Unsupervised Learning under Latent Label Shift

Anonymous Author(s)
Affiliation
Address
email

Abstract

What sorts of structure might enable a learner to discover classes from unlabeled data? Traditional approaches rely on feature-space similarity and heroic assumptions on the data. In this paper, we introduce unsupervised learning under *Latent Label Shift* (LLS), where the label marginals $p_d(y)$ shift but the class conditionals $p(\mathbf{x}|y)$ do not. This work instantiates a new principle for identifying classes: elements that shift together group together. For finite input spaces, we establish an isomorphism between LLS and topic modeling: inputs correspond to words, domains to documents, and labels to topics. Addressing continuous data, we prove that when each label's support contains a separable region, analogous to an anchor word, oracle access to $p(d|\mathbf{x})$ suffices to identify $p_d(y)$ and $p_d(y|\mathbf{x})$ up to permutation. Thus motivated, we introduce a practical algorithm that leverages domain-discriminative models as follows: (i) push examples through domain discriminator $p(d|\mathbf{x})$; (ii) discretize the data by clustering examples in $p(d|\mathbf{x})$ space; (iii) perform non-negative matrix factorization on the discrete data; (iv) combine the recovered $p(y|d)$ with the discriminator outputs $p(d|\mathbf{x})$ to compute $p_d(y|x)\ \forall d$. With semi-synthetic experiments, we show that our algorithm can leverage domain information to improve state of the art unsupervised classification methods. We reveal a failure mode of standard unsupervised classification methods when data-space similarity does not indicate true groupings, and show empirically that our method better handles this case. Our results establish a deep connection between distribution shift and topic modeling, opening promising lines for future work.

1 Introduction

Discovering systems of categories from unlabeled data is a fundamental but ill-posed challenge in machine learning. Typical unsupervised learning methods group instances together based on feature-space similarity. Accordingly, given a collection of photographs of animals, a practitioner might hope that, in some appropriate feature space, images of animals of the same species should be somehow similar to each other. But why should we expect a clustering algorithm to recognize that dogs viewed in sunlight and dogs viewed at night belong to the same category? Why should we expect that butterflies and caterpillars should lie close together in feature space?

In this paper, we offer an alternative principle according to which we might identify a set of classes: we exploit distribution shift across times and locations to reveal otherwise unrecognizable groupings among examples. For example, if we noticed that whenever we found ourselves in a location where butterflies are abundant, caterpillars were similarly abundant, and that whenever butterflies were scarce, caterpillars had a similar drop in prevalence, we might conclude that the two were tied to the same underlying concept, no matter how different they appear in feature space. In short, our principle suggests that latent classes might be uncovered whenever *instances that shift together group together*.

Formalizing this intuition, we introduce the problem of unsupervised learning under *Latent Label Shift* (LLS). Here, we assume access to a collection of domains $d \in \{1, \ldots, r\}$, where the mixture

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

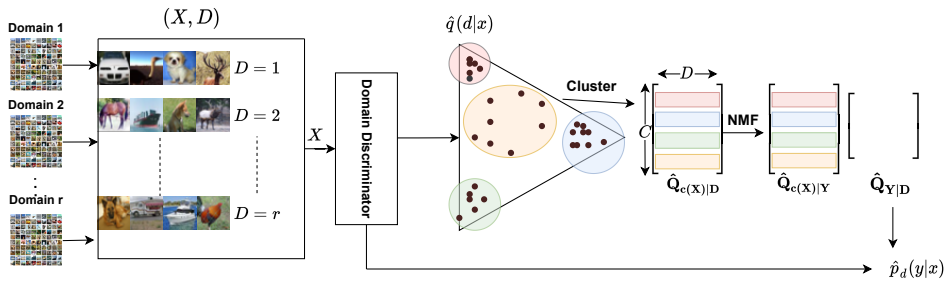

Figure 1: **Schematic of our DDFA algorithm**. After training a domain discriminator, we (i) push all data through the discriminator; (ii) cluster the data based on discriminator outputs; (iii) solve the resulting discrete topic modeling problem and then combine $\widehat{q}(d|x)$ and $\widehat{q}(y, d)$ to estimate $\widehat{p}_d(y|x)$.

proportions $p_d(y)$ vary across domains but the class conditional distribution $p(x|y)$ is domain-invariant. Our goals are to recover the underlying classes up to permutation, and thus to identify both the per-domain mixture proportions $p_d(y)$ and optimally adapted per-domain classifiers $p_d(y|x)$. The essential feature of our setup is that only the true y's, as characterized by their class-conditional distributions $p(x|y)$, could account for the observed shifts in $p_d(x)$. We prove that under mild assumptions, knowledge of this underlying structure is sufficient for inducing the full set of categories.

First, we focus on the *tabular setting*, demonstrating that when the input space is discrete and finite, LLS is isomorphic to topic modeling [8]. Here, each distinct input x maps to a *word* each *latent label* y maps to a *topic* and each domain d maps to a document. In this case, we can apply standard identification results for topic modeling [19, 4, 26, 31, 12] that rely only on the existence of anchor words within each topic (for each label y_i there is at least one x in the support of y_i, that is not in the support of any $y_j \neq y_i$). Here, standard methods based on Non-negative Matrix Factorization (NMF) can recover each domain's underlying mixture proportion $p_d(y)$ and optimal predictor $p_d(y|x)$. [19, 31, 26]. However, the restriction to discrete inputs, while appropriate for topic modeling, proves restrictive when our interests extend to high-dimensional continuous input spaces.

Then, to handle high-dimensional inputs, we propose *Discriminate-Discretize-Factorize-Adjust (DDFA)*, a general framework that proceeds in the following steps: (i) pool data from all domains to produce a mixture distribution $q(x, d)$; (ii) train a domain discriminative model f to predict $q(d|x)$; (iii) push all data through f, cluster examples in the pushforward distribution, and tabularize the data based on cluster membership; (iv) solve the resulting discrete topic modeling problem (e.g., via NMF), estimating $q(y, d)$ up to permutation of the latent labels; (v) combine the predicted $q(d|x)$ and $q(y, d)$ to to estimate $p_d(y)$ and $p_d(y|x)$. In developing this approach, we draw inspiration from recent works on distribution shift and learning from positive and unlabeled data that (i) leverage black box predictors to perform dimensionality reduction [37, 22, 23]; and (ii) work with *anchor sets*, separable subsets of continuous input spaces that belong to only one class's support [49, 38, 20, 6, 23].

