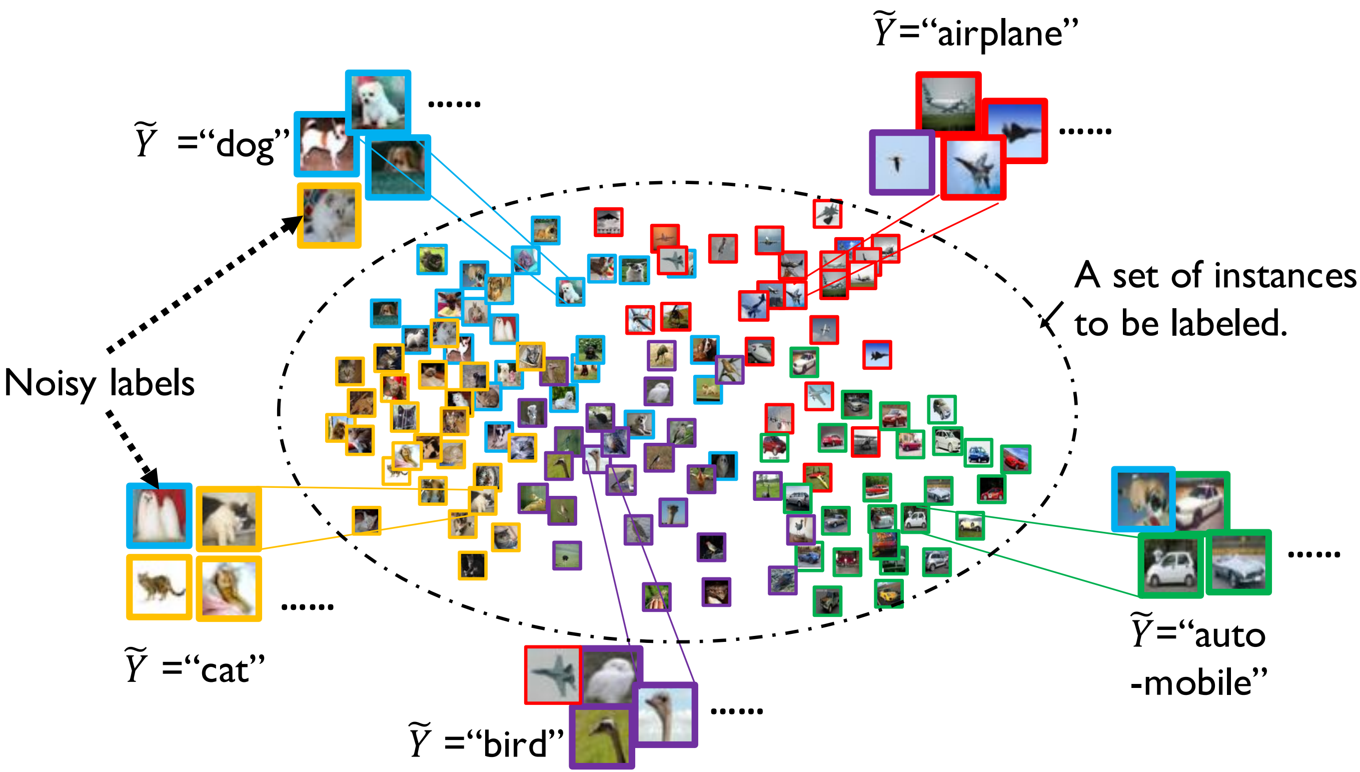
amazon mechanical turk

beta





What is label noise?



What is label noise?

Label noise widely exists even for small data because:

- Labels are provided by non-expert labellers, such as those in the Amazon Mechanical Turk.
- The labelling task is subjective, especially in tasks such as image captioning.
- Insufficient discriminative information for assigning reliable labels. Collecting reliable labels is time-consuming and costly.

