On Selection Bias and Fairness Issues in Machine-Learning

Stephan Clémençon

–

LTCI, Telecom Paris

–

with M. Achab (Pompeu Fabra), G. Ausset (Telecom Paris), A. Bellet (INRIA), P. Bertail (Paris X), E. Chautru (Mines Paris), P. Laforgue (Univ. Milano), G. Papa (G-Research), F. Portier (ENSAI), C. Tillié (UVSQ), R. Vogel (Univ. Edinburgh), M. Limnios (EPFL), N. Vayatis (ENS), Y. Guyonvarch (INRAE)

5/23/2024

The Way of Considering Bias & Fairness in AI Here

- Expertise: Statistical Machine Learning, Theory and Algorithms
- Scientific Goals
  - Predictive issues cast as M-estimation problems:
    - Classification
    - Regression
    - Density level set estimation
    - ... and their numerous variants
  - Complex data - Minimal assumptions on the distribution
  - Algorithms: design feasible M-estimators for specific criteria
  - Many Questions - Theory/Computation/Applicability:
    - Theoretical guarantees: optimal elements, consistency, non-asymptotic excess risk bounds, fast rates of convergence, oracle inequalities
    - Practice: numerical optimization, convexification, randomization, relaxation, scalability (distributed architectures, real-time, memory, *etc.*)
    - Applicability/acceptability: robustness/reliability, explainability, privacy preservation, fairness...

Many applications of these concepts/methods e.g. Facial Recognition

The Flagship Problem: Pattern Recognition

- (X, Y) random pair with unknown distribution P
- $X \in \mathcal{X}$ observation vector
- $Y \in \{-1, +1\}$ binary label/class

- *A posteriori* probability $\sim$ regression function

$$\forall x \in \mathcal{X}, \quad \eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$$

- $g : \mathcal{X} \to \{-1, +1\}$ classifier that **can be coded** using a machine
- Performance measure = classification error

$$L(g) = \mathbb{P}\left\{g(X) \neq Y\right\} \quad \to \min_g$$

- Solution: Bayes rule

$$\forall x \in \mathcal{X}, \quad g^*(x) = 2\mathbb{I}\{\eta(x) > 1/2\} - 1$$

- Bayes error $L^* = L(g^*)$

Main Paradigm: Empirical Risk Minimization

- **Training sample** $(X_1, Y_1), \ldots, (X_n, Y_n)$ with i.i.d. copies of (X, Y)
- Class $\mathcal{G}$ of classifiers (massive catalog)
- Empirical Risk Minimization principle

$$\hat{g}_n = \arg\min_{g \in \mathcal{G}} L_n(g) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{g(X_i) \neq Y_i\}$$

- Mimic the best classifier in the class

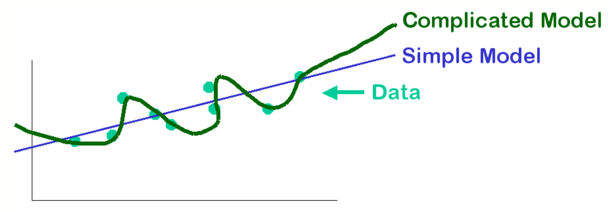
$$\bar{g} = \arg\min_{g \in \mathcal{G}} L(g)$$

Does the ERM principle works?

Predict labels of **past data**

vs

Predict labels of **future data**



Empirical Processes in Statistical Learning

- Bias-variance decomposition

$$\begin{aligned} L(\hat{g}_n) - L^* &\leq (L(\hat{g}_n) - L_n(\hat{g}_n)) + (L_n(\bar{g}) - L(\bar{g})) + (L(\bar{g}) - L^*) \\ &\leq 2\left(\sup_{g\in\mathcal{G}} \mid L_n(g) - L(g) \mid\right) + \left(\inf_{g\in\mathcal{G}} L(g) - L^*\right) \end{aligned}$$

- Concentration inequality

  With probability $1-\delta$:

$$\sup_{g\in\mathcal{G}} \mid L_n(g) - L(g) \mid \leq \mathbb{E} \sup_{g\in\mathcal{G}} \mid L_n(g) - L(g) \mid + \sqrt{\frac{2\log(1/\delta)}{n}}$$

The ERM principle works!

With enough good training examples and computing power!

Machine Learning Implemented at Large Scale

In the Big Data era, massive datasets are available but... the acquisition process may be poorly controlled.

In **facial recognition** (FR), public databases do not represent well the target population in terms of ethnicity and gender.