Our main theoretical result shows that domain discrimination provides a sufficient representation for identifying all parameters of interest. Given oracle access to $q(d|x)$ (which is identified without labels), our procedure is asymptotically consistent. Our analysis reveals that the true $q(d|x)$ maps all points in the same anchor set to a single point mass in the push-forward distribution. This motivates our practical approach of discretizing data by hunting for tight clusters in $\widehat{q}(d|x)$ space.

In semi-synthetic experiments, we adapt existing image classification benchmarks to the LLS setting, sampling without replacement to construct collections of label-shifted domains. We note that training a domain discriminative classifier is a difficult task, and find that warm starting the initial layers of our model with pretrained weights from unsupervised approaches can significantly boost performance. We show that warm-started DDFA outperforms state-of-the-art (SOTA) unsupervised approaches when domain marginals $p_d(y)$ are sufficiently sparse. In particular, we observe improvements of as much as 30% accuracy over unsupervised SOTA on CIFAR-20. Further, on subsets of FieldGuide dataset, where similarity between species and diversity within a species leads to failure of unsupervised learning, we show that DDFA recovers the true distinctions. To be clear, these are not apples-to-apples comparisons: our methods are specifically tailored to the LLS setting. The takeaway is that the structure of the *LLS* setting can be exploited to outperform the best unsupervised learning heuristics.

2 Related Work

Unsupervised Learning Standard unsupervised learning approaches for discovering labels often rely on similarity in the original data space [39, 47]. While distances in feature space become meaningless for high-dimensional data, deep learning researchers have turned to similarity in a representations space learned via self-supervised contrastive tasks [41, 18, 25, 11], or similarity in a feature space learned end-to-end for a clustering task [9, 10, 44, 54]. Our problem setup closely resembles independent component analysis (ICA), where one seeks to identify statistically independent signal components from mixtures [32]. However, ICA's assumption of statistical independence among the components does not obtain in our setup. In topic modeling [8, 4, 31, 12, 43], documents are modeled as mixtures of topics, and topics as categorical distributions over a finite vocabulary. Topic models were pioneered by Latent Dirichlet Allocation (LDA) [8] and closely followed by papers that relaxed assumptions on the distribution of topic mixing coefficients (pLSI) [30, 43]. The topic modeling literature often draws on non-negative Matrix Factorization (NMF) methods [42, 50], which decompose a given matrix into a product of two matrices with non-negative elements [17, 15, 24, 27]. In both Topic Modeling and NMF, a fundamental problem has been to characterize the precise conditions under which the system is uniquely identifiable [19, 4, 31, 12]. The anchor condition (also referred to as separability) is known to be instrumental for identifying topic models [4, 12, 31, 19]. In this work, we extend these ideas, leveraging separable subsets of each label's support (the anchor sets) to produce anchor words in the discretized problem. Existing methods have attempted to extend latent variable modeling to continuous input domains by making assumptions about the functional forms of the class-conditional densities, e.g., restricting to Gaussian mixtures [47, 46]. A second line of approach involves finding an appropriate discretization of the continuous space [53].

Distribution Shift under the Label Shift Assumption The label shift assumption, where $p_d(y)$ can vary but $p(x|y)$ cannot, has been extensively studied in the domain adaptation literature [48, 51, 56, 37, 22] and also obtains in the problem of learning from positive and unlabeled data [21, 7, 23]. For both problems, many classical approaches suffer from the curse of dimensionality, failing in the settings where deep learning prevails. Our solution strategy draws inspiration from recent work on label shift [37, 1, 5, 22] and PU learning [7, 38, 49, 23] that leverage black-box predictors to produce sufficient low-dimensional representations for identifying target distributions of interest (other works leverage black box predictors heuristically [33]). **Key differences:** While PU learning requires identifying *one* new class for which we lack labeled examples provided that the positive class contains an anchor set [23], LLS can identify an arbitrary number of classes (up to permutation) from completely unlabeled data, provided a sufficient number of domains.

Domain Generalization The related problem of Domain Generalization (DG) also addresses learning with data drawn from multiple distributions and where the domain identifiers play a key role [40, 2]. However in DG, we are given *labeled* data from multiple domains, and our goal is to learn a classifier that can generalize to new domains. By contrast, in LLS, we work with unlabeled data only, leveraging the problem structure to identify the underlying labels.

3 Unsupervised Learning under Latent Label Shift

Notation For a vector $v \in \mathbb{R}^p$, we use v_j to denote its jth entry, and for an event E, we let $\mathbb{I}[E]$ denote the binary indicator of the event. By $|A|$, we denote the cardinality of set A. With $[n]$, we denote the set $\{1, 2, \ldots, n\}$. We use $[A]_{i,j}$ to access the element at (i, j) in A. Let $\mathcal{X}$ be the input space and $\mathcal{Y} = \{1, 2, \ldots, k\}$ the output space for multiclass classification. Throughout this paper, we use capital letters to denote random variables and small case letters to denote the corresponding values they take. For example, by X we denote the input random variable and by x, we denote a value that X may take.

We now formally introduce the problem of unsupervised learning under LLS. In LLS, we assume that we observe unlabeled data from r domains. Let $\mathcal{R} = \{1, 2, \ldots, r\}$ be the set of domains. By p_d, we denote the probability density (or mass) function for each domain $d \in \mathcal{R}$.

Definition 1 (Latent label shift)**.** *We observe data from r domains. While the label distribution among these domains can change, for all $d, d' \in \mathcal{R}$ and for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have $p_d(x|y) = p_{d'}(x|y)$.*

Simply put, Definition 1 states that the conditional distribution $p_d(x|y)$ remains invariant across domains, i.e., they satisfy the label shift assumption. Thus, we can drop the subscript on this factor, denoting all $p_d(x|y)$ by $p(x|y)$. Crucially, under LLS, $p_d(y)$ can vary across different domains.

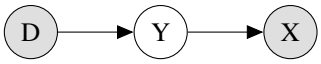

Figure 2: Relationship under Q between observed D, observed X, and latent Y.

