Estimating Accuracy from Unlabeled Data

A Bayesian Approach

Emmanouil Antonios Platanios
e.a.platanios@cs.cmu.edu

Avinava Dubey
akdubey@cs.cmu.edu

Tom Mitchell
tom.mitchell@cs.cmu.edu

MACHINE LEARNING DEPARTMENT

1. Problem

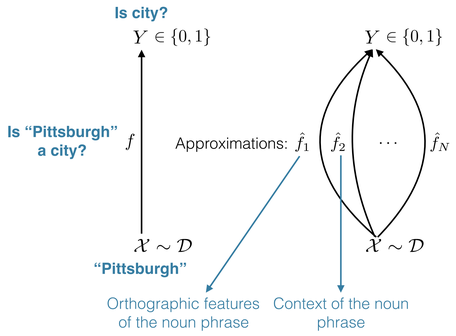
Using **only unlabeled data** we can measure **consistency** but not **correctness**. Therefore:

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**, $X_1, \ldots, X_S$, 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:

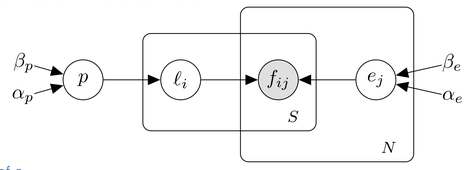
Error Rate: The probability over $\mathbb{P}(\mathcal{X}) = \mathcal{D}$ of disagreeing with the correct output label.

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

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

2. Approach

1 Bayesian Error Estimation



Prior probability of a label being positive ← $p \sim \text{Beta}(\alpha_p, \beta_p)$,

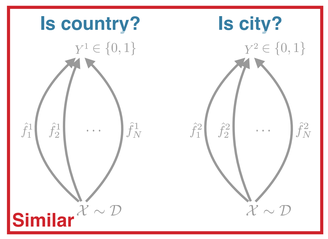
True label ← $\ell_i \sim \text{Bernoulli}(p)$, for $i = 1, \ldots, S$,

Function approximation **error rate** ← $e_j \sim \text{Beta}(\alpha_e, \beta_e)$, for $j = 1, \ldots, N$,

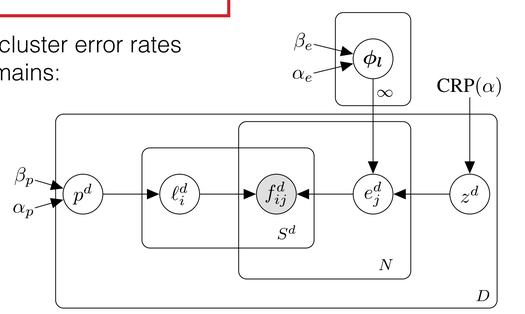
Observation (flipping based on error rate) ← $\hat{f}_{ij} = \begin{cases} \ell_i & \text{, with probability } 1 - e_j, \\ 1 - \ell_i & \text{, otherwise.} \end{cases}$

2 Coupled Bayesian Error Estimation

What about multiple domains?



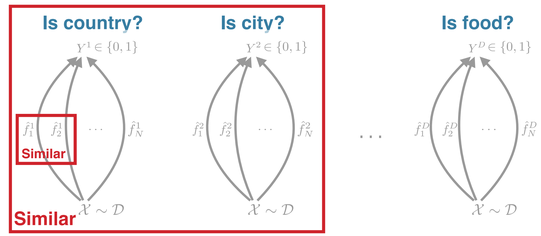
We can cluster error rates over domains:



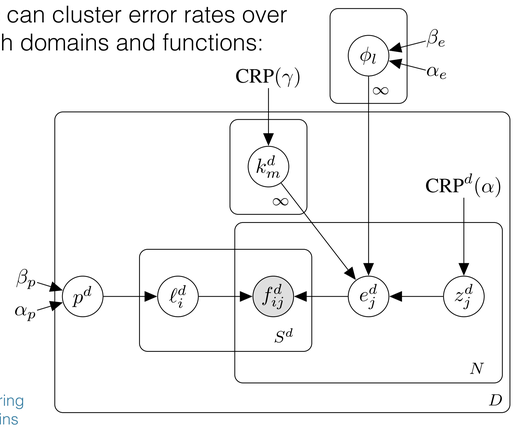
Clustering domains using a Dirichlet Process (DP) prior

$$\begin{aligned}
p^d &\sim \text{Beta}(\alpha_p, \beta_p), \text{ for } d = 1, \ldots, D, \\
\ell_i^d &\sim \text{Bernoulli}(p^d), \text{ for } i = 1, \ldots, S^d, \text{ and } d = 1, \ldots, D, \\
[\phi_l]_j &\sim \text{Beta}(\alpha_e, \beta_e), \text{ for } j = 1, \ldots, N, \text{ and } l = 1, \ldots, \infty, \\
z^d &\sim \text{CRP}(\alpha), \text{ for } d = 1, \ldots, D, \\
e_j^d &= [\phi_{z^d}]_j, \text{ for } j = 1, \ldots, N, \text{ and } d = 1, \ldots, D, \\
\hat{f}_{ij}^d &= \begin{cases} \ell_i^d & \text{, with probability } 1 - e_j^d, \\ 1 - \ell_i^d & \text{, otherwise.} \end{cases}
\end{aligned}$$

3 Hierarchical Coupled Bayesian Error Estimation



We can cluster error rates over both domains and functions:



Clustering domains using a Hierarchical Dirichlet Process (HDP) prior

$$\begin{aligned}
p^d &\sim \text{Beta}(\alpha_p, \beta_p), \text{ for } d = 1, \ldots, D, \\
\ell_i^d &\sim \text{Bernoulli}(p^d), \text{ for } i = 1, \ldots, S^d, \text{ and } d = 1, \ldots, D, \\
\phi_l &\sim \text{Beta}(\alpha_e, \beta_e), \text{ for } l = 1, \ldots, \infty, \\
k_m^d &\sim \text{CRP}(\gamma), \text{ for } d = 1, \ldots, D, \text{ and } m = 1, \ldots, \infty, \\
z_j^d &\sim \text{CRP}^d(\alpha), \text{ for } d = 1, \ldots, D, \text{ and } j = 1, \ldots, N, \\
e_j^d &= \phi_{k_{z_j^d}^d}, \text{ for } j = 1, \ldots, N, \text{ and } d = 1, \ldots, D, \\
\hat{f}_{ij}^d &= \begin{cases} \ell_i^d & \text{, with probability } 1 - e_j^d, \\ 1 - \ell_i^d & \text{, otherwise.} \end{cases}
\end{aligned}$$

Our methods implicitly use agreement rates in order to estimate function error rates. We are using the agreement between the function outputs and the true underlying labels in order to infer both the error rates of our functions and those labels, jointly.

3. Experiments

We report the **error mean squared error ($\text{MSE}_{\text{error}}$)** between:

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

the **target label mean absolute deviation ($\text{MAD}_{\text{label}}$ —** equivalent to **accuracy)**.

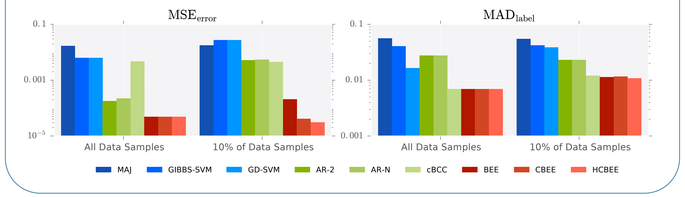
1 NELL Data Set

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



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*

11 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