FR systems perform much better on certain segments of the population

LFW (Huang et al, 2007)

13K images, 5.7K people

MS1M (Guo et al, 2016)

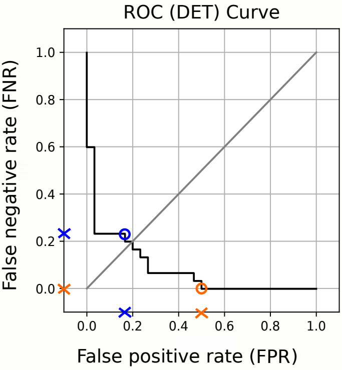
5.2M images, 93.4K people

FR systems are less accurate for certain social groups

In FR, the ROC curve evaluates a similarity function s *w.r.t.* its ability to separate positive and negative observations with thresholding $s > t$. (left)

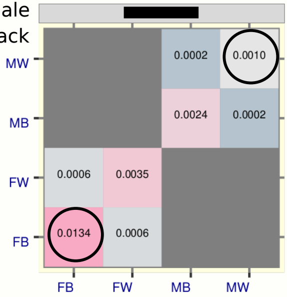
Recent reports of the NIST show discrepancies in error rates between social groups for FR. (right)

At fixed t, $13\times$ more FP for black females than white males.


ROC (DET) Curve

FPR for t s.t. $\mathrm{FPR}_{\mathrm{MW}} = 10^{-3}$

M/F: Male/Female
W/B: White/Black



	FB	FW	MB	MW
MW			0.0002	0.0010
MB			0.0024	0.0002
FW	0.0006	0.0035		
FB	0.0134	0.0006		

(Grother and Ngan, 2019)

Deployment of Machine Learning - Threats

- In many situations, training data are 'easily available' and 'Big'. Poor control of the acquisition process of the training data!

- The generalization ability of predictive rules is established when

$$P_{train} = P_{test}$$

  Otherwise? It requires novel versions of ML algorithms, new dedicated analyses and auxiliary information about the data acquisition process

- Many selection bias issues documented in the literature 'Women also Snowboard: Overcoming Bias in Captioning Models' in ECCV 2018, L.A. Hendricks et al.

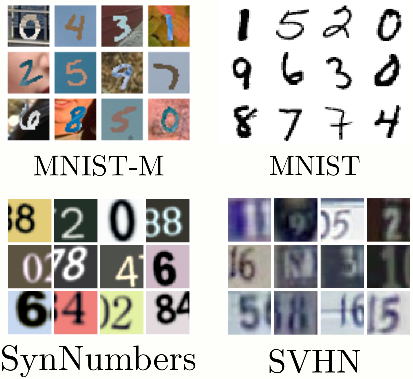
- In the Big Data era, can the ideas of the Scarce Data era, survey theory in particular, be of any help?

Domain Adaptation - Transferring Deep Features

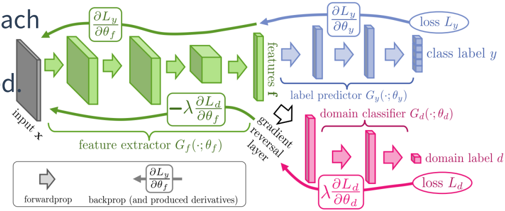
In computer vision, most of the transfer learning work focuses on **Domain Adaptation** (DA).

Most DA work seeks to correct for covariate shift ($p_{\text{train}} \neq p_{\text{test}}$) with invariant deep features, and sometimes also models differences in posterior distributions. [Tzeng et al., 2015] [Long et al., 2016]

In e.g. [Ganin et al., 2016], an approach to learn invariant deep features between visual domains is proposed.



(Ganin et al, 2016)


$\frac{\partial L_y}{\partial \theta_f}$
$\frac{\partial L_y}{\partial \theta_y}$
loss L_y
class label y
features f
label predictor $G_y(\cdot;\theta_y)$
$-\lambda \frac{\partial L_d}{\partial \theta_f}$
gradient reversal layer
domain classifier $G_d(\cdot;\theta_d)$
domain label d
loss L_d
$\lambda \frac{\partial L_d}{\partial \theta_d}$
input x
feature extractor $G_f(\cdot;\theta_f)$
forwardprop
backprop (and produced derivatives)

How to apply ERM to biased data?

