@inproceedings{black2020fliptest,
author = {Black, Emily and Yeom, Samuel and Fredrikson, Matt},
title = {FlipTest: Fairness Testing via Optimal Transport},
year = {2020},
isbn = {9781450369367},
publisher = {Association for Computing Machinery},
address = {New York, NY, USA},
url = {https://doi.org/10.1145/3351095.3372845},
doi = {10.1145/3351095.3372845},
abstract = {We present FlipTest, a black-box technique for uncovering discrimination in classifiers. FlipTest is motivated by the intuitive question: had an individual been of a different protected status, would the model have treated them differently? Rather than relying on causal information to answer this question, FlipTest leverages optimal transport to match individuals in different protected groups, creating similar pairs of in-distribution samples. We show how to use these instances to detect discrimination by constructing a flipset: the set of individuals whose classifier output changes post-translation, which corresponds to the set of people who may be harmed because of their group membership. To shed light on why the model treats a given subgroup differently, FlipTest produces a transparency report: a ranking of features that are most associated with the model's behavior on the flipset. Evaluating the approach on three case studies, we show that this provides a computationally inexpensive way to identify subgroups that may be harmed by model discrimination, including in cases where the model satisfies group fairness criteria.},
booktitle = {Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency},
pages = {111–121},
numpages = {11},
keywords = {disparate impact, optimal transport, fairness, machine learning},
location = {Barcelona, Spain},
series = {FAT* '20}
}

@inproceedings{kusner2017counterfactual,
 author = {Kusner, Matt J and Loftus, Joshua and Russell, Chris and Silva, Ricardo},
 booktitle = {Advances in Neural Information Processing Systems},
 editor = {I. Guyon and U. V. Luxburg and S. Bengio and H. Wallach and R. Fergus and S. Vishwanathan and R. Garnett},
 pages = {4066--4076},
 publisher = {Curran Associates, Inc.},
 title = {Counterfactual Fairness},
 url = {https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf},
 volume = {30},
 year = {2017}
}

@article{wachter2017counterfactual,
  title={Counterfactual explanations without opening the black box: Automated decisions and the GDPR},
  author={Wachter, Sandra and Mittelstadt, Brent and Russell, Chris},
  journal={Harv. JL \& Tech.},
  volume={31},

```
  pages={841},
  year={2017},
  publisher={HeinOnline}
}

@book{pearl2009causality,
  title={Causality},
  author={Pearl, Judea},
  year={2009},
  publisher={Cambridge university press}
}

@book{villani2003topics,
  title={Topics in optimal transportation},
  author={Villani, C{\'e}dric},
  number={58},
  year={2003},
  publisher={American Mathematical Soc.}
}

@book{villani2008optimal,
  address = {Berlin},
  series = {Grundlehren der mathematischen Wissenschaften},
  title = {Optimal Transport: Old and New},
  isbn = {978-3-540-71049-3},
  lccn = {QA402.5 .V538 2009},
  shorttitle = {Optimal Transport},
  timestamp = {2017-03-08T20:25:11Z},
  number = {338},
  publisher = {{Springer}},
  author = {Villani, C{\'e}dric},
  year = {2008},
  note = {OCLC: ocn244421231}
}


@article{besse2019can,
  title={Can Everyday AI be Ethical? Machine Learning Algorithm
Fairness},
  author={Besse, Philippe and Castets-Renard, C{\'e}line and Garivier,
Aur{\'e}lien and Loubes, Jean-Michel},
  journal={Machine Learning Algorithm Fairness (May 20, 2018).
Statistiques et Soci{\'e}t{\'e}},
  volume={6},
  number={3},
  year={2019}
}

@article{chouldechova2017fair,
  title={Fair prediction with disparate impact: A study of bias in
recidivism prediction instruments},
  author={Chouldechova, Alexandra},
  journal={Big data},
  volume={5},
  number={2},
  pages={153--163},
  year={2017},
```

  publisher={Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA}
}

@article{besse2020survey,
  title={A survey of bias in Machine Learning through the prism of Statistical Parity for the Adult Data Set},
  author={Besse, Philippe and del Barrio, Eustasio and Gordaliza, Paula and Loubes, Jean-Michel and Risser, Laurent},
  journal={arXiv preprint arXiv:2003.14263},
  year={2020}
}

@inproceedings{poyiadzi2020face,
  title={FACE: feasible and actionable counterfactual explanations},
  author={Poyiadzi, Rafael and Sokol, Kacper and Santos-Rodriguez, Raul and De Bie, Tijl and Flach, Peter},
  booktitle={Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society},
  pages={344--350},
  year={2020}
}