Label noise widely exists

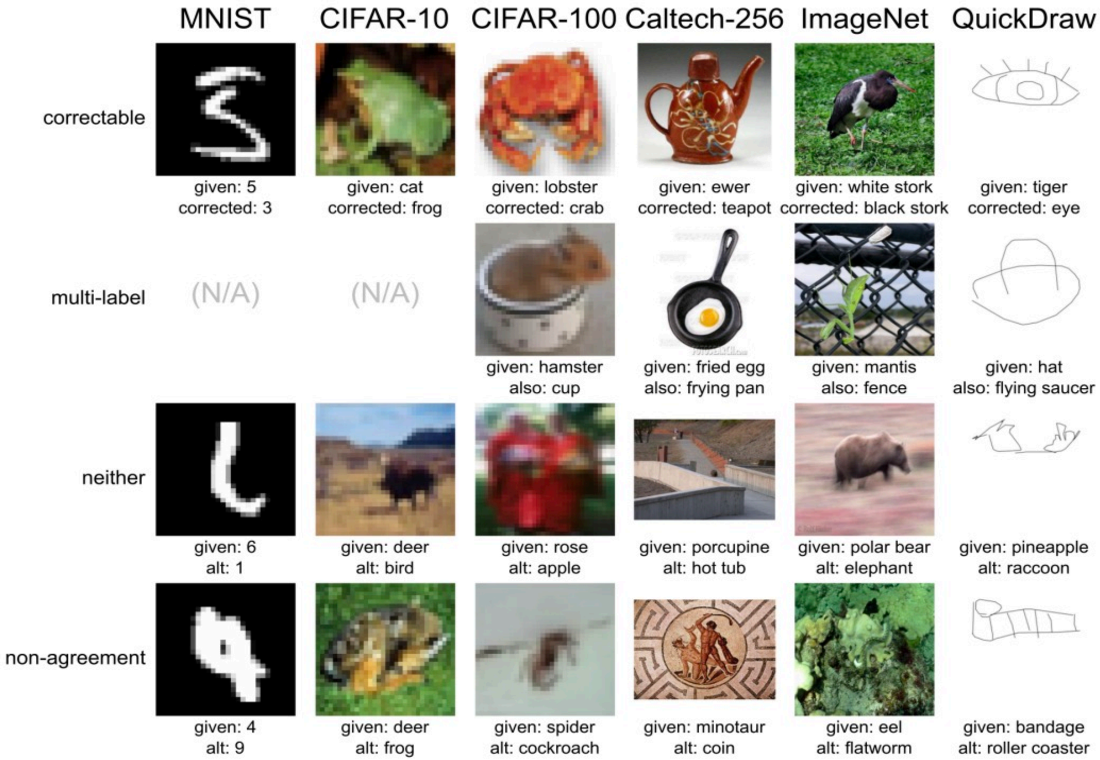


Image credit: Northcutt et al. "Pervasive label errors in test sets destabilize machine learning benchmarks." arXiv preprint arXiv:2103.14749 (2021).

Real-world problems

- ImageNet dataset has noisy labels.

- The WebVision database and the extremely large scale JFT-300M database have about 20% of images which are believed to be incorrectly labelled.

[1] Li, W., Wang, L., Li, W., Agustsson, E., Berent, J., Gupta, A., Sukthankar, R. and Van Gool, L., 2017. WebVision challenge: visual learning and understanding with web data. arXiv preprint arXiv:1705.05640.
[2] Sun, C., Shrivastava, A., Singh, S. and Gupta, A., 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. arXiv preprint arXiv:1707.02968.

Classification

- Expected 0-1 Risk:

clean data.

$$R_D(f) = P(\text{sign}(f(X)) \neq Y) = \mathbb{E}_{(X,Y)\sim D}[\mathbf{1}(\text{sign}(f(X)) \neq Y)].$$

- Expected L-Risk:

$$R_{D,L}(f) = \mathbb{E}_{(X,Y)\sim D}[L(f(X),Y)].$$

- Empirical risk:

no L means 0-1 loss, L means surrogate loss

loss function

$$R_{D,L,n}(f) = \frac{1}{n}\sum_{i=1}^{n} L(f(X_i),Y_i)\,.$$

distribution

#sample

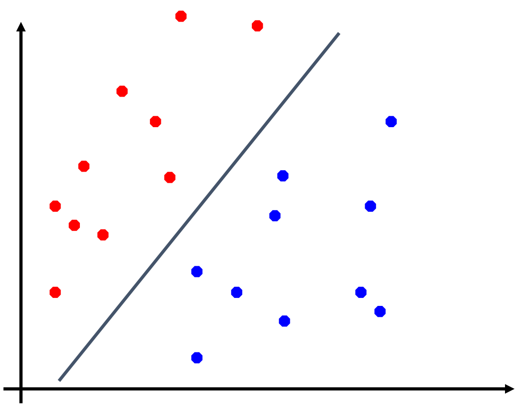
Learning with noisy labels

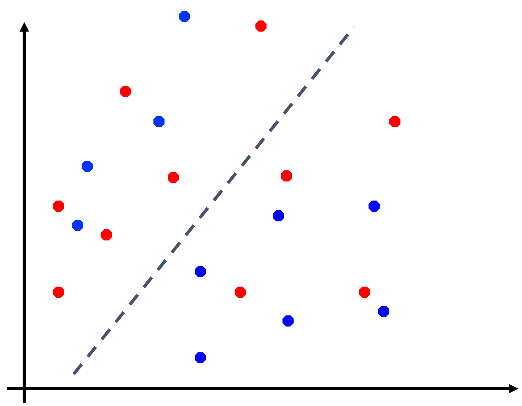
labels are noised ≠ labels are incorrect