- Goal: minimize the risk

  $$L_P(\theta) = \mathbb{E}_{Z \sim P}[\ell(Z,\ \theta)]$$

  over the decision space Θ, where P is the test/target distribution
- Training data available $Z'_1,\ \dots,\ Z'_n \overset{i.i.d.}{\sim} P'$, with $P' \neq P$
- A specific transfer learning problem
- **Heuristic:** solve a minimization problem

  $$\min_{\theta \in \Theta} \widehat{L}_{n,\omega}(\theta),$$

  where the objective is a weighted empirical risk

  $$\widehat{L}_{n,\omega}(\theta) = \sum_{i=1}^{n} \omega_i \ell(Z'_i,\ \theta).$$

How to apply ERM to biased data?

- Ideally, pick the ω_i's, so that

$$\sup_{\theta\in\Theta}\left|\widehat{L}_{n,\omega}(\theta)-L_P(\theta)\right| = O_{\mathbb{P}}(\sqrt{\log n/n})$$

- **Challenges:**
  - Design methods/algorithms to build the debiasing weights
  - Study the fluctuations of the nearly debiased risk

- Some auxiliary information about the biasing mechanism is required!

A First Go: Training from Survey Data

- Framework: original sample $(Z_1, \ \dots, \ Z_N)$ viewed as a superpopulation
- Sampling plan $R_N =$ probability distribution on the ensemble of all nonempty subsets of $\{1, \ \dots, \ N\}$
- Let $S \sim R_N$ and set $\epsilon_i = 1$ if $i \in S$, $\epsilon_i = 0$ otherwise
  The vector $(\epsilon_1, \ \dots, \ \epsilon_n)$ fully describes the training sample S
- First and second order inclusion probabilities:

$$\pi_i(R_N) = \mathbb{P}\{i \in S\} \text{ and } \pi_{i,j}(R_N) = \mathbb{P}\{(i,j) \in S^2\}$$

- Do not rely on the raw empirical risk based on the sample S:
  $\frac{1}{\#S} \sum_{i \in S} \ell(Z_i, \ \theta)$ is a biased estimate of $L_P(\theta)$

Horvitz -Thompson theory

- Suppose that the inclusion probabilities are known
- **Inverse Probability Weighting (IPW)**: Horvitz-Thompson estimator of the empirical distribution of the Z_i's

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\epsilon_i}{\pi_i}\delta_{Z_i}$$

- It is not a probability measure in general but yields an unbiased estimate of the (empirical) risk

$$L_N^{R_N}(\theta)=\frac{1}{N}\sum_{i=1}^{N}\frac{\epsilon_i}{\pi_i}\ell(Z_i,\ \theta)$$

- The Horvitz Thompson empirical risk minimizer

$$\arg\min_{\theta\in\Theta} L_N^{R_N}(\theta)=\hat{\theta}_N^{\epsilon}$$

A functional non-asymptotic Horvitz -Thompson theory

- Due to the dependence structure of the terms averaged in the HT risk, investigating the fluctuations of the supremum

$$\sup_{\theta\in\Theta}\left|L_N^{R_N}(\theta)-\hat{L}_N(\theta)\right|$$

  is not straightforward!

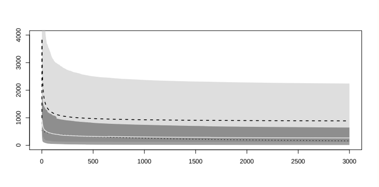
- Many situations can be handled: when data are sampled from
  - a Poisson scheme
  - a survey plan such that the ϵ_i's are negatively associated, e.g. rejective sampling, Srinivasan sampling, Rao-Sampford sampling, Successive sampling, Pareto sampling, Post-stratified sampling ...
  - any plan that can be **tightly coupled** with one of the schemes above

On the use of survey schemes for machine-learning

To minimize $\hat{L}_N(\theta)$, rather than implementing SGD based on mini-batches selected by simple sampling without replacement, use a Poisson scheme with inclusion probabilities positively correlated to

$$\pi_i^*(\theta) = N_0 \frac{||H_N^{-1/2} \nabla \ell(Z_i, \theta)||}{\sum_{j=1}^N ||H_N^{-1/2} \nabla \ell(Z_i, \theta)||},$$

with $H_N = \nabla^2 \hat{L}_N(\theta_N^*)$ to drastically reduce the asymptotic variance



More in Bertail, Chautru, Clémençon & Papa (2018)

Biased training data with known inclusion probabilities: done!

- It works for successive sampling, post-stratified sampling, *etc.*
- When the π_i's are known, the IPW method applied to ERM produces predictive rules with the same performance as that attained by ERM based on unbiased data

  More in e.g. Bertail, Clémençon & Papa (2016), Bertail, Chautru & Clémençon (2016, 2018)
- Well ... but what if the inclusion probabilities are unknown?