Under LLS, we observe unlabeled data with domain label $\{(x_1, d_1), (x_2, d_2), \ldots, (x_n, d_n)\}$. Our goal breaks down into two tasks. Upto permutation of labels, we aim to (i) estimate the label marginal in each domain $p_d(y)$; and (ii) estimate the optimal per-domain predictor $p_d(y|x)$.

Mixing distribution Q A key step in our algorithm will be to train a domain discriminative model. Towards this end we define Q, a distribution over $\mathcal{X} \times \mathcal{Y} \times \mathcal{R}$, constructed by taking a uniform mixture over all domains. By q, we denote the probability density (or mass) function of Q. Define Q such that $q(x, y|D = d) = p_d(x, y)$, i.e., when we condition on $D = d$ we recover the joint distribution over $\mathcal{X} \times \mathcal{Y}$ specific to that domain d. For all $d \in \mathcal{R}$, we define $\gamma_d = q(d)$, i.e., the prevalence of each domain in our distribution Q. Notice that $q(x, y)$ is a mixture over the distributions $\{p_d(x, y)\}_{d\in\mathcal{R}}$, with $\{\gamma_d\}_{d\in\mathcal{R}}$ as the corresponding mixture coefficients. Under LLS (Definition 1), X does not depend on D when conditioned on Y (Fig. 2).

Additional notation for the discrete case To begin, we setup notation for discrete input spaces with $|\mathcal{X}| = m$. Without loss of generality, we assume that $\mathcal{X} = \{1, 2, \ldots, m\}$. The label shift assumption allows us to formulate the label marginal estimation problem in matrix form. Let $\mathbf{Q}_{X|D}$ be an $m \times r$ matrix such that $[\mathbf{Q}_{X|D}]_{i,d} = p_d(X = i)$, i.e., the d-th column of $\mathbf{Q}_{X|D}$ is $p_d(x)$. Let $\mathbf{Q}_{X|Y}$ be an $m \times k$ matrix such that $[\mathbf{Q}_{X|Y}]_{i,j} = p(X = i|Y = j)$, the j-th column is a distribution over X given $Y = j$. Similarly, define $\mathbf{Q}_{Y|D}$ as a $k \times r$ matrix whose d-th column is the domain marginal $p_d(y)$. Now with Definition 1, we have $p_d(x) = \sum_y p_d(x, y) = \sum_y p_d(x|y)p_d(y) = \sum_y p(x|y)p_d(y)$. Since this is true $\forall d \in \mathcal{R}$, we can express this in a matrix form as $\mathbf{Q}_{X|D} = \mathbf{Q}_{X|Y}\mathbf{Q}_{Y|D}$.

Additional assumptions Before we present identifiability results for the LLS problem, we introduce four additional assumptions required throughout the paper:

- A.1 There are at least as many domains as classes, i.e., $|\mathcal{R}| \geqslant |\mathcal{Y}|$.
- A.2 The matrix formed by label marginals (as columns) across different domains is full-rank, i.e., $\text{rank}(\mathbf{Q}_{Y|D}) = k$.
- A.3 Equal representation of domains, i.e., for all $d \in \mathcal{R}, \gamma_d = 1/r$.
- A.4 Fix $\epsilon > 0$. For all $y \in \mathcal{Y}$, there exists a subdomain $A_y \subseteq \mathcal{X}$, such that $q(A_y) \geqslant \epsilon$ with $q(A_y|y) > 0$ and for all $y' \in \mathcal{Y}\backslash\{y\}$, $q(A_y|y') = 0$. We refer to this assumption as ϵ-anchor sub-domain condition.

We now comment on the assumptions. A.1–A.2 are benign, these assumptions just imply that the matrix $\mathbf{P}_{Y|D}$ is full row rank. Without loss of generality, A.3 can be assumed when dealing with data from a collection of domains. When this condition is not satisfied, one could just re-sample data points uniformly at random from each domain d. Intuitively, A.4 states that for each label $y \in \mathcal{Y}$, we have some subset of inputs that only belong to that class y. To avoid vanishing probability of this subset, we ensure at least ϵ probability mass in our mixing distribution Q. The anchor word condition is related to the positive sub-domain in PU learning, which requires that there exists a subset of $\mathcal{X}$ in which all examples only belong to the positive class [49, 38, 20, 6].

4 Theoretical Analysis

In this section, we establish identifiability of LLS problem. We begin by considering the case where the input space is discrete and formalize the isomorphism to topic modeling. Then we establish the identifiability of the system in this discrete setting by appealing to existing results in topic modeling [31]. Finally, extending results from discrete case, we provide novel analysis to establish our identifiability result for the continuous setting.

Isomorphism to topic modeling Recall that for the discrete input setting, we have the matrix formulation: $\mathbf{Q}_{X|D} = \mathbf{Q}_{X|Y}\mathbf{Q}_{Y|D}$. Consider a corpus of r documents, consisting of terms from a vocabulary of size m. Let $\mathbf{D}$ be an $\mathbb{R}^{m\times r}$ matrix representing the underlying corpus. Each column of

$\mathbf{D}$ represents a document, and each row represents a term in the vocabulary. Each element $[\mathbf{D}]_{i,j}$ represents the frequency of term i in document j. Topic modeling [8, 30, 31, 4] considers each document to be composed as a mixture of k topics. Each topic prescribes a frequency with which the terms in the vocabulary occur given that topic. Further, the proportion of each topic varies across documents with the frequency of terms given topic remaining invariant.

We can state the topic modelling problem as: $\mathbf{D} = \mathbf{CW}$, where $\mathbf{C}$ is an $\mathbb{R}^{m\times k}$ matrix, $[\mathbf{C}]_{i,j}$ represents the frequency of term i given topic j, and $\mathbf{W}$ is an $\mathbb{R}^{k\times r}$ matrix, where $[\mathbf{W}]_{i,j}$ represents the proportion of topic i in document j. Note that all three matrices are column normalized. The isomorphism is then between document and domain, topic and label, term and input sample, i.e., $\mathbf{D} = \mathbf{CW} \equiv \mathbf{Q}_{X|D} = \mathbf{Q}_{X|Y}\mathbf{Q}_{Y|D}$. In both the cases, we are interested in decomposing a known matrix into two unknown matrices. This formulation is examined as a non-negative matrix factorization problem with an added simplicial constraint on the columns (columns sum to 1) [3, 26].

