Weighted Empirical Risk Minimization

Robin Vogel[1,2] Mastane Achab [2] Stephan Clémençon[2]
Charles Tillier [2]

[1] IDEMIA, [2] Télécom Paris

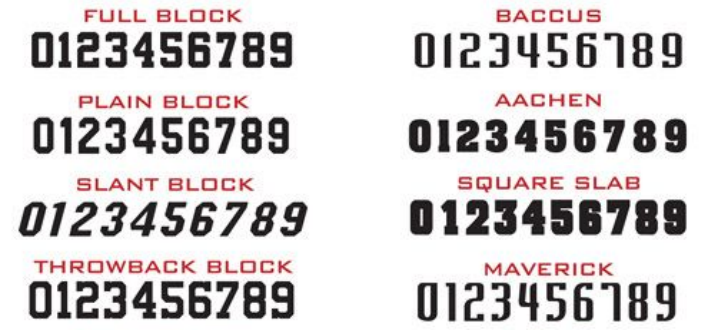
17/08/2020

Outline

Distributional shift

Sometimes, **training data does not match testing data**.

[Ganin et al., 2015]: generated numbers to classify SVHN numbers.



Train & test data contain different ethnicities.
Here: RacialFaces [Wang et al., 2019] and LFW [Huang et al., 2007].

Probabilistic setting

Define $\mathcal{Z}$ as the **input space**.
e.g. in binary classif. $\mathcal{Z} = \mathbb{R}^d \times \{-1, +1\}$

As well as a **distribution** P over $\mathcal{Z}$,
defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

Introduce the *i.i.d.* **sample**:

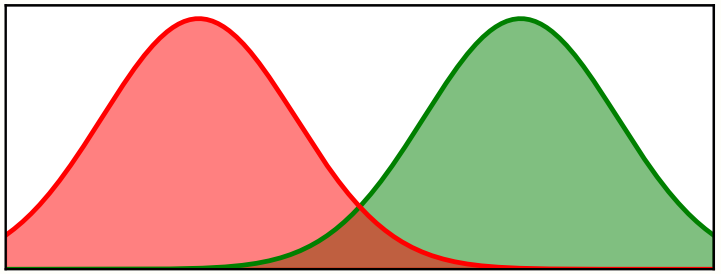
$$\mathcal{D}_n = \{Z_1, \dots, Z_n\} \overset{\text{i.i.d.}}{\sim} P.$$

The **empirical distribution** writes:

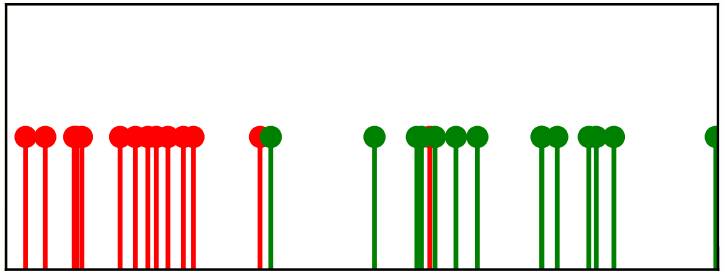
$$\hat{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i},$$

where δ_z is the Dirac distribution
($\delta_z(A) = \mathbb{I}\{z \in A\}$ for any $A \subset \mathcal{Z}$).

Here $\mathcal{Z} = \mathbb{R} \times \{-1, +1\}$,
$P(\mathcal{Z}, +1) = P(\mathcal{Z}, -1)$.



Here $Z_1 = (X_1, Y_1)$
red: *p.d.f.* $X_1 | Y_1 = -1$.
green: *p.d.f.* $X_1 | Y_1 = +1$.



Empirical distribution
$Y_i = -1$, $Y_i = +1$.

Empirical Risk Minimization (ERM)

Usual problems seeks to **minimize an expected risk** in $\theta \in \Theta$:

$$\mathcal{R}_P(\theta) := \mathbb{E}_P[\ell(\theta, Z)], \tag{1}$$

w/ $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}_+$ is a loss function and $\mathbb{E}_P$ the expectation w.r.t. P.