You may estimate them when you have some knowledge of the biasing mechanism...

ERM under Random Censorship

- Regression framework: predict a random duration $Y \geq 0$ based on a random vector X through $f(x)$, so as to minimize

$$L_P(f) = \mathbb{E}[(Y - f(X))^2]$$

- Right censored output data: n independent copies $(X_i, \tilde{Y}_i,\ \delta_i)$ of

$$X,\ \tilde{Y} = \min\{Y, C\},\ \delta = \mathbb{I}\{Y \leq C\},$$

  assuming that Y and C are conditionally independent given X
- Applying ERM to the $(X_i, \tilde{Y}_i)$'s would naturally lead to severe underestimation
- Set $S_C(t \mid X) = \mathbb{P}\{C > t \mid X\}$ and rewrite the risk as

$$L_P(f) = \mathbb{E}\left[\frac{\delta(\tilde{Y} - f(X))^2}{S_C(\tilde{Y}- \mid X)}\right]$$

ERM with IP(C)W

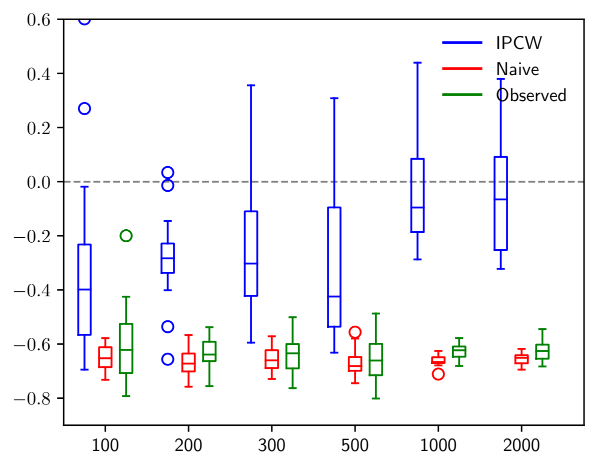
- Inverse of Probability of Censoring Weights: minimize

$$\tilde{L}_n(f) \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\hat{S}_C(\tilde{Y}_i- \mid X_i)} \left(\tilde{Y}_i - f(X_i)\right)^2$$

- The probability of censoring can be estimated by the Kaplan-Meier method
- The concentration properties of the process $\{\tilde{L}_n(f) - L_P(f)\}_{f\in\mathcal{F}}$ can be established by means of linearization techniques
- Neglecting the bias error due to the plug-in step, the classic learning rate $O_{\mathbb{P}}(\sqrt{\log n/n})$ is attained by ERM with IPCW

More in Ausset, Clémençon & Portier (2019)

ERM with approximate IPCW works!



But you need to know something about the selection bias process or to learn it from extra data!

Weighted ERM beyond IPW

- Assume $P << P'$. Let $\Phi = dP/dP'$
- Weights and Importance Sampling

$$\frac{1}{n}\sum_{i=1}^{n}\Phi(Z_i')\ell(Z_i',\theta)$$

  is an unbiased estimate of $L_P(\theta)$

- In various cases (e.g. covariate shift, positive-unlabeled learning, availability of several biased training samples), Φ can be estimated by means of the Z_i''s and auxiliary information about the target population

More in e.g. Clémençon and Laforgue (2019), Achab, Clémençon, Tillier and Vogel (2019) and Bertail, Clémençon, Guyonvarch & Noiry (2021)

Learning the ERM Weights - **Unknown** Poisson Sampling

- Assume $P << P'$ and macro-information about P is known (e.g. moments)
- A two-stage learning procedure:
  1. learn the weights ω_i so as to reproduce the macro-information
  2. solve the weighted ERM problem
- Under appropriate conditions, one gets the **same learning rate as if the sampling scheme was known**
- If the macro-information is rich enough, this an be extended even if $P << P'$ does not hold

More in Bertail, Clémençon, Guyonvarch & Noiry (2021) and in Clémençon, Guyonvarch & Noiry (2024)

Assessing Selection Bias - a ML approach

- How to test that $P \neq P'$? High-dimensional two-sample problem.
- If selection bias is significant, what kind of information is required to correct it?
- A novel ML approach based on bipartite ranking

More in Clémençon, Limnios & Vayatis (2021, 24)

In Facial Recognition, reweighting may be insufficient!

Certain learning subproblems can be much harder than others, possibly causing a great accuracy disparity.