Identifiability of the topic modeling problem is well-established [19, 4, 26, 31, 12]. We leverage the isomorphism to topic modeling to extend this identifiability condition to our LLS setting. We formalize the adaption here:

Theorem 1. *(adapted from Proposition 1 in Huang et al. [31]) Assume A.1, A.2 and A.4 hold (A.4 in the discrete setting is referred to as the anchor word condition). Then the solution to $\mathbf{Q}_{X|D} = \mathbf{Q}_{X|Y}\mathbf{Q}_{Y|D}$ is uniquely identified.*

We refer readers to Huang et al. [31] for a proof of this theorem. Intuitively, Theorem 1 states that if each label y has at least one token in the input space that has support only in y, and A.1, A.2 hold, then the solution to $\mathbf{Q}_{X|Y}$, $\mathbf{Q}_{Y|D}$ is unique. Furthermore, under this condition, there exist algorithms that can recover $\mathbf{Q}_{X|Y}$, $\mathbf{Q}_{Y|D}$ within some permutation [31, 26, 3, 4].

Extensions to the continuous case We will prove identifiability in the continuous setting, when $\mathcal{X} = \mathbb{R}^p$ for some $p \geqslant 1$. In addition to A.1–A.4, we make an additional assumption that we have oracle access to $q(d|x)$, i.e., the true domain discriminator for mixture distribution Q. This is implied by assuming access to the marginal $q(x, d)$ from which we observe our samples. Formally, we define a push forward function f such that $[f(x)]_d = q(d|x)$, then push the data forward through f to obtain outputs in Δ^{r-1}. In the proof of Theorem 2, we will show that these outputs can be discretized in a fashion that maps anchor subdomains to anchor words in a tabular, discrete setting. We separately remark that the anchor word outputs are in fact extreme corners of the convex polytope in Δ^{r-1} which encloses all $f(x)$ mass; we discuss this geometry further in App. F. After constructing the anchor word discretization, we appeal to Theorem 1 to recover $\mathbf{Q}_{Y|D}$. Given $\mathbf{Q}_{Y|D}$, we show that we can use Bayes' rule and the LLS condition (Definition 1) to identify the distribution $q(y|x, d) = p_d(y|x)$ over latent variable y. We formalize this in the following theorem:

Theorem 2. *Let the distribution Q over random variables X, Y, D satisfy Assumptions A.1–A.4. Assuming access to the joint distribution $q(x, d)$, we show that the following quantities are identifiable: (i) $\mathbf{Q}_{Y|D}$, (ii) $q(y|X = x)$, for all $x \in \mathcal{X}$ that lies in the support (i.e. $q(x) > 0$); and (iii) $q(y|X = x, D = d)$, for all $x \in \mathcal{X}$ and $d \in \mathcal{R}$ such that $q(x, d) > 0$.*

Before presenting a proof sketch for Theorem 2, we first present key lemmas (we include their proofs in App. B).

Lemma 1. *Under the same assumptions as Theorem 2, the matrix $\mathbf{Q}_{Y|D}$ and $f(x) = q(d|x)$ uniquely determine $q(y|x)$ for all $y \in \mathcal{Y}$ and $x \in \mathcal{X}$ such that $q(x) > 0$.*

Lemma 1 states that given matrix $\mathbf{Q}_{Y|D}$ and oracle domain discriminator, we can uniquely identify $q(y|x)$. In particular, we show that for any $x \in \mathcal{X}$, $q(d|x)$ can be expressed as a convex combination of the k columns of $\mathbf{Q}_{D|Y}$ (which is computed from $\mathbf{Q}_{Y|D}$ and is column rank k) and the coefficients of the combination are $q(y|x)$. Combining this with the linear independence of the columns of $\mathbf{Q}_{D|Y}$, we show that these coefficients are unique. In the following lemma, we show how the identified $q(y|x)$ can then be used to identify $q(y|x, d)$:

Lemma 2. *Under the same assumptions as Theorem 2, for all $y \in \mathcal{Y}$, and $x \in \mathcal{X}$ such that $q(x, d) > 0$. the matrix $\mathbf{Q}_{Y|D}$ and $q(y|x)$ uniquely determine $q(y|x, d)$.*

To prove Lemma 2, we show that we can combine the conditional distribution over the labels given a sample $x \in \mathcal{X}$ with the prior distribution of the labels in each domain to determine the posterior

distribution over labels given the sample x and the domain of interest. Next, we introduce a key property of the domain discriminator classifier f:

Lemma 3. *Under the same assumptions as Theorem 2, for all x, x' in anchor sub-domain, i.e., $x, x' \in A_y$ for a given label $y \in \mathcal{Y}$, we have $f(x) = f(x')$. Further, for any $y \in \mathcal{Y}$, if $x \in A_y, x' \notin A_y$, then $f(x) \neq f(x')$.*

Lemma 3 implies that the oracle domain discriminator f maps all points in an anchor subdomain, and only those points in that anchor subdomain to the same point in $f(x) = q(d|x)$ space. We can now present a proof sketch for Theorem 2 (full proof in App. B):

Proof sketch of Theorem 2. The key idea of the proof lies in proposing a discretization such that some subset of anchor subdomains for each label y in the continuous space map to distinct anchor words in discrete space. In particular, if there exists a discretization of the continuous space $\mathcal{X}$ that for any $y \in \mathcal{Y}$, maps all $x \in A_y$ to the same point in the discrete space, but no $x \notin A_y$ maps to this point, then this point serves as an anchor word. From Lemma 3, we know that all the $x \in A_y$ and only the $x \in A_y$ get mapped to specific points in the $f(x)$ space. Pushing all the $x \in \mathcal{X}$ through f, we know from A.4 that there exists k point masses of size ϵ, one for each $f(A_y)$ in the $f(x) = q(d|x)$ space. We can now inspect this space for point masses of size at least ϵ to find at most $\mathcal{O}(1/\epsilon)$ such point masses among which are contained the k point masses corresponding to the anchor subdomains. Discretizing this space by assigning each point mass to a group (and non-point masses to a single additional group), we have k groups that have support only in one y each. Thus, we have achieved a discretization with anchor words. Further, since the discrete space arises from a pushforward of the continuous space through f, the discrete space also satisfies the latent label shift assumption A.1. We now use Theorem 1 to claim identifiability of $\mathbf{Q}_{Y|D}$. We then use Lemmas 1 and 2 to prove parts (ii) and (iii).

