Estimating Accuracy from Unlabeled Data

Emmanouil Antonios Platanios — e.a.platanios@cs.cmu.edu
Avrim Blum — avrim@cs.cmu.edu
Tom Mitchell — tom.mitchell@cs.cmu.edu

1. Problem

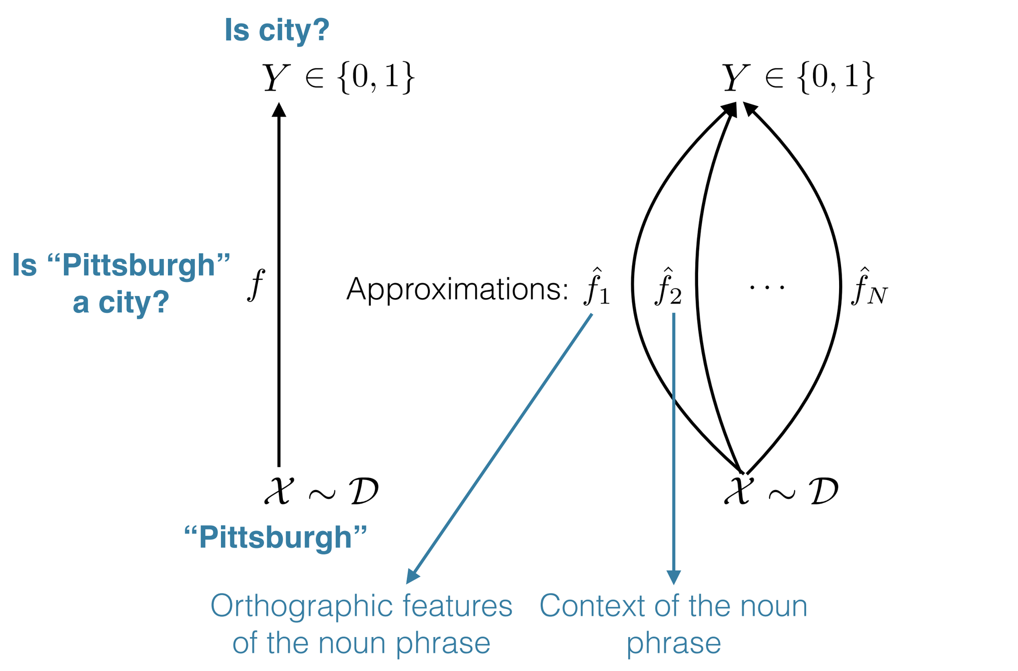
Using **only unlabeled data** we can measure **consistency** but not **correctness**. So:

consistency

Does this implication hold? — If yes, under what conditions?

correctness

There exists a **binary function** f that we do not know. Instead, we have a set of function approximations to that function and **we want to know how accurate they are**.



Consistency definition:

Given **unlabeled input data**, $X_1, \dots, X_S$, we observe the **sample agreement rates**:

$$\hat{a}_{\mathcal{A}} = \frac{1}{S} \sum_{s=1}^{S} \mathbb{I}\left\{\hat{f}_i(X_s) = \hat{f}_j(X_s), \forall i, j \in \mathcal{A} : i \neq j\right\}$$

Correctness definition:

Error Rate: The probability over $\mathbb{P}(\mathcal{X}) = \mathcal{D}$ of disagreeing with the correct output label.

$$e_{\mathcal{A}} = \mathbb{P}_{\mathcal{D}}\Big(\underbrace{\bigcap_{i \in \mathcal{A}} [\hat{f}_i(X) \neq Y]}_{E_{\mathcal{A}}}\Big)$$

Error Event $\leftarrow E_{\mathcal{A}}$

2. Approach

1 Agreement rates

Agreement rate between $\hat{f}_i$ and $\hat{f}_j$:

$$a_{\{i,j\}} = \mathbb{P}_{\mathcal{D}}\big(\overbrace{E_{\{i\}} \cap E_{\{j\}}}^{\text{both are wrong}}\big) + \mathbb{P}_{\mathcal{D}}\big(\overbrace{\bar{E}_{\{i\}} \cap \bar{E}_{\{j\}}}^{\text{both are right}}\big)$$

$$a_{\{i,j\}} = 1 - e_{\{i\}} - e_{\{j\}} + 2e_{\{i,j\}}$$

Probability that $\hat{f}_i$ makes an error — Probability that $\hat{f}_j$ makes an error — Probability that both make an error