Empirical Risk Minimization (ERM) ([Devroye et al., 1996]) approximates the expected risk with the sample $\mathcal{D}_n$:

$$\widehat{\mathcal{R}}_P(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, Z_i) = \mathcal{R}_{\widehat{P}_n}(\theta). \tag{2}$$

The performance of minimizers of Eq. (2) for Eq. (1) can be guaranteed from usual concentration inequalities.

What if the training (P') and test (P) distributions differ ?

Contributions

Our work proposes:

· A **general probabilistic setting** for problems where train and test distribution differ.

· A concrete technique for **several problems**:
$\rightarrow$ learning with different class/strata probabilities,
$\rightarrow$ PU learning,
$\rightarrow$ predicting with censored samples,
with **auxiliary information** on the relation between train and test.

· Illustrative experiments of its effectiveness on **ImageNet** ([Russakovsky et al., 2014]).

Outline

Weighted Empirical Risk Minimization

Introduce a sample $\mathcal{D}'_n = \{Z'_1, \ldots, Z'_n\} \overset{\text{i.i.d.}}{\sim} P'$ and $w = (w_1, \ldots, w_n)$ a weight vector. Then the **weighted empirical risk** is:

$$\mathcal{R}_{\widetilde{P}_{w,n}}(\theta) := \frac{1}{n} \sum_{i=1}^{n} w_i \cdot \ell(\theta, Z_i),$$

and the empirical weighted distribution is $\widetilde{P}_{w,n} := \frac{1}{n} \sum_{i=1}^{n} w_i \cdot \delta_{Z'_i}$.

If $P \ll P'$, let $\Phi(z) := (dP/dP')(z)$ be the likelihood function.
Then the weights w^* such that $w^*_i = \Phi(Z'_i)$ satisfy:

$$\mathbb{E}_{P'}\left[\mathcal{R}_{\widetilde{P}_{w^*,n}}(\theta)\right] = \mathcal{R}_P(\theta),$$

i.e. the weighted risk is an **unbiased estimator** of the expected risk.
We denote the minimizer of $\mathcal{R}_{\widetilde{P}_{w^*,n}}(\theta)$ by $\widetilde{\theta}^*_n$.

Theoretical guarantees

Assume $\sup_{(\theta,z)\in\Theta\times\mathcal{Z}} \ell(\theta, z) \leq L$ and $\sup_{z\in\mathcal{Z}} \Phi(z) \leq M$.

Theorem 1

With probability at least $1-\delta$*, for any* $n \geq 1$*,*

$$\mathcal{R}_P(\widetilde{\theta}_n^*) - \min_{\theta\in\Theta} \mathcal{R}_P(\theta) \leq 2M \cdot \left(\mathbb{E}[\mathfrak{R}'_n(\mathcal{F})] + L\sqrt{\frac{2\log(1/\delta)}{n}} \right).$$

Note: Under complexity assumptions (*e.g.* VC assumptions), the quantity $\mathbb{E}[\mathfrak{R}'_n(\mathcal{F})]$ is of order $O(n^{-1/2})$.

Problems:

· $\Phi : \mathcal{Z} \to \mathbb{R}_+$ is unknown in general.

· if M is large, the bound is null.

Outline

Different proportion of target classes

Let $\mathcal{Z} = \mathcal{X} \times \{1, \ldots, K\}$, thus $Z_1 = (X_1, Y_1)$ with Y_1 a target class.
Then, with $p_k := P(\mathcal{X}, k)$ and $p'_k := P'(\mathcal{X}, k)$, we have:

$$\Phi(z) = \Phi((x, y)) = \frac{1}{n} \sum_{k=1}^{K} \frac{p_k}{p'_k} \cdot \mathbb{I}\{y = k\}.$$

Φ **only depends** on the p_k's and p'_k's.