$$Y \xrightarrow{\text{noise added}} \tilde{Y}$$

some of them are correct
some of them are incorrect
don't know which are incorrect

$$S=(X_1, Y_1), \dots, (X_n, Y_n) \xrightarrow{\text{label noise}} (X_1, \tilde{Y}_1), \dots, (X_n, \tilde{Y}_n)$$





Learning with noisy labels

Problem Setup

- Observation: $X \in \mathcal{X} \subset \mathbb{R}^d$.
- Clean but unobservable label: $Y \in \mathcal{Y} = \{-1, +1\}$.
- Observable but noisy label: $\tilde{Y} \in \mathcal{Y}$.
- Clean distribution: $D(X, Y)$; Noisy distribution: $D_\rho(X, \tilde{Y})$.

Learning with noisy labels

Problem Setup

- Given the training examples $\{(X_i, \tilde{Y}_i)\}_{1 \leq i \leq n} \sim D_\rho(X, \tilde{Y})^n$.

- The target is to learn a discriminant function $f_n \colon \mathcal{X} \to \mathbb{R}$ such that the classifier predicts the correct label y given an observation $\boldsymbol{x}$.

but also best classifier

Model Label Noise

A probabilistic model:

$$\rho_Y(X) = P(\tilde{Y}|Y, X),$$

where X is the feature, Y is the unobservable true label, and $\tilde{Y}$ is the observed noisy label.

true label -1 but flip into 1

$$\rho_{+1}(X) = P(\tilde{Y} = -1|Y = 1, X); \rho_{-1}(X) = P(\tilde{Y} = 1|Y = -1, X).$$

observe
noise label 1

Note that if there is no label noise, we have

$$P(\tilde{Y} = 1|Y = 1, X) = P(\tilde{Y} = -1|Y = -1, X) = 1$$

otherwise

$$P(\tilde{Y} = 1|Y = -1, X), P(\tilde{Y} = -1|Y = 1, X) \in (0,1).$$

Model Label Noise

(1) Random Classification Noise (RCN):

label noise

assume ② flipping probability is independent of Y

$$\rho_Y(X) = P(\tilde{Y}|Y,X) = P(\tilde{Y}|Y);\ \rho_{+1}(X) = \rho_{-1}(X) = \rho.$$

assume ① flipping probability is independent of X

$$\rho_{+1}(X) = P(\tilde{Y} = -1|Y = 1, X);\ \rho_{-1}(X) = P(\tilde{Y} = 1|Y = -1, X).$$

(2) Class-Dependent Noise (CCN):

$$\rho_Y(X) = P(\tilde{Y}|Y,X) = P(\tilde{Y}|Y);\ \rho_{+1}(X) = \rho_{+1},\ \rho_{-1}(X) = \rho_{-1}.$$

flip from +1 to -1

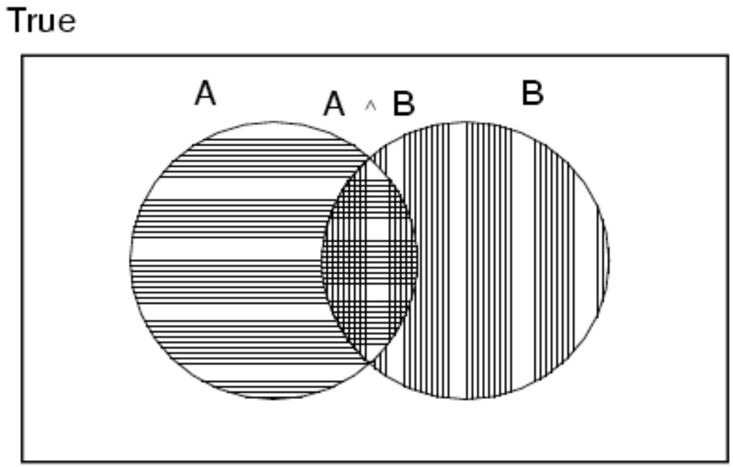
flip from -1 to +1

(3) Instance- and Label-Dependent Noise (ILN):

$$\rho_Y(X) = P(\tilde{Y}|Y,X).$$

Basics in Probability Theory

Axioms of probability



$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cap B) \equiv P(AB) \equiv P(A, B)$$

$$P(A^c) = 1 - P(A)$$

Rules of Probability

- Sum Rule $$p(X) = \sum_Y p(X, Y)$$
- Product Rule $$p(X, Y) = p(Y|X)p(X)$$

$= p(X|Y)p(Y)$

Random Classification Noise

Random Classification Noise (RCN)