@misc{joshi2019realistic,
      title={Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems},
      author={Shalmali Joshi and Oluwasanmi Koyejo and Warut Vijitbenjaronk and Been Kim and Joydeep Ghosh},
      year={2019},
      eprint={1907.09615},
      archivePrefix={arXiv},
      primaryClass={cs.LG}
}

@article{karimi2020algorithmic,
  title={Algorithmic Recourse: from Counterfactual Explanations to Interventions},
  author={Karimi, Amir-Hossein and Sch{\"o}lkopf, Bernhard and Valera, Isabel},
  journal={arXiv preprint arXiv:2002.06278},
  year={2020}
}

@article{peyre2019computational,
  title={Computational Optimal Transport: With Applications to Data Science},
  author={Peyr{\'e}, Gabriel and Cuturi, Marco and others},
  journal={Foundations and Trends{\textregistered} in Machine Learning},
  volume={11},
  number={5-6},
  pages={355--607},
  year={2019},
  publisher={Now Publishers, Inc.}
}

@inproceedings{hardt2016equality,
  title={Equality of opportunity in supervised learning},

  author={Hardt, Moritz and Price, Eric and Srebro, Nati},
  booktitle={Advances in neural information processing systems},
  pages={3315--3323},
  year={2016}
}

@inproceedings{li2019repair,
  title={Repair: Removing representation bias by dataset resampling},
  author={Li, Yi and Vasconcelos, Nuno},
  booktitle={Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition},
  pages={9572--9581},
  year={2019}
}

@misc{delbarrio2020centeroutward,
      title={Center-Outward Distribution Functions, Quantiles, Ranks, and Signs in $\mathbb{R}^d$},
      author={Eustasio del Barrio and Juan A. Cuesta-Albertos and Marc Hallin and Carlos Matrán},
      year={2020},
      eprint={1806.01238},
      archivePrefix={arXiv},
      primaryClass={stat.ME}
}

@article{del2020note,
  title={A note on the regularity of optimal-transport-based center-outward distribution and quantile functions},
  author={del Barrio, Eustasio and Gonz{\'a}lez-Sanz, Alberto and Hallin, Marc},
  journal={Journal of Multivariate Analysis},
  volume={180},
  pages={104671},
  year={2020},
  publisher={Elsevier}
}

@article{bongers2020foundations,
  title={Foundations of structural causal models with cycles and latent variables},
  author={Bongers, Stephan and Forr{\'e}, Patrick and Peters, Jonas and Sch{\"o}lkopf, Bernhard and Mooij, Joris M},
  journal={arXiv preprint arXiv:1611.06221},
  year={2020}
}

@inproceedings{dwork2012fairness,
  title={Fairness through awareness},
  author={Dwork, Cynthia and Hardt, Moritz and Pitassi, Toniann and Reingold, Omer and Zemel, Richard},
  booktitle={Proceedings of the 3rd innovations in theoretical computer science conference},
  pages={214--226},
  year={2012}
}

@misc{scholkopf2019causality,
      title={Causality for Machine Learning},
      author={Bernhard Schölkopf},
      year={2019},
      eprint={1911.10500},
      archivePrefix={arXiv},
      primaryClass={cs.LG}
}

@article{plecko2020fair,
  title={Fair Data Adaptation with Quantile Preservation},
  author={Plecko, Drago and Meinshausen, Nicolai},
  journal={Journal of Machine Learning Research},
  volume={21},
  pages={1--44},
  year={2020}
}

@book{imbens2015causal,
  title={Causal inference in statistics, social, and biomedical
sciences},
  author={Imbens, Guido W and Rubin, Donald B},
  year={2015},
  publisher={Cambridge University Press}
}

@book{barocas-hardt-narayanan,
  title = {Fairness and Machine Learning},
  author = {Solon Barocas and Moritz Hardt and Arvind Narayanan},
  publisher = {fairmlbook.org},
  note = {\url{http://www.fairmlbook.org}},
  year = {2019}
}

@article{cuturi2013sinkhorn,
  title={Sinkhorn distances: Lightspeed computation of optimal
transport},
  author={Cuturi, Marco},
  journal={Advances in neural information processing systems},
  volume={26},
  pages={2292--2300},
  year={2013}
}