The p'_k's can be estimated from the data $Y'_1, \ldots, Y'_n$.
However, we assume the p_k to be known (auxiliary information).

Assume that the p'_k are away from zero, *i.e.* $\max_k p'_k \geq \epsilon$ with $\epsilon > 0$,
then Φ is bounded and we have the **usual learning rate** $O(n^{-1/2})$.

Stratified data

We can reweight on **any discrete attribute** !

$\rightarrow$ We need to know the **proportion of each strata** in the test set.

Other problems

Positive unlabeled (PU) learning

PU learning solves binary classif. with positive and unlabeled data.

Then $\mathcal{Z} = \mathcal{X} \times \{-1, +1\}$, thus $Z_1 = (X_1, Y_1)$.
Set $p := P(\mathcal{X} \times \{-1\})$ and $q := P(\mathcal{X} \times \{+1\})$.

$$\Phi(x, y) = \frac{p}{q}\mathbb{I}\{y = +1\} + \frac{1}{1-q}\mathbb{I}\{y = -1\} - \frac{p}{1-q}\frac{dF_+}{dF}(x) \cdot \mathbb{I}\{y = -1\}.$$

The quantity dF_+/dF is removed by substituting samples.

Learning from censored data

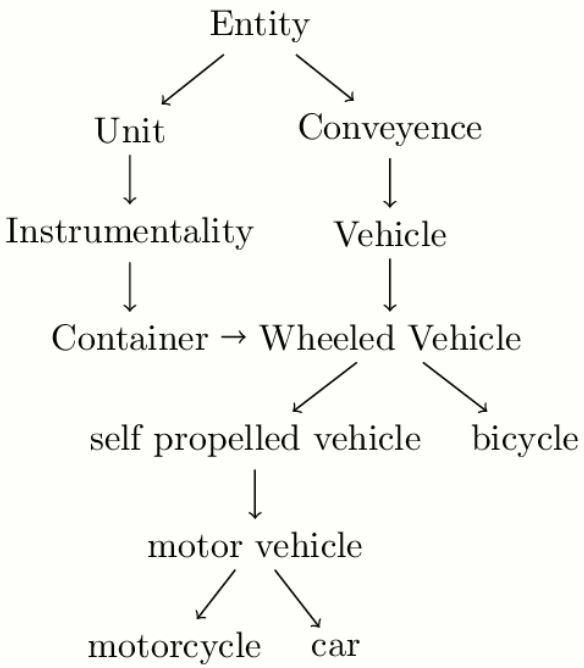
Denote by $(X, \min(Y, C), \mathbb{I}\{Y \leq C\})$ and $(X', \min(Y', C'), \mathbb{I}\{Y' \leq C\})$ the *r.v.* concerned with the distributions of P and P'. Then:

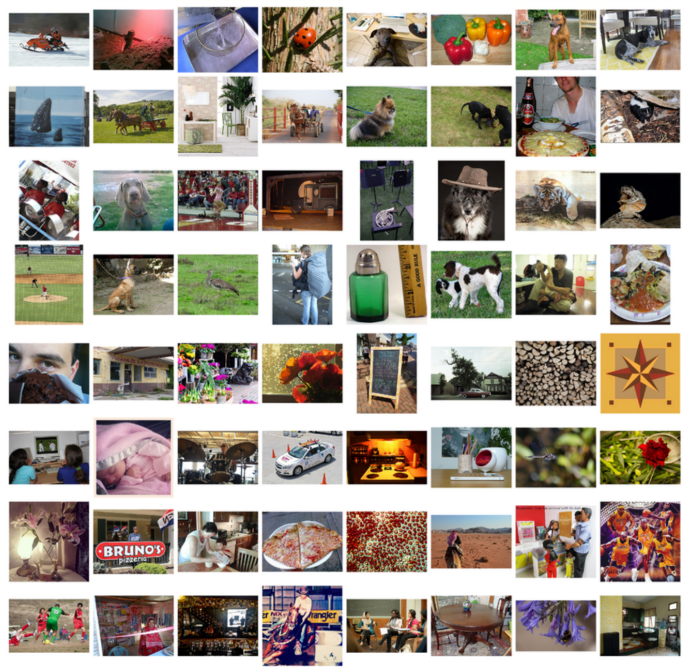
$$\Phi(x, y, \delta) = \frac{\mathbb{I}\{Y \leq C\}}{\mathbb{P}\{C' \geq y | X' = x\}}.$$