The Impact of Label Noise

loss function

Under RCN, minimisation of *any* **convex potential** over a **linear function class** can result in classification performance equivalent to **random guessing**.

loss function

Convex potential: any loss function $\ell: (h(X), Y) \mapsto \phi(Yh(X))$ where $\phi: \mathbb{R} \to \mathbb{R}$ is convex, non-increasing, differentiable with $\phi'(0) < 0$, and $\phi(+\infty) = 0$.

linear function class: $\mathcal{F}_{lin} = \{\boldsymbol{x} \mapsto \boldsymbol{\omega}^{\top}\boldsymbol{x} | \boldsymbol{\omega} \in \mathbb{R}^d\}$.

- input data (with noise)
- pre-defined hypothesis class

Long, Philip M., and Rocco A. Servedio. "Random classification noise defeats all convex potential boosters." Machine learning 78.3 (2010): 287-304.

Random Classification Noise (RCN)

How to Reduce the Effects of RCN?

Symmetric loss function is robust to RCN when the function class $\mathcal{F}_{lin}$ is extended to the universal function space, which means the function in it can be of any form.

Random Classification Noise (RCN)

How to Reduce the Effects of RCN?

convex potential ⇒ symmetric loss

linear function class ⇒ universal function

0-1 loss   sigmoid loss

unhinged loss → only convex & symmetric

Symmetric loss function is robust to RCN when the function class $\mathcal{F}_{lin}$ is extended to the universal function space, which means the function in it can be of any form.

f: hypothesis   f(x) : prediction

Theorem 1. The losses satisfying the following symmetric criterion are robust to RCN:

distance between pred and truth   eg 0-1 loss function

$$L(f(X), +1) + L(f(X), -1) = C,$$

where C is a constant. That is

minimiser with clean data || minimiser with noise data

robust →

$$\arg\min_f R_{D,L}(f) = \arg\min_f R_{D_\rho,L}(f).$$

Van Rooyen, Brendan, Aditya Menon, and Robert C. Williamson. "Learning with symmetric label noise: The importance of being unhinged." Advances in Neural Information Processing Systems. 2015.

Random Classification Noise (RCN)

We will prove that

RCN ρ constant

$$R_{D_\rho,L}(f) = \mathbb{E}_{(X,\tilde{Y})\sim D_\rho}[L(f(X),\tilde{Y})] = (1-2\rho)R_{D,L}(f) + \rho C.$$

f_ρ f

proportional

→ minimiser will be the same

Random Classification Noise (RCN)

Proof

$$P(\tilde{Y} = 1|X)$$

sum rul

$$= P(\tilde{Y} = 1, Y = 1|X) + P(\tilde{Y} = 1, Y = -1|X)$$

product rule

$= 1 - P(\tilde{Y} = -1 | Y = 1, X)$

$$= P(\tilde{Y} = 1|Y = 1, X)P(Y = 1|X) + P(\tilde{Y} = 1|Y = -1, X)P(Y = -1|X)$$

flip prob

$= 1 - P(Y = 1|X).$

$$= (1 - \rho_{+1}(X))P(Y = 1|X) + \rho_{-1}(X)P(Y = -1|X)$$

$$= (1 - \rho_{+1}(X) - \rho_{-1}(X))P(Y = 1|X) + \rho_{-1}(X).$$

Random Classification Noise (RCN)

Proof Cont'd

$$P(\tilde{Y} = -1|X) = (1 - \rho_{+1}(X) - \rho_{-1}(X))P(Y = -1|X) + \rho_{+1}(X)$$

Under RCN

$\rho_{+1}(x) = \rho_{-1}(x) = \rho$

expect risk w.r.t noisy data
= expect risk w.r.t clean data

$$P(\tilde{Y} = 1|X) = (1 - 2\rho)P(Y = 1|X) + \rho$$

$$P(\tilde{Y} = -1|X) = (1 - 2\rho)P(Y = -1|X) + \rho$$

Random Classification Noise (RCN)

