Mandoline: Model Evaluation under Distribution Shift

Mayee Chen*, Karan Goel*, Nimit Sohoni*, Fait Poms, Kayvon Fatahalian, Christopher Ré

Motivation

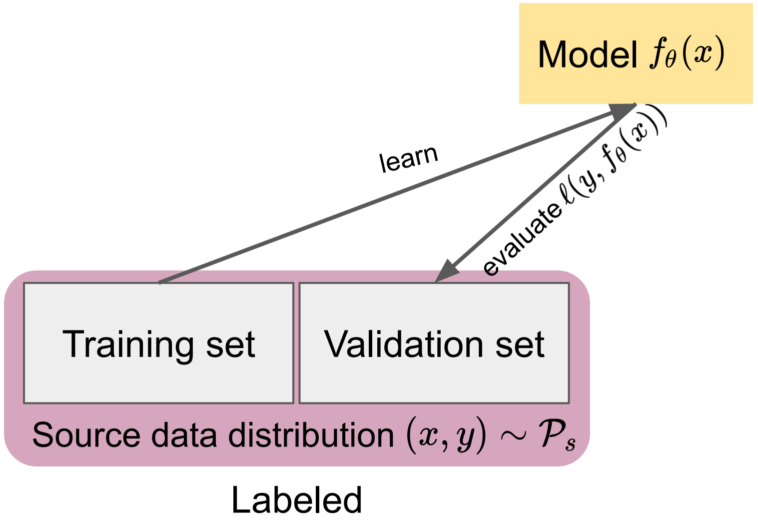
Q: How do we **evaluate** model performance during deployment?

- Model's deployment setting ≠ training setting

Motivation

Q: How do we **evaluate** model performance during deployment?

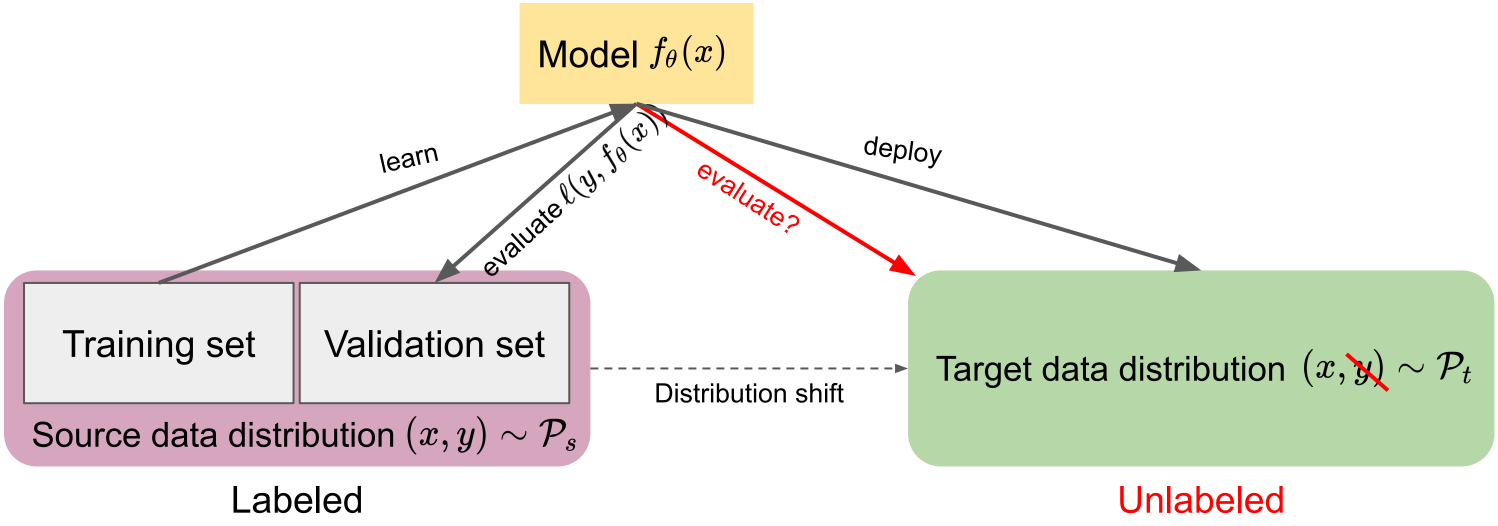
- Model's deployment setting ≠ training setting



Motivation

Q: How do we **evaluate** model performance during deployment?