Outline

The ImageNet dataset

ImageNet is based on **WordNet** ([Fellbaum, 1998]).
WordNet is a hierarchical database of english nouns (synsets).
We use the ImageNet dataset with **high-level synsets as strata**.





The train and test splits of ImageNet have same strata distribution.
We **induce strata bias by removing data**.

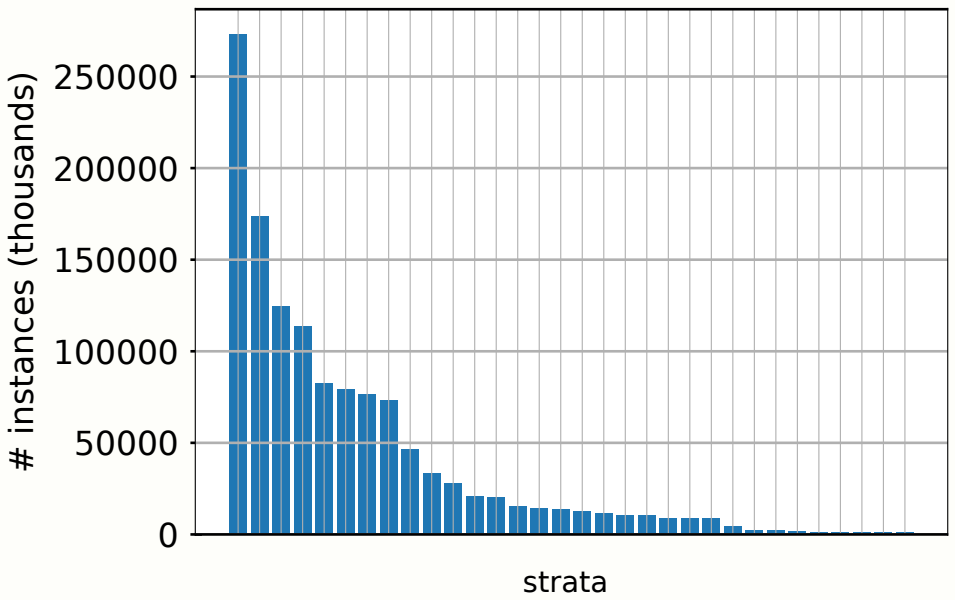
Imbalanced data for ImageNet

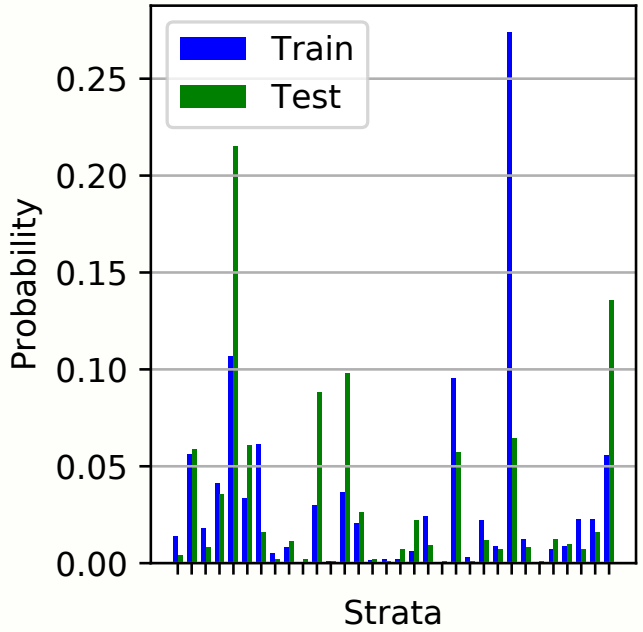
We induce strata bias using a **power law**.