Proof Cont'd

$$R_{D_\rho,L}(f) = \mathbb{E}_{(X,\tilde{Y})\sim D_\rho}[L(f(X),\tilde{Y})]$$

Def

$$= \int \left(P(\tilde{Y}=1,X)L(f(X),1) + P(\tilde{Y}=-1,X)L(f(X),-1)\right)dX$$

$$= \int \left(P(\tilde{Y}=1|X)P(X)L(f(X),1) + P(\tilde{Y}=-1|X)P(X)L(f(X),-1)\right)dX$$

$$= \int P(X)[(1-2\rho)P(Y=1|X)L(f(X),1) + \rho L(f(X),1)]dX$$

$$+\int P(X)[(1-2\rho)P(Y=-1|X)L(f(X),-1) + \rho L(f(X),-1)]dX$$

clean data

$$= (1-2\rho)\boxed{\int \left(P(Y=1,X)L(f(X),1) + P(Y=-1,X)L(f(X),-1)\right)dX}$$

$$+\rho\int P(X)\boxed{[\ell(h(X),+1) + \ell(h(X),-1)]}dX$$

if loss function is symmetric func then constant

Random Classification Noise (RCN)

Proof Cont'd

$$R_{D_\rho,L}(f) = \mathbb{E}_{(X,\tilde{Y})\sim D_\rho}[L(f(X),\tilde{Y})]$$

$$= (1-2\rho)\boxed{\int \big(P(Y=1,X)L(f(X),1) + P(Y=-1,X)L(f(X),-1)\big)dX}$$

$$+\rho\int P(X)\boxed{[\ell(h(X),+1) + \ell(h(X),-1)]}dX$$

$$= (1-2\rho)\mathbb{E}_{(X,Y)\sim D}[L(f(X),Y)]+\rho C$$

$$= (1-2\rho)R_{D,L}(f) + \rho C.$$

Completed!

if data size is large
empirical risk
→ close to expected risk

normally $\rho < 0.5$
we believe the majority label are correct.

Random Classification Noise (RCN)

RCN-Robust Losses

The symmetric losses are robust to RCN:

(1) 0-1 Loss: $L(f(X), Y) = \mathbf{1}(\text{sign}(f(X)) \neq Y)$;

(2) Unhinged Loss: $L(f(X), Y) = 1 - Yf(X)$; Note hinge loss $\max\{0, 1-Yf(x)\}$

(3) Sigmoid Loss: $L(f(X), Y) = \frac{1}{1+e^{Yf(X)}}$;

$L(f(x), 1) = \frac{1}{1+e^{f(x)}}$
$L(f(x), -1) = \frac{1}{1+e^{-f(x)}} = \frac{e^{f(x)}}{e^{f(x)}+1}$
$\Sigma = 1$

(4) Ramp Loss: $L(f(X), Y) = \frac{1}{2}\max(0, \min(2, 1 - Yf(X)))$ ...

Ghosh, Aritra, Naresh Manwani, and P. S. Sastry. "Making risk minimization tolerant to label noise." Neurocomputing 160 (2015): 93-107.

Class-dependent Label Noise: Binary

Class-dependent Label Noise

(2) Class-Dependent Noise (CCN):

$$\rho_Y(X) = P(\tilde{Y}|Y,X) = P(\tilde{Y}|Y);\ \rho_{+1}(X) = \rho_{+1},\ \rho_{-1}(X) = \rho_{-1}.$$

Class-dependent Label Noise

Modifying the loss function L to $\tilde{L}$ such that

$$\arg\min_{f\in\mathcal{F}} R_{D,L}(f) = \arg\min_{f\in\mathcal{F}} R_{D_\rho,\tilde{L}}(f).$$

Methods: **Importance reweighting**, unbiased estimator, cost-sensitive loss, rank pruning……

Liu, Tongliang, and Dacheng Tao. "Classification with noisy labels by importance reweighting." IEEE Transactions on pattern analysis and machine intelligence 38.3 (2016): 447-461.
Natarajan, Nagarajan, et al. "Learning with noisy labels." Advances in neural information processing systems. 2013.

Class-dependent Label Noise

Viewing the noisy data and clean data are sampled from two domains, importance reweighting can be applied.

$$R_{D,L}(f) = \mathbb{E}_{(X,Y)\sim D}[L(f(X),Y)] = \int P_D(X,Y)\, L(f(X),Y)\text{dXdY}$$

$$= \int P_{D_\rho}(X,Y)\frac{P_D(X,Y)}{P_{D_\rho}(X,Y)} L(f(X),Y)\text{dXdY}$$

$$= \mathbb{E}_{(X,Y)\sim D_\rho}[\frac{P_D(X,Y)}{P_{D_\rho}(X,Y)} L(f(X),Y)]$$

$$= \mathbb{E}_{(X,Y)\sim D_\rho}[\beta(X,Y)L(f(X),Y)] \quad \text{where } \beta(\boldsymbol{x},y) = \frac{P_D(X=\boldsymbol{x},Y=y)}{P_{D_\rho}(X=\boldsymbol{x},\widetilde{Y}=y)}.$$

Liu, Tongliang, and Dacheng Tao. "Classification with noisy labels by importance reweighting." IEEE Transactions on pattern analysis and machine intelligence 38.3 (2016): 447-461.