Independent errors: $e_{\{i\}}e_{\{j\}}$

3 functions that make independent errors:

$\binom{3}{2} = 3$ equations

3 unknowns

$$e_{\{i\}} = \frac{c \pm (1 - 2\hat{a}_{\{j,k\}})}{\pm 2 (1 - 2\hat{a}_{\{j,k\}})}$$

where $i \in \{1, 2, 3\}$ $j, k \in \{1, 2, 3\} \backslash i$ with $j < k$ and:

$$c = \sqrt{(2\hat{a}_{\{1,2\}} - 1)(2\hat{a}_{\{1,3\}} - 1)(2\hat{a}_{\{2,3\}} - 1)}$$

Independent errors $\longrightarrow$ **Too strong** assumption $\longrightarrow$ We do not make it

But we end up with **more unknowns than equations**

Minimize the dependence between the error rates:

$$c(\boldsymbol{e}) = \sum_{i,j \in \{1,\dots,N\}} \left(e_{\{i,j\}} - e_{\{i\}}e_{\{j\}}\right)^2$$

Relaxes the independence assumption

Agreement Rates Equations $\xrightarrow{\text{Equality Constraints}}$ **Constrained Optimization Problem** $\leftarrow$ **Objective Function**

Inequality constraints:

$$e_{\{i,j\}} \leq \min\left\{e_{\{i\}}, e_{\{j\}}\right\}$$

Agreement rate for a bigger set of functions:

$$a_{\mathcal{A}} = \mathbb{P}_{\mathcal{D}}\Big(\overbrace{\bigcap_{i \in \mathcal{A}} E_i}^{\text{all are wrong}}\Big) + \mathbb{P}_{\mathcal{D}}\Big(\overbrace{\bigcap_{i \in \mathcal{A}} \bar{E}_i}^{\text{all are right}}\Big)$$

$$a_{\mathcal{A}} = e_{\mathcal{A}} + 1 + \sum_{k=1}^{|\mathcal{A}|} \left[(-1)^k \sum_{\substack{I \subseteq \mathcal{A} \\ |I| = k}} e_I\right]$$

Inequality constraints for a bigger set of functions:

$$e_{\mathcal{A}} \leq \min_{i \in \mathcal{A}} e_{\mathcal{A} \backslash i}$$

Objective function for a bigger set of functions:

$$c(\boldsymbol{e}) = \sum_{\mathcal{A} : |\mathcal{A}| \geq 2} \left(e_{\mathcal{A}} - \prod_{i \in \mathcal{A}} e_i\right)^2$$

Apply to the maximum likelihood method as well

2 Maximum likelihood

Likelihood definition: $L(\boldsymbol{e}) = \mathbb{P}_{\mathcal{D}}\left(\hat{\boldsymbol{Y}}_1, \dots, \hat{\boldsymbol{Y}}_S | \boldsymbol{e}\right) = \prod_{s=1}^{S} \mathbb{P}_{\mathcal{D}}\left(\hat{\boldsymbol{Y}}_s | \boldsymbol{e}\right)$

where $\boldsymbol{e}$ contains all of the function error rates, and $\hat{\boldsymbol{Y}}_s = \left[\hat{f}_1(X_s), \dots, \hat{f}_N(X_s)\right]$.

For each sample, $\hat{\boldsymbol{Y}}_s$, we have **two cases**:

1. All of the functions agree with each other. $\longrightarrow$ **One Clique Case**
2. The functions can be split into two non-empty groups: those that output 1 and those that output 0. $\longrightarrow$ **Two Cliques Case**

One Clique Case

$$\mathbb{P}_{\mathcal{D}}(\hat{\boldsymbol{Y}}_s | \boldsymbol{e}) = e_{\mathcal{C}} + 1 + \sum_{k=1}^{|\mathcal{C}|} \left[(-1)^k \sum_{\substack{I \subseteq \mathcal{C} \\ |I| = k}} e_I\right]$$

where $\mathcal{C}$ contains all of the functions. That is:

$$\mathcal{C} = \{1, \dots, N\}$$

The probabilities here are computed in the same way as for the agreement rates.