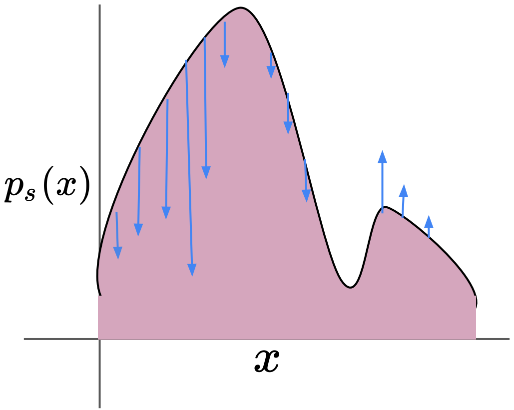
- Model's deployment setting ≠ training setting



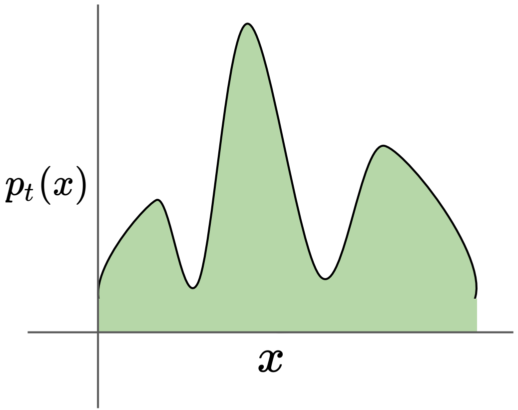
Mandoline: user-guided framework for evaluation under distribution shift

Common approach: importance weighting

Source data distribution $\mathcal{P}_s$



Target data distribution $\mathcal{P}_t$



$$\mathbb{E}_t[\ell(y, f_\theta(x))] = \mathbb{E}_s[\tfrac{p_t(x)}{p_s(x)}\ell(y, f_\theta(x))] \approx \tfrac{1}{n}\sum_{i=1}^{n} \boxed{\frac{p_t(x_i)}{p_s(x_i)}} \ell(y_i, f_\theta(x_i))$$

Density ratio

Problems:

- Support shift - what if $p_s(x) = 0, p_t(x) \neq 0$?
- High dimensional data $x \in \mathbb{R}^d$: harder to compute $\frac{p_t(x)}{p_s(x)}$

Mandoline: Slice-based reweighting framework

Slice: user-defined grouping of data $g(x) \in \{-1, 1\}$

negation
contains not, n't
$g_1(x)$

male pronoun
contains he, him
$g_2(x)$

strong sentiment
contains love, adore
$g_3(x)$

Mandoline: Slice-based reweighting framework

Slice: user-defined grouping of data $g(x) \in \{-1, 1\}$

negation contains not, n't	**male pronoun** contains he, him	**strong sentiment** contains love, adore
$g_1(x)$	$g_2(x)$	$g_3(x)$