5 DDFA Framework

Motivated by our identifiability analysis, in this section, we present an algorithm to estimate $\mathbf{Q}_{Y|D}, q(y|x)$, and $q(y|x, d)$ when X is continuous by exploiting domain structure and approximating the true domain discriminator f. Intuitively, $q(y|x, d)$ is the domain specific classifier $p_d(y|x)$ and $q(y|x)$ is the classifier for data from aggregated domains. $\mathbf{Q}_{Y|D}$ captures label marginal for individual domains. A naive approach would be to aggregate data from different domains and exploit recent advancements in unsupervised learning [54, 44, 9, 10]. However, aggregating data from multiple domains loses the domain structure that we hope to leverage. We highlight this failure mode of the unsupervised clustering method in Sec. 6.

Discriminate We begin Algorithm 1 by creating a split of the unlabeled samples into the training and validation sets. Using the unlabeled data samples and the domain that each sample originated from, we first train a domain discriminative classifier $\widehat{f}$. The domain discriminative classifier outputs a distribution over domains for a given input. This classifier is trained with cross-entropy loss to predict the domain label of each sample on the training set. With unlimited data, the minimizer of this loss is the true f, as we prove in App. C. To avoid overfitting, we stop training $\widehat{f}$ when the cross-entropy loss on the validation set stops decreasing. Note that here the validation set also only contains domain labels (and no information about true labels).

Discretize We now push forward all the samples from the training and validation sets through the domain discriminator to get vector $\widehat{f}(x_i)$ for each sample x_i. In the proof of Theorem 2, we argue that when working with true f, and the entire marginal $q(x)$, we can choose a discretization satisfying the anchor word assumption by identifying point masses in the distribution of $f(x)$ and filtering to include those of at least ϵ size. In the practical setting, because we have only a finite set of data points and a noisy $\widehat{f}$, we use clustering to approximately find point masses. We choose $m \geqslant k$ and recover m clusters with any standard clustering procedure (e.g. K-means). If the noise in $\widehat{f}$ is sufficiently small and the clustering sufficiently granular, we intuit that our m discovered clusters include k *pure* clusters, each of which only contains data points from a different anchor subdomain which are tightly arranged around the true $f(A_y)$ for the corresponding label y. This clustering is superior to a naive clustering on the input space because close proximity in this space indicates similarity in $q(d|x)$.

Let us denote the learned clustering function as c, where $c(x)$ is the cluster assigned to a datapoint x. We now leverage the cluster id $c(x_i)$ of each sample x_i to discretize sample into a finite discrete

Algorithm 1 DDFA Training

input $k \geqslant 1, r \geqslant k, \{(x_i, d_i)\}_{i\in[n]} \sim q(x, d),$ A class of functions $\mathcal{F}$ from $\mathbb{R}^p \to \mathbb{R}^r$

1: Split into train set T and validation set V

2: Train $\widehat{f} \in \mathcal{F}$ to minimize cross entropy loss for predicting $d|x$ on T with early stopping on V

3: Push all $\{x_i\}_{i\in[n]}$ through $\widehat{f}$

4: Train clustering algorithm on the n points $\{\widehat{f}(x_i)\}_{i\in[n]},$ obtain m clusters.

5: $c(x_i) \leftarrow$ Cluster id of $\widehat{f}(x_i)$

6: $\widehat{q}(c(X) = a|D = b) \leftarrow \frac{\sum_{i\in[n]} \mathbb{I}[c(x_i)=a,\ d_i=b]}{\sum_{j\in[n]} \mathbb{I}[d_j=b]}$

7: Populate $\widehat{\mathbf{Q}}_{c(X)|D}$ as $[\widehat{\mathbf{Q}}_{c(X)|D}]_{a,b} \leftarrow \widehat{q}(c(X) = a|D = b)$

8: $\widehat{\mathbf{Q}}_{c(X)|D}, \widehat{\mathbf{Q}}_{Y|D} \leftarrow$ NMF $(\widehat{\mathbf{Q}}_{c(X)|D})$

output $\widehat{\mathbf{Q}}_{Y|D}, \widehat{f}$

Algorithm 2 DDFA Prediction

input $\widehat{\mathbf{Q}}_{Y|D}, \widehat{f}, (x', d') \sim q(x, d)$

1: Populate $\widehat{\mathbf{Q}}_{D|Y}$ as $[\widehat{\mathbf{Q}}_{D|Y}]_{d,y} \leftarrow \frac{[\widehat{\mathbf{Q}}_{Y|D}]_{y,d}}{\sum_{d''=1}^{d''=r}[\widehat{\mathbf{Q}}_{Y|D}]_{y,d''}}$

2: Assign $\widehat{q}(y|X = x') \leftarrow \left[\left(\widehat{\mathbf{Q}}_{D|Y}\right)^{\dagger} \widehat{f}(x')\right]_y$

3: Assign $\widehat{q}(y|X = x', D = d') \leftarrow \dfrac{[\widehat{\mathbf{Q}}_{D|Y}]_{d',y}\widehat{q}(y|X = x')}{\sum\limits_{y''\in[k]} [\widehat{\mathbf{Q}}_{D|Y}]_{d',y''}\widehat{q}(y''|X = x')}$

4: $y_{\text{pred}} \leftarrow \arg\max_{y\in[k]} \widehat{q}(y|X = x', D = d')$

output : $\widehat{q}(y|X = x', D = d') = \widehat{p}_{d'}(y|x'), \widehat{q}(y|X = x'), y_{\text{pred}}$

space $[m]$. Combining cluster id with the domain source d_i for each sample, we estimate $\widehat{\mathbf{Q}}_{c(X)|D}$ by simply computing, for each domain, the fraction of its samples assigned to each cluster.

Factorize We apply an NMF algorithm to $\widehat{\mathbf{Q}}_{c(X)|D}$ to obtain our estimates of $\widehat{\mathbf{Q}}_{c(X)|Y}$ and $\widehat{\mathbf{Q}}_{Y|D}$.