Class-dependent Label Noise

Viewing the noisy data and clean data are sampled from two domains, importance reweighting can be applied.

Recall that

$$P_{D_\rho}(\tilde{Y} = y | X = \boldsymbol{x}) = (1 - \rho_{+1} - \rho_{-1}) P_D(Y = y | X = \boldsymbol{x}) + \rho_{-y}$$

Then

instances are clean

√ $$\beta(\boldsymbol{x}, y) = \frac{P_D(X=\boldsymbol{x}, Y=y)}{P_{D_\rho}(X=\boldsymbol{x}, \tilde{Y}=y)} = \frac{P_D(Y=y|X=\boldsymbol{x})}{P_{D_\rho}(\tilde{Y}=y|X=\boldsymbol{x})} = \frac{P_{D_\rho}(\tilde{Y}=y|X=\boldsymbol{x}) - \rho_{-y}}{(1-\rho_{+1}-\rho_{-1}) P_{D_\rho}(\tilde{Y}=y|X=\boldsymbol{x})}.$$

Liu, Tongliang, and Dacheng Tao. "Classification with noisy labels by importance reweighting." IEEE Transactions on pattern analysis and machine intelligence 38.3 (2016): 447-461.

Class-dependent Label Noise

Noise rate estimation

Recall that

equal when $P(Y=-1 | X = x_{+1}) = 0$

the instance is alway +1

$$P(\tilde{Y} = -1 | X = \boldsymbol{x}_{+1}) = (1 - \rho_{+1} - \rho_{-1}) P(Y = -1 | X = \boldsymbol{x}_{+1}) + \rho_{+1}$$

$$P(\tilde{Y} = +1 | X = \boldsymbol{x}_{-1}) = (1 - \rho_{+1} - \rho_{-1}) P(Y = +1 | X = \boldsymbol{x}_{-1}) + \rho_{-1}$$

We also assume that the flip rate is small such that

$$\rho_{+1} + \rho_{-1} \leq 1.$$

→ so factor $(1 - \rho_{+1} - \rho_{-1}) > 0$.

We have

$$P(\tilde{Y} = -1 | X = \boldsymbol{x}) \geq \rho_{+1}$$

$$P(\tilde{Y} = +1 | X = \boldsymbol{x}) \geq \rho_{-1}$$

Liu, Tongliang, and Dacheng Tao. "Classification with noisy labels by importance reweighting." IEEE Transactions on pattern analysis and machine intelligence 38.3 (2016): 447-461.

Class-dependent Label Noise

Noise rate estimation

To estimate flip rate, find instances ⇒ assume specific instance exists, then

$$P(\tilde{Y}=-1|X=\boldsymbol{x}_{+1}) = (1-\rho_{+1}-\rho_{-1})\boxed{P(Y=-1|X=\boldsymbol{x}_{+1})}+\rho_{+1}$$

$$P(\tilde{Y}=+1|X=\boldsymbol{x}_{-1}) = (1-\rho_{+1}-\rho_{-1})\boxed{P(Y=+1|X=\boldsymbol{x}_{-1})}+\rho_{-1}$$

0

$$P(\tilde{Y}=-1|X=\boldsymbol{x}_{+1}) = \rho_{+1} \qquad P(\tilde{Y}=-1|X=\boldsymbol{x}) \geq \rho_{+1}$$

$$P(\tilde{Y}=+1|X=\boldsymbol{x}_{-1}) = \rho_{-1} \qquad P(\tilde{Y}=+1|X=\boldsymbol{x}) \geq \rho_{-1}$$

We designed the following estimator: $\rho_{-y} = \min_{X\in\mathcal{X}} P(\tilde{Y}=y|X)$

Liu, Tongliang, and Dacheng Tao. "Classification with noisy labels by importance reweighting." IEEE Transactions on pattern analysis and machine intelligence 38.3 (2016): 447-461.