Fig. 1. Images of Infant from the Database

See e.g. Bharadwaj et a. (2020)

Fairness constraints must be incorporated to the ERM program.

Why facial recognition algorithms can’t be perfectly fair? Clémençon & Maxwell (the Conversation, 2020).

Fair Machine Learning, beyond Biometrics

Algorithmic decisions are increasingly used in many domains:

Banking (*e.g.* loans)	Recruting (*e.g.*, hiring)
Insurance (*e.g.* cars)	Judiciary (*e.g.*, bail)

Recently, the fairness of algorithms has gathered lots of attention.

05/2016: The COMPAS system predicts recidivism likelihood for US courts.

Algorithms are designed for the interest of some party,
fairness in ML suggests confronting those to the law.

“Predictive models are really just opinions embedded in math.” C. O’Neil.

Lack of fairness is not always a consequence of selection bias.

An illustration is given by **age performance gaps** in biometrics.

See e.g. Achab, Clémençon, Tillié & Vogel (2020).

Fairness Definitions in Binary Classification

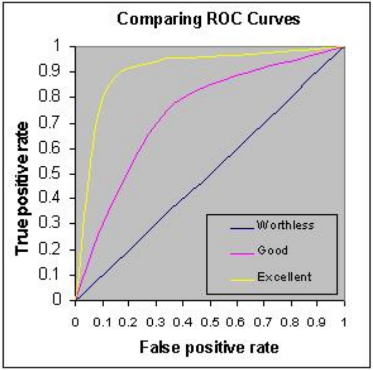
A lot of recent works considered fairness in binary classification, with two sensitive groups.
[Donini et al., 2018, Menon and Williamson, 2018, Zafar et al., 2019]

They add a **sensitive variable** $Z \in \{0, 1\}$ to the usual binary classification model (X, Y), and learn $g(X)$ from:

$$\mathcal{D}_n = \{(X_1, Y_1, Z_1), \dots, (X_n, Y_n, Z_n)\}.$$

Many definitions of fairness exist, and apply to specific use-cases.
· Treatment: $g(X, Z) = g(X)$ a.s.
· Impact: $\mathbb{P}\{g(X) = +1 \mid Z = 0\} = \mathbb{P}\{g(X) = +1 \mid Z = 1\}$
· Error: $\mathbb{P}\{g(X) \neq Y \mid Z = 0\} = \mathbb{P}\{g(X) \neq Y \mid Z = 1\}$
· FPR: $\mathbb{P}\{g(X) = +1 \mid Y = -1, Z = 0\} = \mathbb{P}\{g(X) = +1 \mid Y = -1, Z = 1\}$

On the Design of Fair Scoring Rules



- True positive rate:

$$1 - G_s(t) = \mathbb{P}\{s(X) \geq t \mid Y = 1\}$$

- False positive rate:

$$1 - H_s(t) = \mathbb{P}\{s(X) \geq t \mid Y = -1\}$$

ROC curve: $t \mapsto (1 - H_s(t), 1 - G_s(t))$

AUC = Area Under the ROC Curve

Fairness issues concern specific FPR ranges.

Learning with Pointwise ROC Constraints Bellet, Clémençon, Vogel (2021)

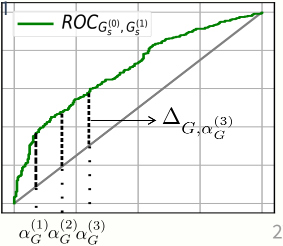
To measure the difference between cdfs for $Z = 0$ and $Z = 1$:

$$\Delta_{H,\alpha}(s) = \mathrm{ROC}_{H_s^{(0)},H_s^{(1)}}(\alpha) - \alpha \quad \text{and} \quad \Delta_{G,\alpha}(s) = \mathrm{ROC}_{G_s^{(0)},G_s^{(1)}}(\alpha) - \alpha.$$

Incorporate m_H pointwise constraints for $\Delta_{H,\cdot}$ and m_G for $\Delta_{G,\cdot}$ as a penalization, and maximize L_Λ in $\mathcal{S}$, where:

$$L_\Lambda(s) := \mathrm{AUC}_{H_s,G_s} - \sum_{k=1}^{m_H} \lambda_H^{(k)} |\Delta_{H,\alpha_H^{(k)}}(s)| - \sum_{k=1}^{m_G} \lambda_G^{(k)} |\Delta_{G,\alpha_G^{(k)}}(s)|.$$

Finite-sample generalization bounds of order $O_{\mathbb{P}}(n^{-1/2})$ have been proved.