Adjust We begin Algorithm 2 by considering a test point (x', d'). To make a prediction, if we had access to oracle f and true $\mathbf{Q}_{Y|D}$, we could precisely compute $q(y|x')$ (Lemma 1). However, in place of these true quantities, we plug in the estimates $\widehat{f}$ and $\widehat{\mathbf{Q}}_{Y|D}$. Since our estimates contain noise, the estimate $\widehat{q}(y|x')$ is found by left-multiplying $\widehat{f}(x')$ with the pseudo-inverse of $\widehat{\mathbf{Q}}_{D|Y}$, as opposed to solving a sufficient system of equations. As our estimates $\widehat{f}$ and $\widehat{\mathbf{Q}}_{D|Y}$ approach the true values, the projection of $\widehat{f}(x')$ into the column space of $\widehat{\mathbf{Q}}_{D|Y}$ tends to $\widehat{f}(x')$ itself, so the pseudo-inverse approaches the true solution. Now we can use the constructive procedure introduced in the proof of Lemma 2 to compute the plug-in estimate $\widehat{q}(y|x', d') = \widehat{p}_{d'}(y|x')$.

6 Experiments

Baselines We select the unsupervised classification method SCAN as a state-of-the-art baseline [54]. SCAN pretrains a ResNet [28] backbone using SimCLR [11] and MoCo [29] setups (pretext tasks). SCAN then trains a clustering head to minimize the SCAN loss (refer [54] for more details) [1]. We make sure to evaluate SCAN on the same potentially class-imbalanced test subset we create for each experiment. Since SCAN is fit on a superset of the data DDFA sees, we believe this gives a slight data advantage to the SCAN baseline (although we acknowledge that the class balance for SCAN training is also potentially different from its evaluation class balance). To evaluate SCAN, we use the public pretrained weights available for CIFAR-10, CIFAR-20, and ImageNet-50. We also train SCAN ourselves on the train and validation portions of the FieldGuide2 and FieldGuide28 datasets with a ResNet18 backbone and SimCLR pretext task. We replicate the hyperparameters used for CIFAR training.

[1] SCAN code: `https://github.com/wvangansbeke/Unsupervised-Classification`

Datasets First we examine standard multiclass image datasets CIFAR-10, CIFAR-20 [35], and ImageNet-50 [16] containing images from 10, 20, and 50 classes respectively. Images in these datasets typically focus on a single large object which dominates the center of the frame, so unsupervised classification methods which respond strongly to similarity in visual space are well-suited to recover true classes up to permutation. These datasets are often believed to be separable (i.e., single true label applies to each image), so every example falls in an anchor subdomain (satisfying A.4).

Motivated by the application of LLS problem, we consider the FieldGuide dataset [2], which contains images of moths and butterflies. The true classes in this dataset are species, but each class contains images taken in immature (caterpillar) and adult stages of life. Based on the intuition that butterflies from a given species look more like butterflies from other species than caterpillars from their own species, we hypothesize that unsupervised classification will learn incorrect class boundaries which distinguish caterpillars from butterflies, as opposed to recovering the true class boundaries. Due to high visual similarity between members of different classes, this dataset may indeed have slight overlap between classes. However, we hypothesize that anchor subdomain still holds, i.e., there exist some images from each class that could only come from that class. Additionally, if we have access to data from multiple domains, it is natural to assume that within each domain the relative distribution of caterpillar to adult stages of each species stay relatively constant as compared to prevalence of different species. We create two subsets of this dataset: FieldGuide2, with two species, and FieldGuide28, with 28 species.

LLS Setup The full sampling procedure for semisynthetic experiments is described in App. D. Roughly, we sample $p_d(y)$ from a symmetric Dirichlet distribution with concentration α/k, and enforce maximum condition number κ on $\mathbf{Q}_{\mathbf{Y}|\mathbf{D}}$. Small α and small κ encourages sparsity in $\mathbf{Q}_{\mathbf{Y}|\mathbf{D}}$, so each label tends to only appear in a few domains. Larger parameters encourages $p_d(y)$ to tend toward uniform. We draw from test, train, and valid datasets without replacement to match these distributions, but discard some examples due to class imbalance.

Training and Evaluation The algorithm uses train and validation data consisting of pairs of images and domain indices. We train ResNet50 [28] (with added dropout) on images x_i with domain indices d_i as the label, choose best iteration by valid loss, pass all training and validation data through $\widehat{f}$, and cluster pushforward predictions $\widehat{f}(x_i)$ into $m \geqslant k$ clusters with Faiss K-Means [34]. We compute the $\widehat{\mathbf{Q}}_{c(X)|D}$ matrix and run NMF to obtain $\widehat{\mathbf{Q}}_{c(X)|Y}$, $\widehat{\mathbf{Q}}_{Y|D}$. To make columns sum to 1, we normalize columns of $\widehat{\mathbf{Q}}_{c(X)|Y}$, multiply each column's normalization coefficient over the corresponding row of $\widehat{\mathbf{Q}}_{Y|D}$ (to preserve correctness of the decomposition), and then normalize columns of $\widehat{\mathbf{Q}}_{Y|D}$. Some NMF algorithms only output solutions satisfying the anchor word property [3, 36, 26]. We found the strict requirement of an exact anchor word solution to lead to low noise tolerance. We therefore use the Sklearn implementation of standard NMF [13, 52, 45].

We instantiate the domain discriminator as ResNet18, and preseed its backbone with SCAN [54] pre-trained weights or [54] contrastive pre-text weights. We denote these models DDFA (SI) and DDFA (SPI) respectively. We predict class labels with Algorithm 2. With the Hungarian algorithm, implemented in [14, 55], we compute the highest true accuracy among any permutation of these labels (denoted "Test acc"). With the same permutation, we reorder rows of $\widehat{P}_{Y|D}$, then compute the average absolute difference between corresponding entries of $\widehat{\mathbf{Q}}_{Y|D}$ and $\mathbf{Q}_{Y|D}$ (denoted "$\mathbf{Q}_{Y|D}$ err").