Introducing a strata bias parameter $0 \leq \gamma \leq 1$, we set:

$$p'_k = \gamma^{1-\lfloor K/2 \rfloor / k} \cdot p_k,$$

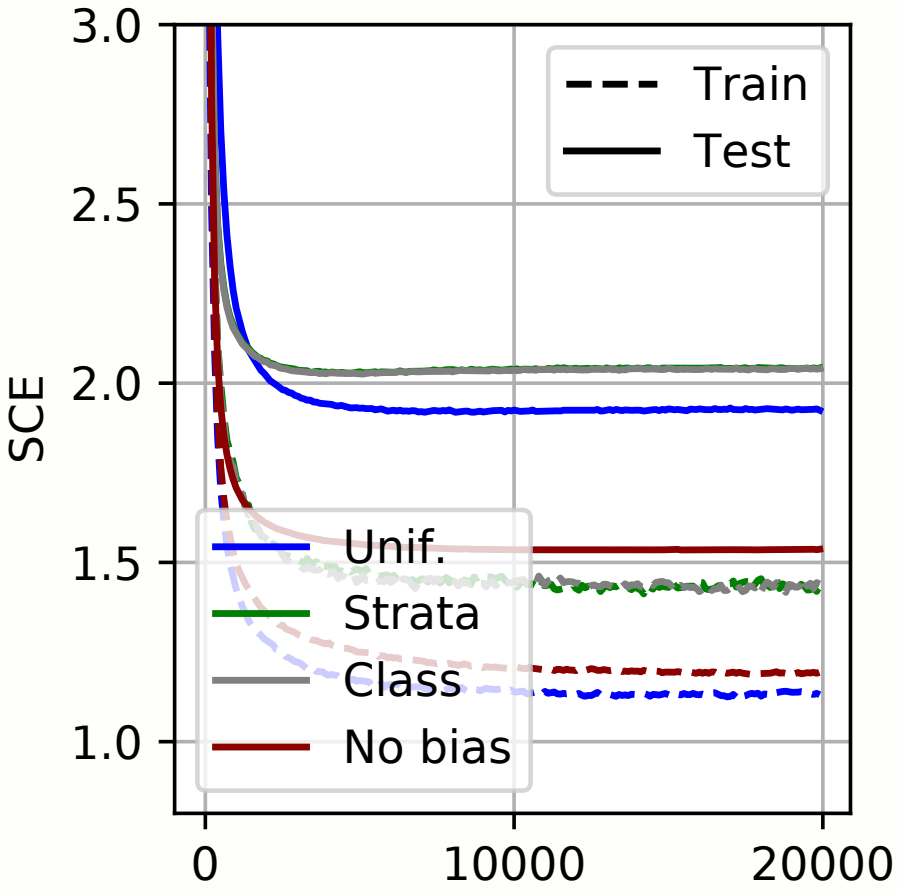
and remove instances of the train set to get right proportions p'_k.