Accuracy *vs* Fairness: Satisfactory trade-offs?

German Credit Dataset (German) in [Zafar et al., 2019, Zehlike et al., 2017, Singh and Joachims, 2019, Donini et al., 2018]. Sensitive variable: gender.

Bank Marketing Dataset (Bank) in [Zafar et al., 2019]: predict whether a client shall subscribe to a term deposit. Sensitive variable: age.

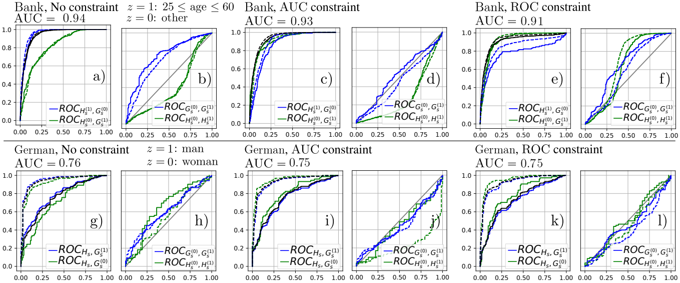


Figure 10: ROC curves for Bank and German for a score learned without and with fairness constraints. On all plots, dashed and solid lines represent respectively training and test sets. Black curves represent ROC_{H_s,G_s}, and above the curves we report the corresponding ranking performance AUC_{H_s,G_s}.

Some references

- Empirical processes in survey sampling. Bertail, Chautru and Clémençon (2016). Scandinavian J. Stat.
- Learning from Survey Training Samples: Rate Bounds for Horvitz-Thompson Risk Minimizers. Bertail, Clémençon & Papa. In ACML 2016.
- Optimal Survey Schemes for SGD with Applications to *M*-estimation. Bertail, Chautru, Clémençon & Papa. In ESAIM, 2018.
- Sampling and Empirical Risk Minimization. Bertail, Chautru and Clémençon (2016). In Statistics.
- Empirical Risk Minimization under Random Censorship. Ausset, Clémençon & Portier JMLR, 2022.
- Learning from Biased Training Samples. Clémençon & Laforgue. EJS, 2022
- Weighted ERM: Transfer Learning based on Importance Sampling. Achab, Clémençon, Tillier and Vogel. In ICMA, 2020.
- Learning Fair Scoring Functions: Bipartite Ranking under ROC-based Fairness Constraints. Bellet, Clémençon and Vogel (2021). AISTATS, 2021.
- Fighting selection bias in statistical learning: application to visual recognition from biased image databases. Clémençon, P. Laforgue and R. Vogel. In JNPS, 2023.
- Learning from Biased Data: A Semi-Parametric Approach. Clémençon et al., ICML, 2021.
- Mitigating Gender Bias in Face Recognition Using the von Mises-Fisher Mixture Model. Clémençon et al. ICML, 2022
- Assessing Uncertainty in Similarity Scoring: Performance Fairness in Face Recognition. Clémençon et al. ICLR, 2024.

References I

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., and Pontil, M. (2018).
Empirical risk minimization under fairness constraints.
In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, pages 2796–2806.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016).
Domain-adversarial training of neural networks.
Journal of Machine Learning Research.

Guo, Y., Zhang, L., Hu, Y., He, X., and Gao, J. (2016).
Ms-celeb-1m: A dataset and benchmark for large-scale face recognition.
In *Computer Vision - ECCV 2016 - 14th European Conference*, volume 9907 of *Lecture Notes in Computer Science*, pages 87–102. Springer.

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.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. (2016).
Unsupervised domain adaptation with residual transfer networks.
In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 136–144.

References II

Menon, A. K. and Williamson, R. C. (2018).
The cost of fairness in binary classification.
In *Conference on Fairness, Accountability and Transparency, FAT 2018*, volume 81 of *Proceedings of Machine Learning Research*, pages 107–118. PMLR.

Singh, A. and Joachims, T. (2019).
Policy learning for fairness in ranking.
In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 5427–5437.

Tzeng, E., Hoffman, J., Darrell, T., and Saenko, K. (2015).
Simultaneous deep transfer across domains and tasks.
In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV ’15, page 4068âĂŞ4076, USA. IEEE Computer Society.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2019).
Fairness constraints: A flexible approach for fair classification.
Journal of Machine Learning Research, 20(75):1–42.

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., and Baeza-Yates, R. (2017).
Fa*ir: A fair top-k ranking algorithm.
In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017*, pages 1569–1578. ACM.