Results On CIFAR-10, we observe that DDFA alone is incapable of matching highly competitive state-of-the-art baseline SCAN performance—however, in suitably sparse problem settings (small α), it comes substantially close, indicating good recovery of true classes. Due to space constraints, we include CIFAR-10 results in App. E. DDFA (SI) combines SCAN's strong pretrain with domain discrimination fine-tuning to outperform SCAN in the easiest, sparsest setting and *certain* denser settings. On CIFAR-20, baseline SCAN is much less competitive, so our DDFA(SI) dominates baseline SCAN in all settings except the densest (Table 1). These results demonstrate how adding domain information can dramatically boost unsupervised baselines.

On FieldGuide-2, DDFA (SPI) outperforms SCAN baselines across all problem settings and domain counts (Table 2); in sparser settings, the accuracy gap is 20-30%. In this dataset, SCAN performs only slightly above chance, reflecting perhaps a total misalignment of learned class distinctions with true species boundaries. We do not believe that SCAN is too weak to effectively detect image groupings

[2] FieldGuide: `https://sites.google.com/view/fgvc6/competitions/butterflies-moths-2019`

Table 1: *Results on CIFAR-20*. With DDFA (RI) we refer to DDFA with randomly initialized backbone. With DDFA (SI) we refer to DDFA's backbone initialized with SCAN. Note that in DDFA (SI), we do not leverage SCAN for clustering. α is the Dirichlet parameter used for generating label marginals in each domain, κ is the maximum allowed condition number of the generated $\mathbf{Q}_{Y|D}$ matrix.

r	Approaches	$\alpha : 0.5,\ \kappa : 8$		$\alpha : 3,\ \kappa : 12$		$\alpha : 10,\ \kappa : 20$	
		Test acc	$\mathbf{Q}_{Y\|D}$ err	Test acc	$\mathbf{Q}_{Y\|D}$ err	Test acc	$\mathbf{Q}_{Y\|D}$ err
20	SCAN	0.4241	-	0.4313	-	**0.4362**	-
	DDFA (RI)	0.4872	0.0452	0.3047	0.043	0.1036	0.0727
	DDFA (SI)	**0.7507**	**0.0248**	**0.503**	**0.0284**	0.3252	**0.0348**
25	SCAN	0.437	-	0.4631	-	0.4434	-
	DDFA (RI)	0.4678	0.0505	0.3157	0.0483	0.0788	0.0802
	DDFA (SI)	**0.8364**	**0.0201**	**0.7403**	**0.0195**	**0.4819**	**0.0322**
30	SCAN	0.4387	-	0.4514	-	0.4214	-
	DDFA (RI)	0.5406	0.0453	0.287	0.05	0.091	0.0789
	DDFA (SI)	**0.8158**	**0.0219**	**0.7247**	**0.0222**	**0.5473**	**0.0258**

Table 2: *Results on FieldGuide*. With DDFA (SPI) we refer to DDFA's backbone initialized with pretext training done adopted by SCAN. Note that only the backbone of DDFA is initialized with SCAN-pretext weights and not the final layers. α is the Dirichlet parameter used for generating label marginals in each domain, κ is the maximum allowed condition number of the generated $\mathbf{Q}_{Y|D}$ matrix.

Approaches	**FieldGuide-2**			**FieldGuide-28**				
	r	$\alpha : 3,\ \kappa : 5$		r	$\alpha : 0.5,\ \kappa : 12$		$\alpha : 3,\ \kappa : 20$	
		Test acc	$\mathbf{Q}_{Y\|D}$ err		Test acc	$\mathbf{Q}_{Y\|D}$ err	Test acc	$\mathbf{Q}_{Y\|D}$ err
SCAN	2	0.6000	-	28	0.3634	-	0.3042	-
DDFA (SPI)		**0.8576**	0.1211		**0.6002**	0.0309	**0.3062**	0.0268
SCAN	3	0.6123	-	37	0.3253	-	0.2709	-
DDFA (SPI)		**0.8478**	0.0351		**0.7337**	0.0263	**0.4987**	0.0327
SCAN	5	0.5892	-	42	0.358	-	0.2956	-
DDFA (SPI)		**0.6827**	0.3395		**0.6068**	0.0377	**0.4407**	0.0341
SCAN	10	0.5865	-	47	0.3897	-	0.3187	-
DDFA (SPI)		**0.8094**	0.2799		**0.5624**	0.0414	**0.4573**	0.0341

on this data; instead we acknowledge that the domain information available to DDFA (SPI) (and not to SCAN) is informative for ensuring recovery of the true class distinction between species as opposed to the visually striking distinction between adult and immature life stages. Results from more domains are available in App. E. We also show the results on FieldGuide-28 over a range of domains in (Table 2). Our method outperforms SCAN on all settings with the highest observed accuracy difference ranging up to 30%.

7 Conclusion

Our theoretical results demonstrate that under LLS, we can leverage shifts among previously seen domains to recover labels in a purely unsupervised manner, and our practical instantiation of the DDFA framework demonstrates both (i) the practical efficacy of our approach; and (ii) that generic unsupervised methods can play a key role both in clustering discriminator outputs, and providing weights for initializing the discriminator. We believe that this work is just the first step in a new direction for leveraging structural assumptions together with distribution shift to perform unsupervised learning. Within the LLS setup, several components of the DDFA framework warrant further investigation: (i) the deep domain discriminator can be enhanced in myriad ways; (ii) for clustering discriminator outputs, we might develop methods specially tailored to our setting; (iii) clustering might be replaced altogether with geometrically informed methods that directly identify the corners of the polytope; (iv) the theory of LLS can be extended beyond identification to provide statistical results that might hold when $q(d|y)$ can only be noisily estimated, and when only finite samples are available for the induced topic modeling problem.

References

[1] Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Adapting to label shift with bias-corrected calibration. In *International Conference on Machine Learning (ICML)*, 2019.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[3] Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization – provably. In *Symposium on Theory of Computing (STOC)*, 2012. doi: 10.1145/2213977.2213994. URL `https://doi.org/10.1145/2213977.2213994`.

[4] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *Foundations of Computer Science (FOCS)*. IEEE, 2012.

[5] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations (ICLR)*, 2019.

[6] Jessa Bekker and Jesse Davis. Estimating the class prior in positive and unlabeled data through decision tree induction. In *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 32, 2018. URL `https://ojs.aaai.org/index.php/AAAI/article/view/11715`.

