Lecture 2: Bootstrapping

Steven Abney

Presenter: Zhihang Dong;
Adv. Natural Lang. Proc.

April 25, 2018

Table of contents

Motivation

Bootstrapping: Research Question

Given a small set of labeled data and a large set of unlabeled data,

The **goal** of bootsrapping is to **induce a classifier**, such classifier is necessary because

1. paucity of labeled data

... in NLP

Bootstrapping: Research Question

Given a small set of labeled data and a large set of unlabeled data,

The **goal** of bootsrapping is to **induce a classifier**, such classifier is necessary because

1. paucity of labeled data
2. plenitude of unlabeled data

... in NLP

Boostrapping: Settings

Given the unlabeled natural language data $\mathcal{X}$, our purposes are

- given a ”trained” or ”labeled” natural language data $\mathcal{X}'$ and their corresponding labels $\mathcal{L}'$

Boostrapping: Settings

Given the unlabeled natural language data $\mathcal{X}$, our purposes are

- given a "trained" or "labeled" natural language data $\mathcal{X}'$ and their corresponding labels $\mathcal{L}'$
- our goal is to find a function $\mathcal{F}$, that produces labels for the unlabeled NL data: $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{L}$

Boostrapping: Settings

Given the unlabeled natural language data $\mathcal{X}$, our purposes are

- given a "trained" or "labeled" natural language data $\mathcal{X}'$ and their corresponding labels $\mathcal{L}'$
- our goal is to find a function $\mathcal{F}$, that produces labels for the unlabeled NL data: $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{L}$
- ... such function map instances to labels through a space $\mathcal{S}$ of classifiers...

Boostrapping: Conditional Independence Assumptions (I)

What is conditional independence?

What is the conditional independence assumptions in the context of bootstrapping algorithms?

Boostrapping: Conditional Independence Assumptions (II)

We can divide the **Yarowsky algorithm** into two smaller algorithms:

- A *decision list algorithm* that identify other reliable collocations, which calculates the probability Pr(Sense | Collocation); this produces a decision list that is ranked by log-likelihood ratio:

Boostrapping: Conditional Independence Assumptions (II)

We can divide the **Yarowsky algorithm** into two smaller algorithms:

- A *decision list algorithm* that identify other reliable collocations, which calculates the probability Pr(Sense | Collocation); this produces a decision list that is ranked by log-likelihood ratio:
- $$\log\left(\frac{\Pr(\text{Sense}_A|\text{Collocation}_i)}{\Pr(\text{Sense}_B|\text{Collocation}_i)}\right)$$

Boostrapping: Conditional Independence Assumptions (II)

We can divide the **Yarowsky algorithm** into two smaller algorithms:

- A *decision list algorithm* that identify other reliable collocations, which calculates the probability Pr(Sense | Collocation); this produces a decision list that is ranked by log-likelihood ratio:
- $$\log\left(\frac{\Pr(\text{Sense}_A|\text{Collocation}_i)}{\Pr(\text{Sense}_B|\text{Collocation}_i)}\right)$$
- it uses only the most reliable piece of evidence rather than the whole matching collocation set

Boostrapping: Conditional Independence Assumptions (II)

We can divide the **Yarowsky algorithm** into two smaller algorithms:

- A *decision list algorithm* that identify other reliable collocations, which calculates the probability Pr(Sense | Collocation); this produces a decision list that is ranked by log-likelihood ratio:
- $$\log\left(\frac{\Pr(\text{Sense}_A|\text{Collocation}_i)}{\Pr(\text{Sense}_B|\text{Collocation}_i)}\right)$$
- it uses only the most reliable piece of evidence rather than the whole matching collocation set
- Then, a *smoothing algorithm* to avoid 0 values

Boostrapping: Conditional Independence Assumptions (III)

Another work is followed by (Blum and Mitchell, 1998):

- BM propose a conditional independence assumption to account for the efficacy of their algorithm, called **co-training**

Boostrapping: Conditional Independence Assumptions (III)

Another work is followed by (Blum and Mitchell, 1998):

- BM propose a conditional independence assumption to account for the efficacy of their algorithm, called **co-training**
- suggest that the Yarowsky algorithm is a special case of the co-training algorithm

Boostrapping: Conditional Independence Assumptions (III)

Another work is followed by (Blum and Mitchell, 1998):