Class-dependent Label Noise

$$R_{D,L}(f) = \mathbb{E}_{(X,Y)\sim D_\rho}[\beta(X,Y)L(f(X),Y)]$$

where $\beta(\boldsymbol{x}, y) = \frac{P_D(X=\boldsymbol{x}, Y=y)}{P_{D_\rho}(X=\boldsymbol{x}, \tilde{Y}=y)} = \frac{P_{D_\rho}(\tilde{Y}=y|X=\boldsymbol{x}) - \rho_{-y}}{(1-\rho_{+1}-\rho_{-1})P_{D_\rho}(\tilde{Y}=y|X=\boldsymbol{x})}.$

$\approx \frac{1}{n}\sum_{i=1}^{n} \beta(x_i, \tilde{Y}_i) L(f(x_i), \tilde{Y}_i)$

Liu, Tongliang, and Dacheng Tao. "Classification with noisy labels by importance reweighting." IEEE Transactions on pattern analysis and machine intelligence 38.3 (2016): 447-461.

THE UNIVERSITY OF SYDNEY

Class-dependent Label Noise: Multi-class

Transition matrix

We can obtain the following by using the product rule and the sum rule:

Law of total prob

c^2 prob to flip if c classes

$$\begin{bmatrix} P(\tilde{Y}=1|\boldsymbol{x}) \\ \vdots \\ P(\tilde{Y}=C|\boldsymbol{x}) \end{bmatrix} = \boxed{\begin{bmatrix} P(\tilde{Y}=1|Y=1,\boldsymbol{x}) & \cdots & P(\tilde{Y}=1|Y=C,\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ P(\tilde{Y}=C|Y=1,\boldsymbol{x}) & \cdots & P(\tilde{Y}=C|Y=C,\boldsymbol{x}) \end{bmatrix}} \begin{bmatrix} P(Y=1|\boldsymbol{x}) \\ \vdots \\ P(Y=C|\boldsymbol{x}) \end{bmatrix}$$

noisy posterior

clean posterior

to learn

Learning with noisy labels

Let T be the following flip matrix (also called transition matrix), e.g.,

$$T = \begin{bmatrix} P(\tilde{Y}=1|Y=1) & P(\tilde{Y}=1|Y=2) & \dots & P(\tilde{Y}=1|Y=C) \\ P(\tilde{Y}=2|Y=1) & P(\tilde{Y}=2|Y=2) & \dots & P(\tilde{Y}=2|Y=C) \\ \vdots & \vdots & \vdots & \vdots \\ P(\tilde{Y}=C|Y=1) & P(\tilde{Y}=C|Y=2) & \dots & P(\tilde{Y}=C|Y=C) \end{bmatrix}.$$

If we assume that given the clean label, the noisy label is independent with the instance, we have that $P(\tilde{Y}|Y) = P(\tilde{Y}|Y,X)$, and that

Forward

$$[P(\tilde{Y}=1|X), \dots, P(\tilde{Y}=C|X)]^\top = T\,[P(Y=1|X), \dots, P(Y=C|X)]^\top,$$

or $$[P(Y=1|X), \dots, P(Y=C|X)]^\top = T^{-1}[P(\tilde{Y}=1|X), \dots, P(\tilde{Y}=C|X)]^\top.$$

Backward ① inverse is hard ② noise will be amplified

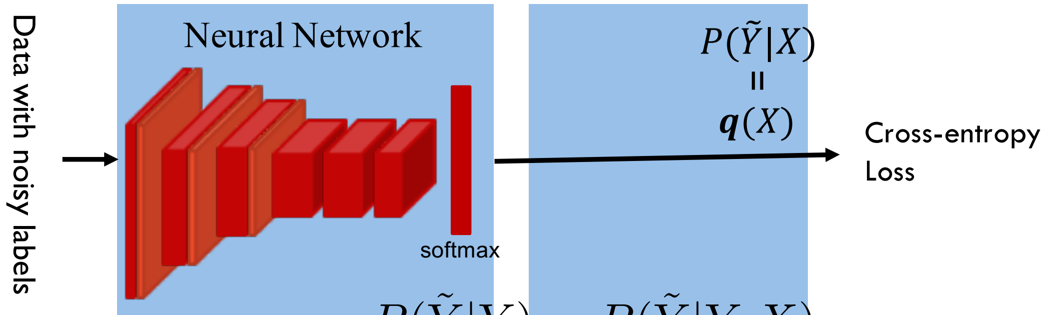
The above means that we can infer the clean class posterior by employing the noisy class posterior and the inverse transition matrix.