Two Cliques Case

Either all function in clique 1 ($\mathcal{C}_1$) make an error and all functions in clique 2 ($\mathcal{C}_2$) are correct:

$$\mathbb{P}^1_{\mathcal{D}}(\hat{\boldsymbol{Y}}_s | \boldsymbol{e}) = e_{\mathcal{C}_1} + \sum_{k=1}^{|\mathcal{C}_2|} \left[(-1)^k \sum_{\substack{I \subseteq \mathcal{C}_2 \\ |I| = k}} e_{\{I \cup \mathcal{C}_1\}}\right]$$

Or, the opposite:

$$\mathbb{P}^2_{\mathcal{D}}\left(\hat{\boldsymbol{Y}}_s | \boldsymbol{e}\right) = e_{\mathcal{C}_2} + \sum_{k=1}^{|\mathcal{C}_1|} \left[(-1)^k \sum_{\substack{I \subseteq \mathcal{C}_1 \\ |I| = k}} e_{\{I \cup \mathcal{C}_2\}}\right]$$

The two events are **mutually exclusive** and so:

$$\mathbb{P}_{\mathcal{D}}(\hat{\boldsymbol{Y}}_s | \boldsymbol{e}) = \mathbb{P}^1_{\mathcal{D}}(\hat{\boldsymbol{Y}}_s | \boldsymbol{e}) + \mathbb{P}^2_{\mathcal{D}}(\hat{\boldsymbol{Y}}_s | \boldsymbol{e})$$

Regularization

We add a **prior**: $c(\boldsymbol{e}) = -\log L(\boldsymbol{e}) + \lambda \sum_{\mathcal{A} : |\mathcal{A}| \geq 2} \left(e_{\mathcal{A}} - \prod_{i \in \mathcal{A}} e_i\right)^2$

3. Experiments

1 NELL Data Set

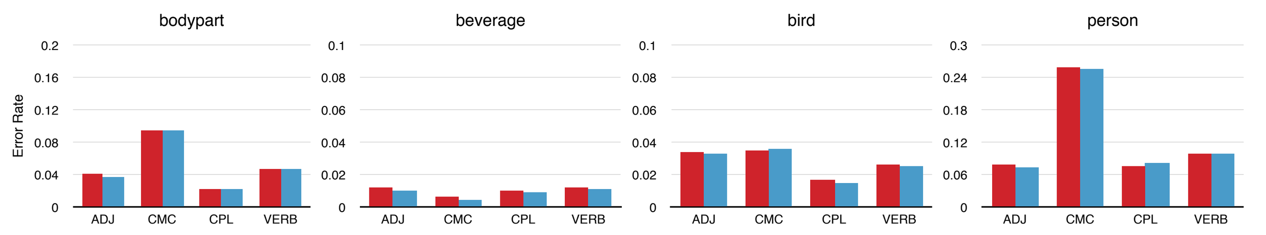
Task: *Predict whether a noun phrase (NP) belongs to a category (e.g. "city")*

4 logistic regression classifiers using different features:

ADJ: Adjectives that occur with the NP
CMC: Orthographic features of the NP
CPL: Phrases that occur with the NP
VERB: Verbs that appear with the NP

Category	# Examples
animal	20,733
beverage	18,932
bird	19,263
bodypart	21,840
city	21,778
disease	21,827
drug	20,452
fish	19,162
food	19,566
fruit	18,911
muscle	21,606
person	21,700
protein	21,811
river	21,723
vegetable	18,826

True error rates (estimated from labeled data) — **Error rates estimated from unlabeled data**



We report the **mean absolute deviation (MAD)** between:

- True error rates (estimated from labeled data)
- Error rates estimates from unlabeled data

3 functions under independence assumption: **2.82x10^{-2}**.
N functions without that assumption:

$\times 10^{-2}$	All Data Samples			50 Data Samples		
	Ind.	Pair.	All	Ind.	Pair.	All
AR	**0.49**	**0.31**	**0.29**	0.82	0.39	0.40
MLE	2.77	2.19	1.84	20.06	19.96	15.42
MAP	1.54	1.30	1.08	13.11	15.17	11.14