- BM propose a conditional independence assumption to account for the efficacy of their algorithm, called **co-training**
- suggest that the Yarowsky algorithm is a special case of the co-training algorithm
- the author identifies several flaws of B&M's statements, which we will discuss a bit later

View Independence vs. Rule Independence

View Independence (I)

In **view independence**, each instance x consists of two "views" x_1, x_2. We can take this as the assumption of functions X_1 and X_2 such that $X_1(x) = x_1$ and $X_2(x) = x_2$

Definition 1 A pair of views x_1, x_2 satisfy view independence if:

- $Pr[X_1 = x_1 | X_2 = x_2; Y = y] = Pr[X_1 = x_1 | Y = y]$

View Independence (I)

In **view independence**, each instance x consists of two "views" x_1, x_2. We can take this as the assumption of functions X_1 and X_2 such that $X_1(x) = x_1$ and $X_2(x) = x_2$

Definition 1 A pair of views x_1, x_2 satisfy view independence if:

- $Pr[X_1 = x_1 | X_2 = x_2; Y = y] = Pr[X_1 = x_1 | Y = y]$
- $Pr[X_2 = x_2 | X_1 = x_1; Y = y] = Pr[X_2 = x_2 | Y = y]$

View Independence (I)

In **view independence**, each instance x consists of two "views" x_1, x_2. We can take this as the assumption of functions X_1 and X_2 such that $X_1(x) = x_1$ and $X_2(x) = x_2$

Definition 1 A pair of views x_1, x_2 satisfy view independence if:

- $Pr[X_1 = x_1 | X_2 = x_2; Y = y] = Pr[X_1 = x_1 | Y = y]$
- $Pr[X_2 = x_2 | X_1 = x_1; Y = y] = Pr[X_2 = x_2 | Y = y]$
- A classification problem instance satisfies view independence if all pairs x_1, x_2 satisfy view independence

View Independence (II)

The first definition can be expanded into a generalization that:

Definition 2 A pair of rules $F \in \mathcal{H}_1$, $G \in \mathcal{H}_2$ satisfies rule independence just in case, for all $u; v; y$

- $Pr[F = u | G = v; Y = y] = Pr[F = u | Y = y]$

View Independence (II)

The first definition can be expanded into a generalization that:

Definition 2 A pair of rules $F \in \mathcal{H}_1$, $G \in \mathcal{H}_2$ satisfies rule independence just in case, for all $u; v; y$

- $Pr[F = u | G = v; Y = y] = Pr[F = u | Y = y]$
- A classification problem instance satisfies rule independence just in case **all opposing-view rule pairs** satisfy rule independence

View Independence (II)

The first definition can be expanded into a generalization that:

Definition 2 A pair of rules $F \in \mathcal{H}_1$, $G \in \mathcal{H}_2$ satisfies rule independence just in case, for all $u; v; y$

- $Pr[F = u | G = v; Y = y] = Pr[F = u | Y = y]$
- A classification problem instance satisfies rule independence just in case **all opposing-view rule pairs** satisfy rule independence
- assume a set of features $\mathcal{F} : \{\mathcal{H}_1, \mathcal{H}_2\}$, rule independence reduces to the *Naive Bayes independence assumption*

View Independence (III)

Theorem 1:

View independence implies rule independence

Agreement Rates (i): Contrary Argument

Blum and Mitchell's paper suggests that rules that agree on unlabeled instances are useful in bootstrapping:

Definition 3 The **agreement rate** between rules F and G is:

$$Pr[F = G | F; G \neq \perp]$$

- ... does *not* explicitly search for rules with good agreement

Agreement Rates (i): Contrary Argument

Blum and Mitchell's paper suggests that rules that agree on unlabeled instances are useful in bootstrapping:

Definition 3 The **agreement rate** between rules F and G is:

$$Pr[F = G | F; G \neq \perp]$$

- ... does *not* explicitly search for rules with good agreement
- ... does *not* play any direct role in the learnability proof given in the B& M paper

Agreement Rates (i): Contrary Argument

Blum and Mitchell's paper suggests that rules that agree on unlabeled instances are useful in bootstrapping:

Definition 3 The **agreement rate** between rules F and G is:

$$Pr[F = G | F; G \neq \perp]$$

- ... does *not* explicitly search for rules with good agreement
- ... does *not* play any direct role in the learnability proof given in the B& M paper
- if view independence is satisfied, then the agreement rate between opposing-view rules F and G upper bounds the error of F/G