Learning with noisy labels

known $(x_i, \tilde{y}_i)_{i=1}^{n}$

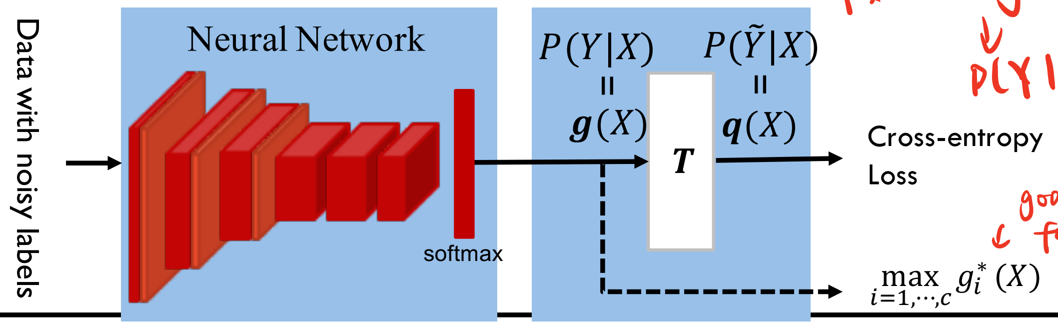
Forward learning:

$$[P(\tilde{Y}=1|X),\ldots,P(\tilde{Y}=C|X)]^\top = T\,[P(Y=1|X),\ldots,P(Y=C|X)]^\top.$$



$$P(\tilde{Y}|Y) = P(\tilde{Y}|Y,X),$$

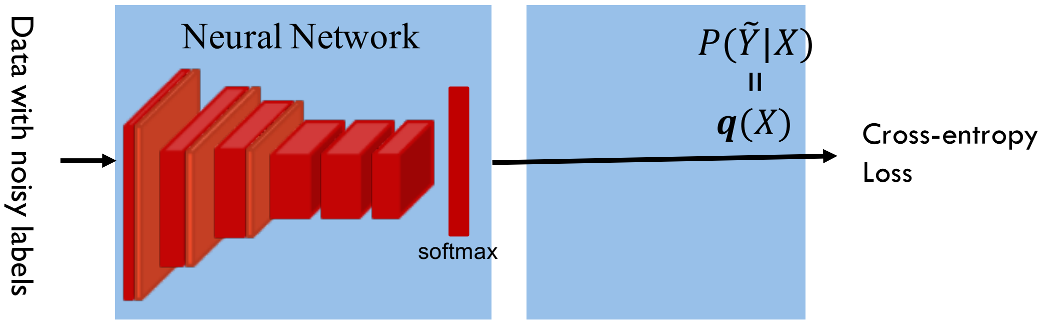
T × something = $P(\tilde{Y}|X)$
↓
$P(Y|X)$



good approximation
for clean
posterior

Learning with noisy labels

Backward learning:



We have that $P(Y|X) = \boldsymbol{g}(X) = T^{-1}\boldsymbol{q}(X)$

because $[P(Y=1|X),\ldots,P(Y=C|X)]^\top = T^{-1}[P(\tilde{Y}=1|X),\ldots,P(\tilde{Y}=C|X)]^\top$.

Instance- and Class-dependent Label Noise

(3) Instance- and Label-Dependent Noise (ILN):

$$\rho_Y(X) = P(\tilde{Y} | Y, X).$$

Relationship: noisy data, flip rates, and clean data

$$P(\tilde{Y} = 1|X)$$

$$= P(\tilde{Y} = 1, Y = 1|X) + P(\tilde{Y} = 1, Y = -1|X)$$

$$= P(\tilde{Y} = 1|Y = 1, X)P(Y = 1|X) + P(\tilde{Y} = 1|Y = -1, X)P(Y = -1|X)$$

$$= (1 - \rho_{+1}(X))P(Y = 1|X) + \rho_{-1}(X)P(Y = -1|X)$$

$$= \left(1 - \rho_{+1}(X) - \rho_{-1}(X)\right)P(Y = 1|X) + \rho_{-1}(X).$$

Instance- and Class-dependent Label Noise

Estimating flip rate for instance- and class-dependent label noise is ill-posed, e.g., if we assume

① noisy posterial large

② label noise rate small

⇓ infer

clean posterior $p(Y=1|X) > 0.5$

$$P(\tilde{Y}=1|X)=0.8,$$

We may have different possible solutions for $P(Y=1|X)$ and $\rho(X)$, e.g.,

flip rate is small

$$P(Y=1|X)=1; \rho_{-1}(X)=\rho_{+1}(X)=0.2.$$

$$P(Y=1|X)=0.875; \rho_{-1}(X)=\rho_{+1}(X)=0.1.$$

Instance- and Class-dependent Label Noise

Estimating flip rate for instance- and class-dependent label noise is ill-posed. → no exact transition matrix

Open problem: Can we make some reasonable assumptions such that the flip rate is identifiable?