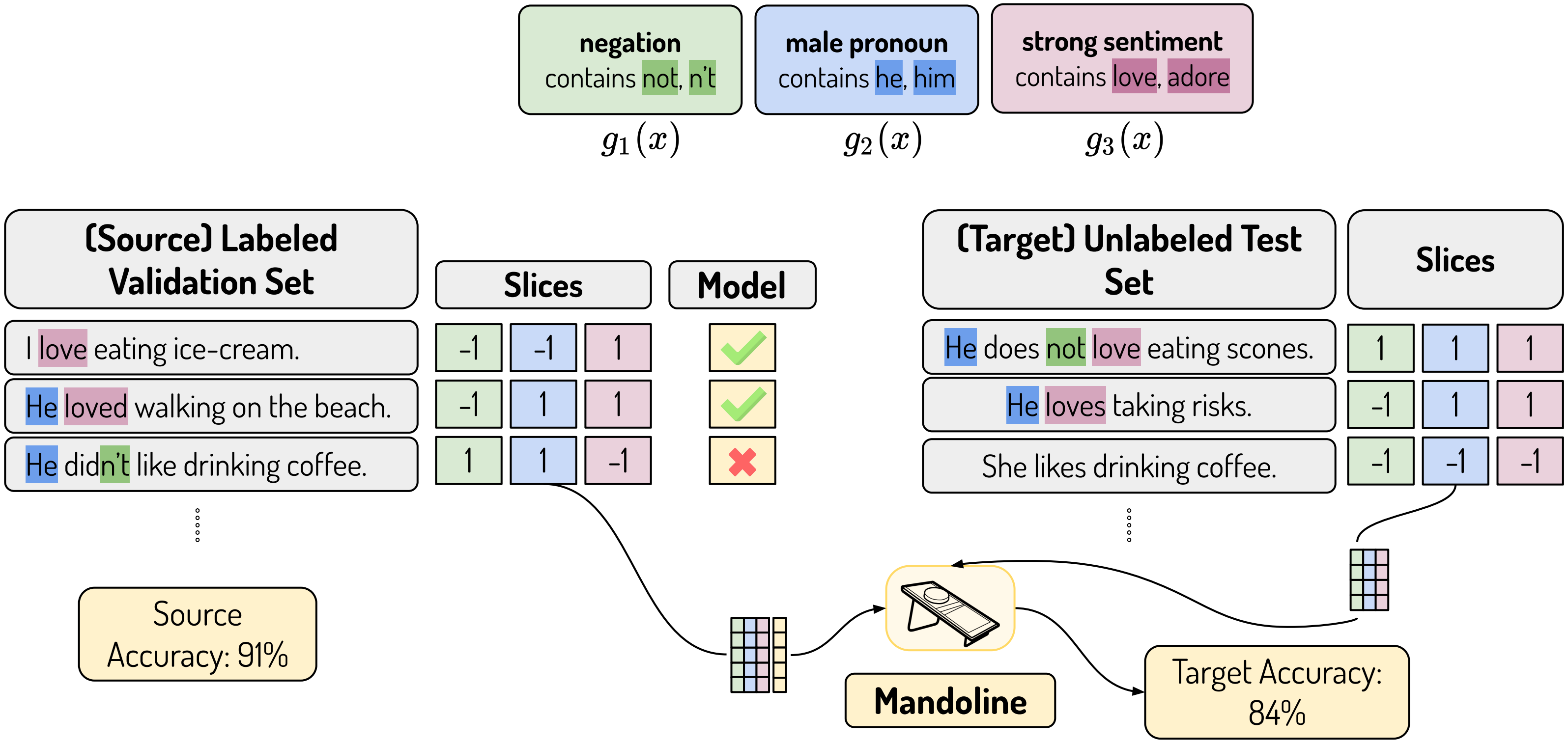
(Source) Labeled Validation Set	Slices			Model
I love eating ice-cream.	-1	-1	1	✓
He loved walking on the beach.	-1	1	1	✓
He didn't like drinking coffee.	1	1	-1	✗
⋮				

Source Accuracy: 91%

(Target) Unlabeled Test Set	Slices		
He does not love eating scones.	1	1	1
He loves taking risks.	-1	1	1
She likes drinking coffee.	-1	-1	-1
⋮			

Mandoline: Slice-based reweighting framework

Slice: user-defined grouping of data $g(x) \in \{-1, 1\}$



Results

Prop 1: if the slices $g = \{g_1, \ldots, g_k\}$ *capture all "relevant" distributional shift* between $\mathcal{P}_s$ and $\mathcal{P}_t$, then reweighting with $\frac{p_t(g(x))}{p_s(g(x))}$ recovers $\mathbb{E}_t[\ell(y, f_\theta(x))]$.

- If support shift occurs on irrelevant slices (i.e. slices independent of *Y*), it can be corrected!
- Dimensionality: reduce from *d* → *k*

- How to compute $\frac{p_t(g(x))}{p_s(g(x))}$? Use any density ratio estimation method on *g(x)*
  - **Kullback-Leibler Importance Estimation Procedure (KLIEP)**
    - Can modify to correct for noisily defined slices

Results

Task	Task Labels	Distribution Shift	Slices
CELEBA *image classification*	*male vs. female*	↑ blurry images	METADATA LABELS *blurry / not blurry*
SNLI→MNLI *natural language inference*	*entailment, neutral or contradiction*	single-genre → multi-genre examples	PROGRAMMATIC *task model predictions, task model entropy*

AVERAGE ESTIMATION ERROR (%)		
METHOD	CELEBA	
	RESNET18	RESNET50
SOURCE	1.96%	1.74%
CBIW	0.54%	0.52%
KMM	1.78%	1.67%
MANDOLINE	**0.14%**	**0.10%**

Importance weighting on x

On g(x)

CelebA

METHOD	STANDARD ACCURACY	
	AVG. ERROR	MAX. ERROR
SOURCE	$6.2\% \pm 3.8\%$	15.6%
CBIW	$5.5\% \pm 4.5\%$	17.9%
KMM	$5.7\% \pm 3.6\%$	14.6%
MANDOLINE	$\mathbf{3.6\% \pm 1.6\%}$	**5.9%**

SNLI → MNLI

Future directions

Slice design

- Mandoline relies on sufficient *g* to capture all axes along which the distribution shift occurs
- How to construct *g?*
    - Users write them using domain knowledge
    - Metadata
    - Model-based (threshold based on entropy)
    - Algorithmically (subset selection, check independence)

A new way of thinking about evaluation under distribution shift: from improving methods to developing ***better slices***

Contact: *mfchen@stanford.edu*