Agreement Rates (ii): Minimizer Inequality

Theorem 2

For all $F \in \mathcal{H}_1$, $G \in \mathcal{H}_2$ that satisfy rule independence and are nontrivial predictors in the sense that $\min_u Pr[F = u] > Pr[F \neq G]$, one of the following inequalities holds:

- $$Pr[F \neq Y] \leq Pr[F \neq G]$$

Agreement Rates (ii): Minimizer Inequality

Theorem 2

For all $F \in \mathcal{H}_1$, $G \in \mathcal{H}_2$ that satisfy rule independence and are nontrivial predictors in the sense that $\min_u Pr[F = u] > Pr[F \neq G]$, one of the following inequalities holds:

- $$Pr[F \neq Y] \leq Pr[F \neq G]$$
- $$Pr[\hat{F} \neq Y] \leq Pr[F \neq G]$$

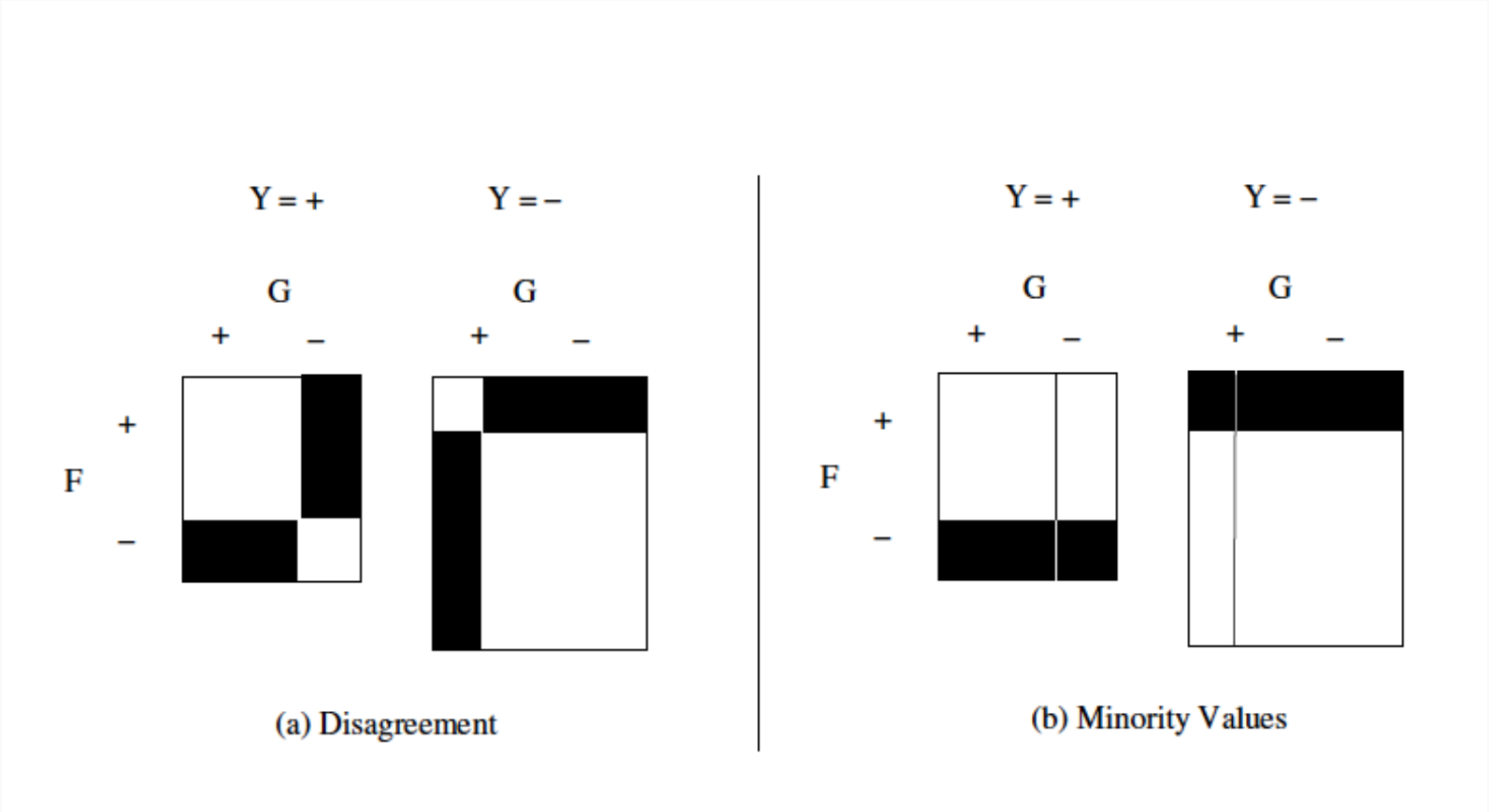
Agreement Rates (ii): Minimizer Inequality