@article{mccann1995,
author = "McCann, Robert J.",
doi = "10.1215/S0012-7094-95-08013-2",
fjournal = "Duke Mathematical Journal",
journal = "Duke Math. J.",
month = "11",
number = "2",
pages = "309--323",
publisher = "Duke University Press",
title = "Existence and uniqueness of monotone measure-preserving maps",
url = "https://doi.org/10.1215/S0012-7094-95-08013-2",
volume = "80",

year = "1995"
}

@book{pearl2016causal,
  title={Causal inference in statistics: A primer},
  author={Pearl, Judea and Glymour, Madelyn and Jewell, Nicholas P},
  year={2016},
  publisher={John Wiley \& Sons}
}

@misc{mahajan2020preserving,
      title={Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers},
      author={Divyat Mahajan and Chenhao Tan and Amit Sharma},
      year={2020},
      eprint={1912.03277},
      archivePrefix={arXiv},
      primaryClass={cs.LG}
}

@article{lewis1973,
  author = {David Lewis},
  title = {Causation},
  journal = {Journal of Philosophy},
  year = {1973},
  volume = 70,
  number = 17,
  pages = {556-567}}

@article{galles1998axiomatic,
  title={An axiomatic characterization of causal counterfactuals},
  author={Galles, David and Pearl, Judea},
  journal={Foundations of Science},
  volume={3},
  number={1},
  pages={151--182},
  year={1998},
  publisher={Springer}
}

@inproceedings{lundberg2017shap,
 author = {Lundberg, Scott M and Lee, Su-In},
 booktitle = {Advances in Neural Information Processing Systems},
 editor = {I. Guyon and U. V. Luxburg and S. Bengio and H. Wallach and R. Fergus and S. Vishwanathan and R. Garnett},
 pages = {4765--4774},
 publisher = {Curran Associates, Inc.},
 title = {A Unified Approach to Interpreting Model Predictions},
 url = {https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf},
 volume = {30},
 year = {2017}
}

@InProceedings{perov2020multiverse,

  title =   {MultiVerse: Causal Reasoning using Importance Sampling in Probabilistic Programming },
  author =   {Perov, Yura and Graham, Logan and Gourgoulias, Kostis and Richens, Jonathan and Lee, Ciaran and Baker, Adam and Johri, Saurabh},
  booktitle =   {Proceedings of The 2nd Symposium on
 Advances in Approximate Bayesian Inference},
  pages =   {1--36},
  year =   {2020},
  editor = {Cheng Zhang and Francisco Ruiz and Thang Bui and Adji Bousso Dieng and Dawen Liang},
  volume = {118},
  series = {Proceedings of Machine Learning Research},
  address =   {},
  month =   {08 Dec},
  publisher =   {PMLR},
  pdf =   {http://proceedings.mlr.press/v118/perov20a/perov20a.pdf},
  url =   {http://proceedings.mlr.press/v118/perov20a.html},
  abstract =   { We elaborate on using importance sampling for causal reasoning, in particular for counterfactual inference. We show how this can be implemented natively in probabilistic programming. By considering the structure of the counterfactual query, one can signicantly optimise the inference process. We also consider design choices to enable further optimisations. We introduce MultiVerse, a probabilistic programming prototype engine for approximate causal reasoning. We provide experimental results and compare with Pyro, an existing probabilistic programming framework with some of causal reasoning tools.}
}

@Inbook{Chickering1996,
author="Chickering, David Maxwell",
editor="Fisher, Doug
and Lenz, Hans-J.",
title="Learning Bayesian Networks is NP-Complete",
bookTitle="Learning from Data: Artificial Intelligence and Statistics V",
year="1996",
publisher="Springer New York",
address="New York, NY",
pages="121--130",
abstract="Algorithms for learning Bayesian networks from data have two components: a scoring metric and a search procedure. The scoring metric computes a score reflecting the goodness-of-fit of the structure to the data. The search procedure tries to identify network structures with high scores. Heckerman et al. (1995) introduce a Bayesian metric, called the BDe metric, that computes the relative posterior probability of a network structure given data. In this paper, we show that the search problem of identifying a Bayesian network---among those where each node has at most K parents---that has a relative posterior probability greater than a given constant is NP-complete, when the BDe metric is used.",
isbn="978-1-4612-2404-4",
doi="10.1007/978-1-4612-2404-4_12",
url="https://doi.org/10.1007/978-1-4612-2404-4_12"
}