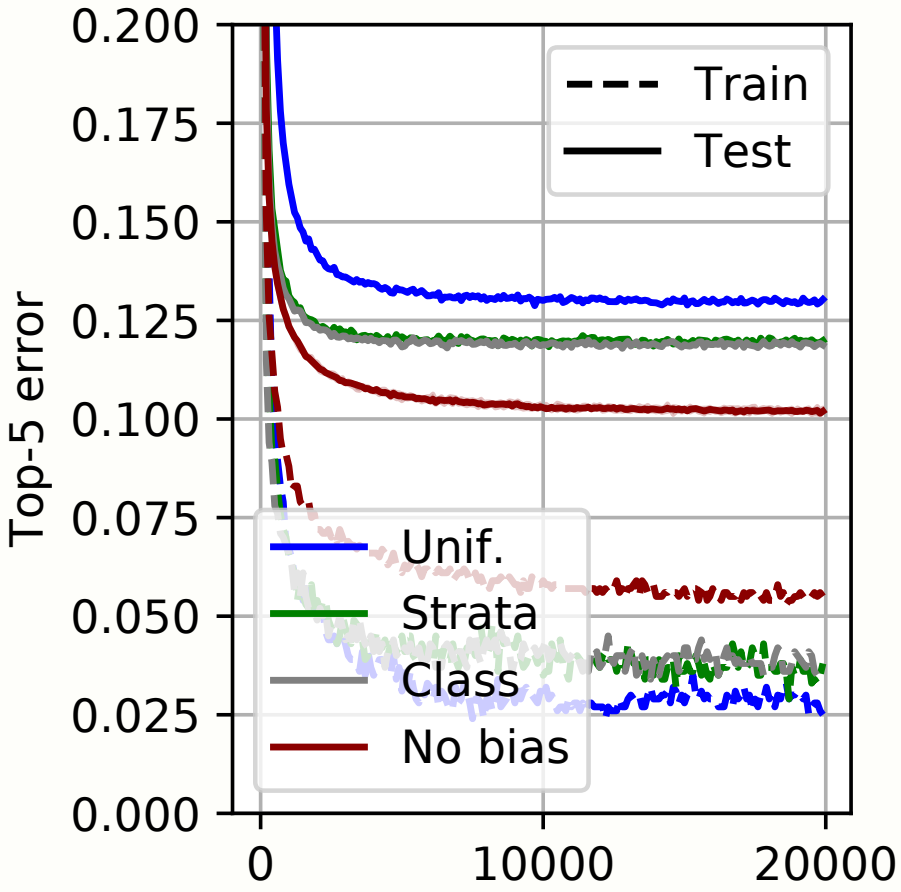




Results

We optimize a softmax cross-entropy (SCE) with ADAM optimizer for a linear model on the last convolutional layer of the ResNet50 network of [He et al., 2015].





Outline

Limitations and future work

Limitations:

· the hypothesis $P \ll P'$.

$\rightarrow$ relaxed in [Laforgue and Clémençon, 2019].

$\rightarrow$ some works account for $\text{dom}(P) \cap \text{dom}(P') = \emptyset$ with geometric interpretations (e.g. optimal transport [Redko et al., 2016]).

· the limited nature of the auxiliary information.

$\rightarrow$ small sample of the target dataset in [Sugiyama et al., 2008].

Future work:

Tackle learning with a small dataset that has the test distribution.

Thank you !

References I


Devroye, L., Györfi, L., and Lugosi, G. (1996).
A Probabilistic Theory of Pattern Recognition.
Springer.


Fellbaum, C., editor (1998).
WordNet: an electronic lexical database.
MIT Press.


Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2015).
Domain-adversarial training of neural networks.


He, K., Zhang, X., Ren, S., and Sun, J. (2015).
Deep residual learning for image recognition.
CoRR, abs/1512.03385.


Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007).
Labeled faces in the wild: A database for studying face recognition in unconstrained environments.
Technical Report 07-49, University of Massachusetts, Amherst.


Laforgue, P. and Clémençon, S. (2019).
Statistical learning from biased training samples.


Redko, I., Habrard, A., and Sebban, M. (2016).
Theoretical analysis of domain adaptation with optimal transport.

References II

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Li, F. (2014).
Imagenet large scale visual recognition challenge.
CoRR, abs/1409.0575.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. v., and Kawanabe, M. (2008).
Direct importance estimation with model selection and its application to covariate shift adaptation.
In *NIPS*, pages 1433–1440.

Wang, M., Deng, W., Hu, J., Tao, X., and Huang, Y. (2019).
Racial faces in the wild: Reducing racial bias by information maximization adaptation network.
In *The IEEE International Conference on Computer Vision (ICCV)*.