Theorem 2

For all $F \in \mathcal{H}_1$, $G \in \mathcal{H}_2$ that satisfy rule independence and are nontrivial predictors in the sense that $\min_u Pr[F = u] > Pr[F \neq G]$, one of the following inequalities holds:

- $$Pr[F \neq Y] \leq Pr[F \neq G]$$
- $$Pr[\hat{F} \neq Y] \leq Pr[F \neq G]$$
- If F agrees with G on all but ε unlabeled instances, then either F or $\hat{F}$ predicts Y with error no greater than ε

Agreement Rate (iii): Theorem 2 w.r.t. Upper Bound



Rule Independence: Problems

Rule independence is a very strong assumption:

- Let F and G be arbitrary rules based on independent views

Rule Independence: Problems

Rule independence is a very strong assumption:

- Let F and G be arbitrary rules based on independent views
- The precision of a rule F is defined to be $Pr[Y = +|F = +]$

Rule Independence: Problems

Rule independence is a very strong assumption:

- Let F and G be arbitrary rules based on independent views
- The precision of a rule F is defined to be $Pr[Y = +|F = +]$
- If rule independence holds, we can compute the precision of every other rule given unlabeled data and knowledge of the size of the target concept with the precision of one rule

Rule Independence: Problems

Rule independence is a very strong assumption:

- Let F and G be arbitrary rules based on independent views
- The precision of a rule F is defined to be $Pr[Y = +|F = +]$
- If rule independence holds, we can compute the precision of every other rule given unlabeled data and knowledge of the size of the target concept with the precision of one rule
- this is easily proven given the knowledge from STAT 001... (Probability lesson 1)

Empirical Exam: Data

The task is to classify names in text as person, location, or organization. There is an unlabeled training set containing 89,305 instances, and a labeled test set containing 289 persons, 186 locations, 402 organizations, and 123 “other”, for a total of 1,000 instances.

Empirical Exam: Results

F	Co-training	Yarowsky	Truth
`M:chairman`	-12.7	.068	.030
`X:Company-of`	10.2	.979	.989
`X:court-in`	-.183	1.00	.981
`X:Company-in`	75.7	1.00	.986
`X:firm-in`	-9.94	.952	.949
`X:%-in`	-15.2	.875	.192
`X:meeting-in`	-2.25	1.00	.753