2 Brain Data Set

Task: *Find which of two 40 second long story passages corresponds to an unlabeled 40 second time series of fMRI neural activity*

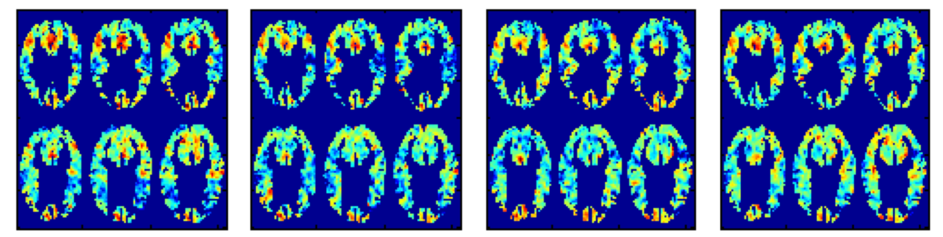
8 classifiers using different text representation:

- Number of letters in each word:
- Parts of speech:
- …

Passage #1: … They were hoping for a reason to fight Malfoy … / … 4 4 6 3 1 6 2 5 6 … / … PRP VBD VBG IN DT NN TO VB NNP …

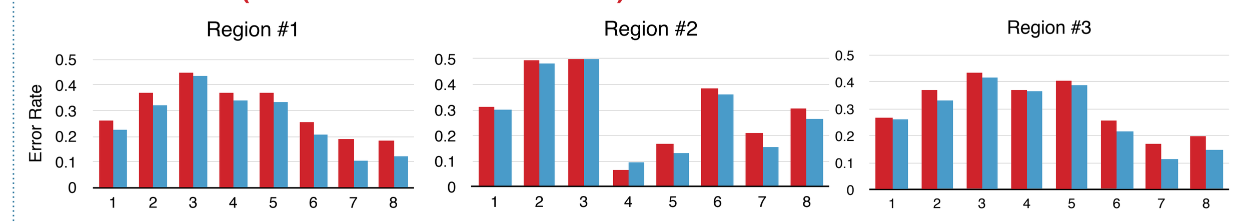
Passage #2: … Harry had heard Fred and George Weasley complain … / … 5 3 5 4 3 6 7 8 … / … NNP VBD VBN NNP CC NNP NNP VB …

Which passage corresponds to this fMRI recording?



1,000 labeled samples for 11 brain regions

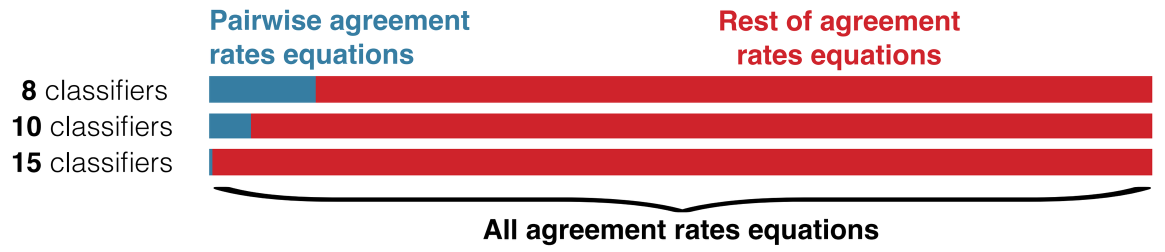
True error rates (estimated from labeled data) — **Error rates estimated from unlabeled data**



$\times 10^{-2}$	4 Classifiers			8 Classifiers		
	Ind.	Pair.	All	Ind.	Pair.	All
AR	10.97	6.60	6.50	**4.36**	**4.14**	**2.01**
MLE	10.60	8.34	7.64	32.02	12.33	4.50
MAP	9.61	18.19	11.16	27.95	18.60	7.26

The more functions we have, the better the estimates we obtain!

High order sample agreement rates are often bad estimates of the actual agreement rates



Runs 8 times faster for 8 classifiers and performs better!

$\times 10^{-2}$	Pairwise Agreement Rates			All Agreement Rates		
	Ind.	Pair.	All	Ind.	Pair.	All
MAD	**4.40**	**4.06**	**1.90**	4.36	4.14	2.01