@inproceedings{claassen2013learning,
author = {Claassen, Tom and Mooij, Joris M. and Heskes, Tom},
title = {Learning Sparse Causal Models is Not NP-Hard},
year = {2013},

publisher = {AUAI Press},
address = {Arlington, Virginia, USA},
abstract = {This paper shows that causal model discovery is not an NP-hard problem, in the sense that for sparse graphs bounded by node degree k the sound and complete causal model can be obtained in worst case order N2(k+2) independence tests, even when latent variables and selection bias may be present. We present a modification of the well-known FCI algorithm that implements the method for an independence oracle, and suggest improvements for sample/real-world data versions. It does not contradict any known hardness results, and does not solve an NP-hard problem: it just proves that sparse causal discovery is perhaps more complicated, but not as hard as learning minimal Bayesian networks.},
booktitle = {Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence},
pages = {172–181},
numpages = {10},
location = {Bellevue, WA},
series = {UAI'13}
}

@article{cooper1990computational,
author = {Cooper, Gregory F.},
title = {The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks (Research Note)},
year = {1990},
issue_date = {March 1990},
publisher = {Elsevier Science Publishers Ltd.},
address = {GBR},
volume = {42},
number = {2–3},
issn = {0004-3702},
url = {https://doi.org/10.1016/0004-3702(90)90060-D},
doi = {10.1016/0004-3702(90)90060-D},
journal = {Artif. Intell.},
month = mar,
pages = {393–405},
numpages = {13}
}

@article{scutari2019learning,
  title={Learning Bayesian networks from big data with greedy search: computational complexity and efficient implementation},
  author={Scutari, Marco and Vitolo, Claudia and Tucker, Allan},
  journal={Statistics and Computing},
  volume={29},
  number={5},
  pages={1095--1108},
  year={2019},
  publisher={Springer}
}

@article{cooper1992bayesian,
  title={A Bayesian method for the induction of probabilistic networks from data},
  author={Cooper, Gregory F and Herskovits, Edward},
  journal={Machine learning},
  volume={9},
  number={4},

  pages={309--347},
  year={1992},
  publisher={Springer}
}

@article{chickering2004large,
  title={Large-sample learning of Bayesian networks is NP-hard},
  author={Chickering, David Maxwell and Heckerman, David and Meek,
Christopher},
  journal={Journal of Machine Learning Research},
  volume={5},
  number={Oct},
  pages={1287--1330},
  year={2004}
}

@misc{delara2021consistent,
      title={A Consistent Extension of Discrete Optimal Transport Maps
for Machine Learning Applications},
      author={Lucas De Lara and Alberto González-Sanz and Jean-Michel
Loubes},
      year={2021},
      eprint={2102.08644},
      archivePrefix={arXiv},
      primaryClass={math.ST}
}

@inproceedings{russell2017when,
 author = {Russell, Chris and Kusner, Matt J and Loftus, Joshua and
Silva, Ricardo},
 booktitle = {Advances in Neural Information Processing Systems},
 editor = {I. Guyon and U. V. Luxburg and S. Bengio and H. Wallach and R.
Fergus and S. Vishwanathan and R. Garnett},
 pages = {},
 publisher = {Curran Associates, Inc.},
 title = {When Worlds Collide: Integrating Different Counterfactual
Assumptions in Fairness},
 url =
{https://proceedings.neurips.cc/paper/2017/file/1271a7029c9df08643b631b02
cf9e116-Paper.pdf},
 volume = {30},
 year = {2017}
}

@inproceedings{ribeiro2016should,
author = {Ribeiro, Marco Tulio and Singh, Sameer and Guestrin, Carlos},
title = {"Why Should I Trust You?": Explaining the Predictions of Any
Classifier},
year = {2016},
isbn = {9781450342322},
publisher = {Association for Computing Machinery},
address = {New York, NY, USA},
url = {https://doi.org/10.1145/2939672.2939778},
doi = {10.1145/2939672.2939778},
abstract = {Despite widespread adoption, machine learning models remain
mostly black boxes. Understanding the reasons behind predictions is,
however, quite important in assessing trust, which is fundamental if one
plans to take action based on a prediction, or when choosing whether to

deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one.In this work, we propose LIME, a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally varound the prediction. We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem. We demonstrate the flexibility of these methods by explaining different models for text (e.g. random forests) and image classification (e.g. neural networks). We show the utility of explanations via novel experiments, both simulated and with human subjects, on various scenarios that require trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier should not be trusted.},
booktitle = {Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining},
pages = {1135–1144},
numpages = {10},
keywords = {black box classifier, explaining machine learning, interpretability, interpretable machine learning},
location = {San Francisco, California, USA},
series = {KDD '16}
}