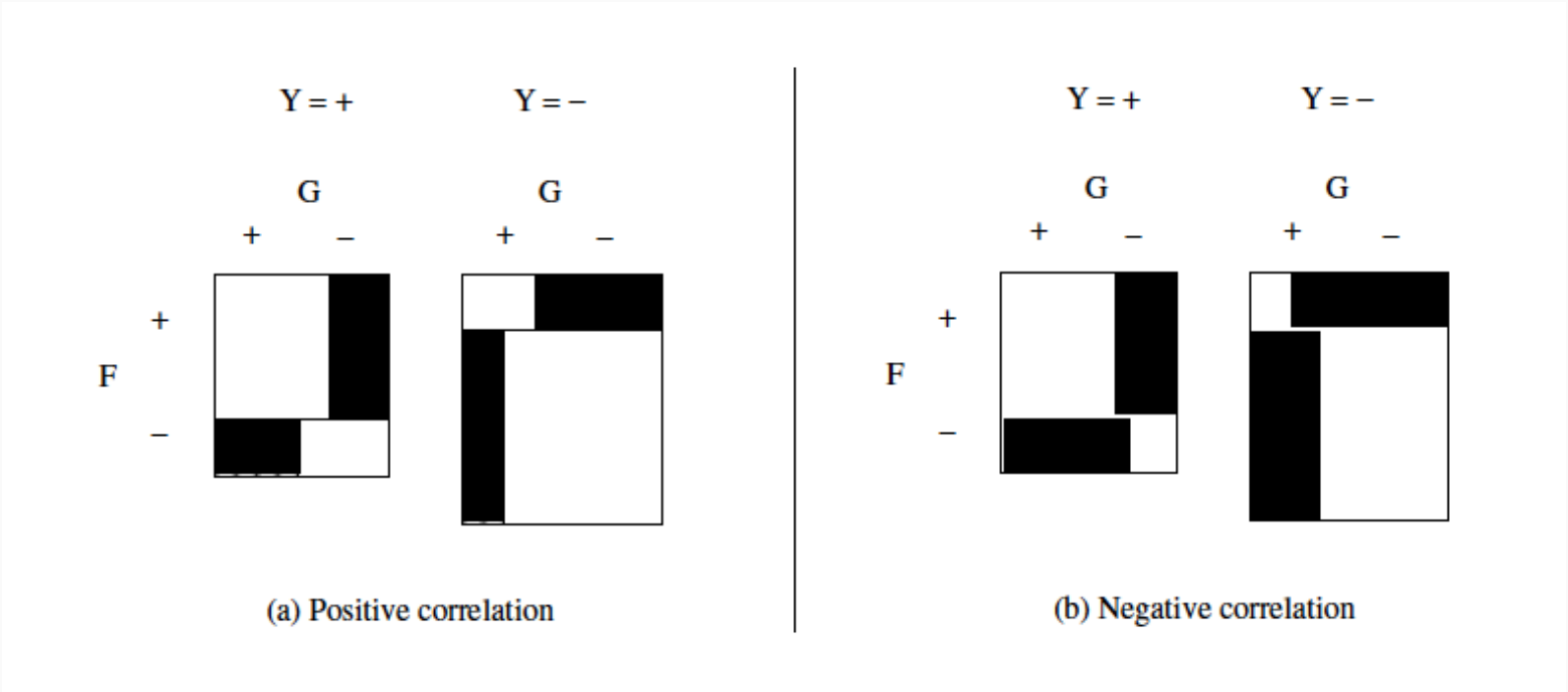
Table 1: Some data

Assumption Relaxed

How to relax the assumption

- There are two ways in which the data can diverge from conditional independence: the rules may either be *positively or negatively correlated*, given the class value
- If the rules are negatively correlated, then their disagreement (shaded in figure 3) is larger than if they are conditionally independent
- … but the data is positively correlated

Positive Correlation vs. Negative Correlation



A Weaker Assumption

Let us quantify the amount of deviation from conditional independence. We define the conditional dependence of F and G given Y = y to be

$$d_y = \frac{\sum_{u,v} |Pr[G = v|Y = y; F = u] - Pr[G = v|Y = y]|}{2}$$

If F and G are conditionally independent, then $d_y = 0$

Definition 4 Rules F and G satisfy weak rule dependence if, for $y \in \{+, -\}$

$$d_y \leq p_2 \frac{q_1 - p_1}{2p_1 q_1}$$

Let's rewrite Theorem 2!

Theorem 3

For all $F \in \mathcal{H}_1$, $G \in \mathcal{H}_2$ that satisfy **weak rule dependence and are nontrivial predictors** in the sense that $\min_u Pr[F = u] > Pr[F \neq G]$, one of the following inequalities holds:

- $$Pr[F \neq Y] \leq Pr[F \neq G]$$

Let's rewrite Theorem 2!

Theorem 3

For all $F \in \mathcal{H}_1$, $G \in \mathcal{H}_2$ that satisfy **weak rule dependence and are nontrivial predictors** in the sense that $\min_u Pr[F = u] > Pr[F \neq G]$, one of the following inequalities holds:

- $$Pr[F \neq Y] \leq Pr[F \neq G]$$
- $$Pr[\hat{F} \neq Y] \leq Pr[F \neq G]$$

A New Bound!