[7] Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: a survey. *Machine Learning*, 109(4):719–760, apr 2020. doi: 10.1007/s10994-020-05877-5. URL `https://doi.org/10.1007%2Fs10994-020-05877-5`.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning research*, 3(Jan):993–1022, 2003.

[9] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.

[10] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data, 2019. URL `https://arxiv.org/abs/1905.01278`.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning (ICML)*, pages 1597–1607. PMLR, 2020.

[12] Yinyin Chen, Shishuang He, Yun Yang, and Feng Liang. Learning topic models: Identifiability and finite-sample analysis. *arXiv preprint arXiv:2110.04232*, 2021.

[13] Andrzej Cichocki and Anh-Huy Phan. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE Transactions*, 92-A:708–721, 03 2009. doi: 10.1587/transfun.E92.A.708.

[14] David F Crouse. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016.

[15] Thiago de Paulo Faleiros and Alneu de Andrade Lopes. On the equivalence between algorithms for non-negative matrix factorization and latent dirichlet allocation. In *European Symposium on Artificial Neural Networks (ESANN)*, 2016.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Ieee, 2009.

[17] Chris Ding, Tao Li, and Wei Peng. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics & Data Analysis*, 52(8): 3913–3927, 2008.

[18] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction, 2015. URL `https://arxiv.org/abs/1505.05192`.

[19] David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? *Advances in Neural Information Processing Systems (NeurIPS)*, 16, 2003.

[20] Marthinus C. du Plessis, Gang Niu, and Masashi Sugiyama. Class-prior estimation for learning from positive and unlabeled data. *Machine Learning*, 106(4):463–492, nov 2016. doi: 10.1007/s10994-016-5604-6. URL https://doi.org/10.1007%2Fs10994-016-5604-6.

[21] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 213–220, 2008.

[22] Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A unified view of label shift estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2003.07554.

[23] Saurabh Garg, Yifan Wu, Alex Smola, Sivaraman Balakrishnan, and Zachary Lipton. Mixture proportion estimation and PU learning: A modern approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://arxiv.org/abs/2111.00980.

[24] Eric Gaussier and Cyril Goutte. Relation between plsa and nmf and implications. In *Conference on Research and Development in Information Retrieval (SIGIR)*, 2005.

[25] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2018. URL https://arxiv.org/abs/1803.07728.

[26] Nicolas Gillis and Stephen A. Vavasis. Fast and robust recursive algorithms for separable non-negative matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):698–714, apr 2014. doi: 10.1109/tpami.2013.226. URL https://doi.org/10.1137%2F130946782.

[27] Mark Girolami and Ata Kabán. On an equivalence between plsi and lda. In *Conference on Research and Development in Information Retrieval (SIGIR)*, 2003.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[30] Thomas Hofmann. Probabilistic latent semantic indexing. In *Conference on Research and Development in Information Retrieval (SIGIR)*, 1999.

[31] Kejun Huang, Xiao Fu, and Nikolaos D Sidiropoulos. Anchor-free correlated topic modeling: Identifiability and algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[32] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[33] Dmitry Ivanov. DEDPUL: Difference-of-estimated-densities-based positive-unlabeled learning. *arXiv preprint arXiv:1902.06965*, 2019.

[34] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[35] Alex Krizhevsky and Geoffrey Hinton. Learning Multiple Layers of Features from Tiny Images. Technical report, Citeseer, 2009.

[36] Abhishek Kumar, Vikas Sindhwani, and Prabhanjan Kambadur. Fast conical hull algorithms for near-separable non-negative matrix factorization. In *International Conference on Machine Learning*, pages 231–239. PMLR, 2013.

[37] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning (ICML)*. PMLR, 2018.

[38] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 38(3):447–461, mar 2016. doi: 10.1109/tpami.2015.2456899. URL `https://doi.org/10.1109%2Ftpami.2015.2456899`.

[39] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[40] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning (ICML)*. PMLR, 2013.

[41] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[42] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[43] Christos H Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 61(2): 217–235, 2000.

[44] Sungwon Park, Sungwon Han, Sundong Kim, Danu Kim, Sungkyu Park, Seunghoon Hong, and Meeyoung Cha. Improving unsupervised image clustering with robust learning. *CoRR*, abs/2012.11150, 2020. URL `https://arxiv.org/abs/2012.11150`.

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research (JMLR)*, 12, 2011.

[46] Kedar S Prabhudesai, Boyla O Mainsah, Leslie M Collins, and Chandra S Throckmorton. Augmented latent dirichlet allocation (lda) topic model with gaussian mixture topics. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2451–2455. IEEE, 2018.

[47] Douglas A Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

[48] Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. *Neural Computation*, 2002.

[49] Clayton Scott. A Rate of Convergence for Mixture Proportion Estimation, with Application to Learning from Noisy Labels. In *Artificial Intelligence and Statistics (AISTATS)*, 2015. URL `https://proceedings.mlr.press/v38/scott15.html`.

[50] D Seung and L Lee. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems (NeurIPS)*, 2001.

[51] Amos Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30:3–28, 2009.

[52] Vincent YF Tan and Cédric Févotte. Automatic relevance determination in nonnegative matrix factorization with the/spl beta/-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(7), 2012.

[53] Dongping Tian. Research on plsa model based semantic image analysis: A systematic review. *J. Inf. Hiding Multim. Signal Process.*, 9(5):1099–1113, 2018.

[54] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels, 2020. URL https://arxiv.org/abs/2005.12320.

[55] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[56] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning (ICML)*. PMLR, 2013.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We believe that this work, which proposes a novel unsupervised learning problem does not present a significant societal concern. While this could potentially guide practitioners to improve classification, we do not believe that it will fundamentally impact how machine learning is used in a way that could conceivably be socially salient.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Sec. 3
   (b) Did you include complete proofs of all theoretical results? [Yes] See Sec. 4 and appendices
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We include all the necessary details to replicate our experiments in appendices. We plan to release code with an updated version of the draft.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Yes, we describe crucial details in Sec. 6 and defer precise details to appendices.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We include results with multiple seeds in appendices
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Refer to experimental setup in App. D.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Refer to experimental setup in App. D.
   (b) Did you mention the license of the assets? [Yes] Refer to experimental setup in App. D.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Refer to experimental setup in App. D.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]