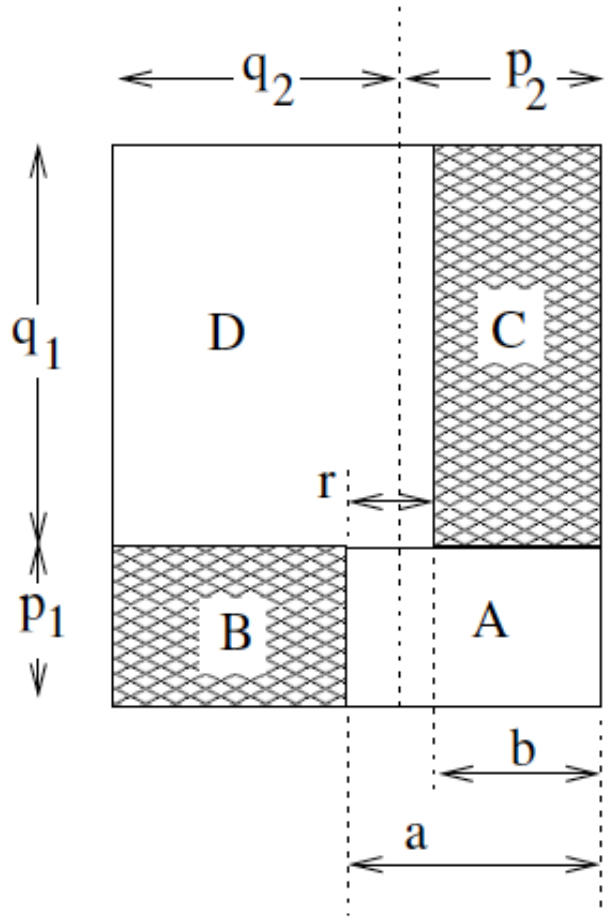


Figure 3: Positive correlation, $Y = +$.

Greedy Agreement Algorithm and Yarowsky Algorithm

Greedy Agreement Algorithm (i): Specifics

The algorithm can be written as:

```
Input: seed rules F, G
loop
for each atomic rule H G' = G + H
evaluate cost of (F,G')
keep lowest-cost G'
if G' is worse than G,
quit
swap F, G'
```

GAA (ii): One another generalization

We explain this by:

- begins with two seed rules, one for each view
- for each iteration, each possible extension to one of the rules is considered and scored
- The best one is kept, and attention shifts to the other rule.

The cost of a classifier pair $(F; G)$ explains a new ground for the generalization of Theorem 2

GAA(iii): Theorem

Theorem 4: If view independence is satisfied, and if F and G are rules based on different views, then one of the following holds:

- $Pr[F \neq Y | F \neq \perp] \leq \frac{\delta}{\mu - \delta}$

GAA(iii): Theorem

Theorem 4: If view independence is satisfied, and if F and G are rules based on different views, then one of the following holds:

- $Pr[F \neq Y | F \neq \perp] \leq \frac{\delta}{\mu - \delta}$
- $Pr[\hat{F} \neq Y | F \neq \perp] \leq \frac{\delta}{\mu - \delta}$

GAA(iii): Theorem

Theorem 4: If view independence is satisfied, and if F and G are rules based on different views, then one of the following holds:

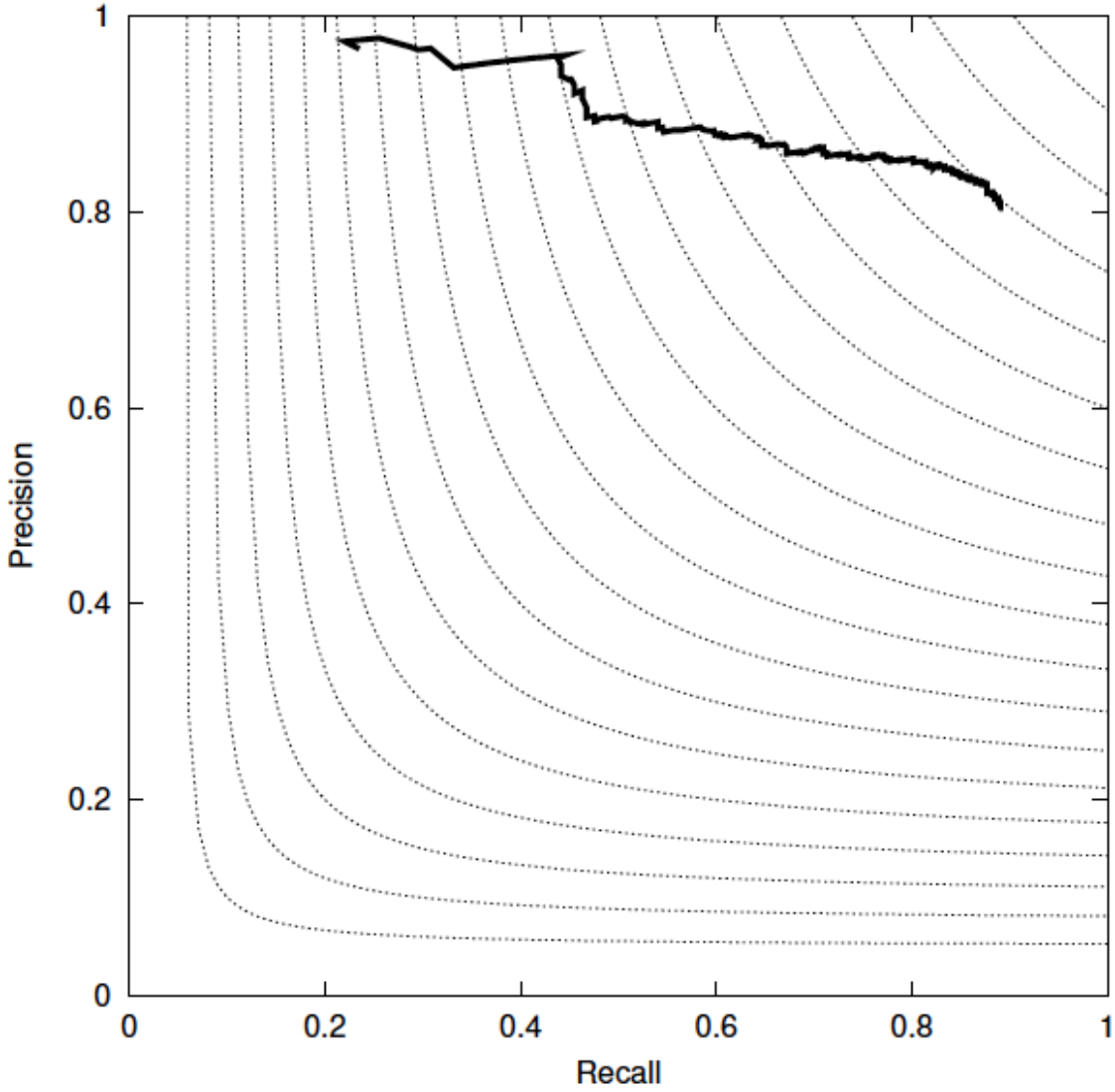
- $Pr[F \neq Y | F \neq \perp] \leq \frac{\delta}{\mu - \delta}$
- $Pr[\hat{F} \neq Y | F \neq \perp] \leq \frac{\delta}{\mu - \delta}$
- where $\delta = Pr[F \neq G | F, G \neq \perp]$, and $\mu = \min_u Pr[F = u | F \neq \perp]$

GAA(iii): Theorem

Theorem 4: If view independence is satisfied, and if F and G are rules based on different views, then one of the following holds:

- $Pr[F \neq Y | F \neq \perp] \leq \frac{\delta}{\mu - \delta}$
- $Pr[\hat{F} \neq Y | F \neq \perp] \leq \frac{\delta}{\mu - \delta}$
- where $\delta = Pr[F \neq G | F, G \neq \perp]$, and $\mu = \min_u Pr[F = u | F \neq \perp]$
- In other words, for a given binary rule F, a pessimistic estimate of the number of errors made by F is $\delta/(\mu - \delta)$ times the number of instances labeled by F, plus the number of instances left unlabeled by F

Greedy Agreement Algorithm (iv): Viz



Yarowsky Algorithm (i) : Justification

Yarowsky did not give a justification, but let's give one...

Let F represents an atomic rule under consideration, and G represents the current classifier. Recall that Y_ℓ is the set of instances whose true label is ℓ, and G_ℓ is the set of instances$\{x : G(x) = \ell\}$ We write G_* for the set of instances labeled by the current classifier, that is,$\{x : G(x) \neq \perp\}$

So let's provide the grounds for precision independence and balanced errors assumptions...

Yarowsky Algorithm (ii) : Assumptions

Definition 5

Candidate rule F_ℓ and classifier G satisfy precision independence just in case

$$P(Y_\ell|F_\ell, G_*) = P(Y_\ell|F_\ell)$$

A bootstrapping problem instance satisfies precision independence just in case all rules G and all atomic rules F_ℓ that nontrivially overlap with G (both $F_\ell \cap G_*$ and $F\ell - G_*$are nonempty) satisfy precision independence.

In fact, it is only "half" an independence assumption—for precision independence, it is not necessary that the conditional on G_* is equal.

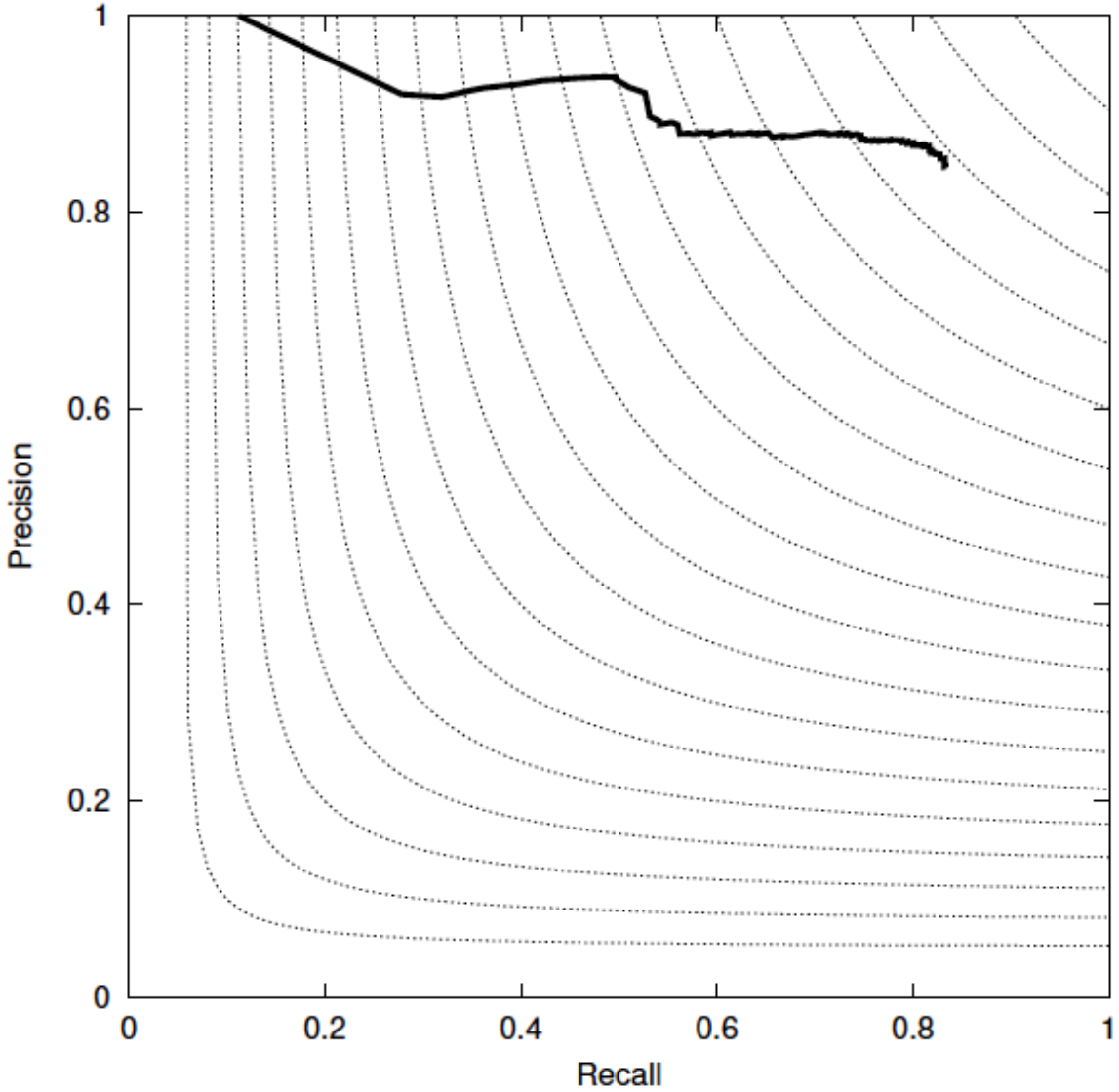
The second assumption is that classifiers make balanced errors. That is:

$$P(Y_\ell, G_{\hat{\ell}}|F_\ell) = P(Y_{\hat{\ell}}, G_\ell|F_\ell)$$

Yarowsky Algorithm (iii) : Assumptions

Theorem 5 If the assumptions of precision independence and balanced errors are satisfied, then the Yarowsky algorithm with threshold θ obtains a final classifier whose precision is at least μ. Moreover, recall is bounded below by $N_t\theta/N_\ell$, a quantity which increases at each round.

Yarowsky Algorithm (iv)



Questions?

References i