PU-learning disjunctive concepts in ILP

Hendrik Blockeel
Dept. of Computer Science, KU Leuven

Motivation

- "Relational grounded language learning" (Becerra-Bonache et al., IDA 2015, ECAI 2016)
- Learn meaning of phrases ("n-grams") from occurrences in some context
- **Mentioned** ⇒ **occurs**, but **occurs** ⇏ **mentioned**

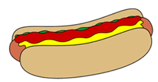
Data set from Zitnick et al., 2013

[$start, mike, is, kicking, the, ball, $stop]

[object(o1), sky(o1, sun), color(o1, yellow), size(o1,big), ..., object(o3), human(o3,boy), pose(o3,pose2), expression(o3,happy), object(o4), human(o4,girl), pose(o4,pose3), expression(o4,surprised), ..., object(o6), clothing(o6,glasses), color(o6,violet), object(o7), toy(o7,ball), sport(o7,soccer), act(o3,wear,o6), ...]

Motivation

- Earlier work: "Meaning" of an n-gram = least general generalization under subsumption (lgg, Plotkin 1970) of all contexts where it occurred. Intuitively: "maximal common pattern"
  - E.g.: "mike" -> object(X), human(X,boy), pose(X, _), expression(X,_)
  - E.g. "mike is wearing a hat" -> object(A),human(A,c_boy),pose(A,_),expression(A,_),object(D),clothing(D,c_hat),style(D,_),act(A,c_wear,D)
- This does not allow for disjunctive concepts
- Those are more common than expected! E.g.: "dog"



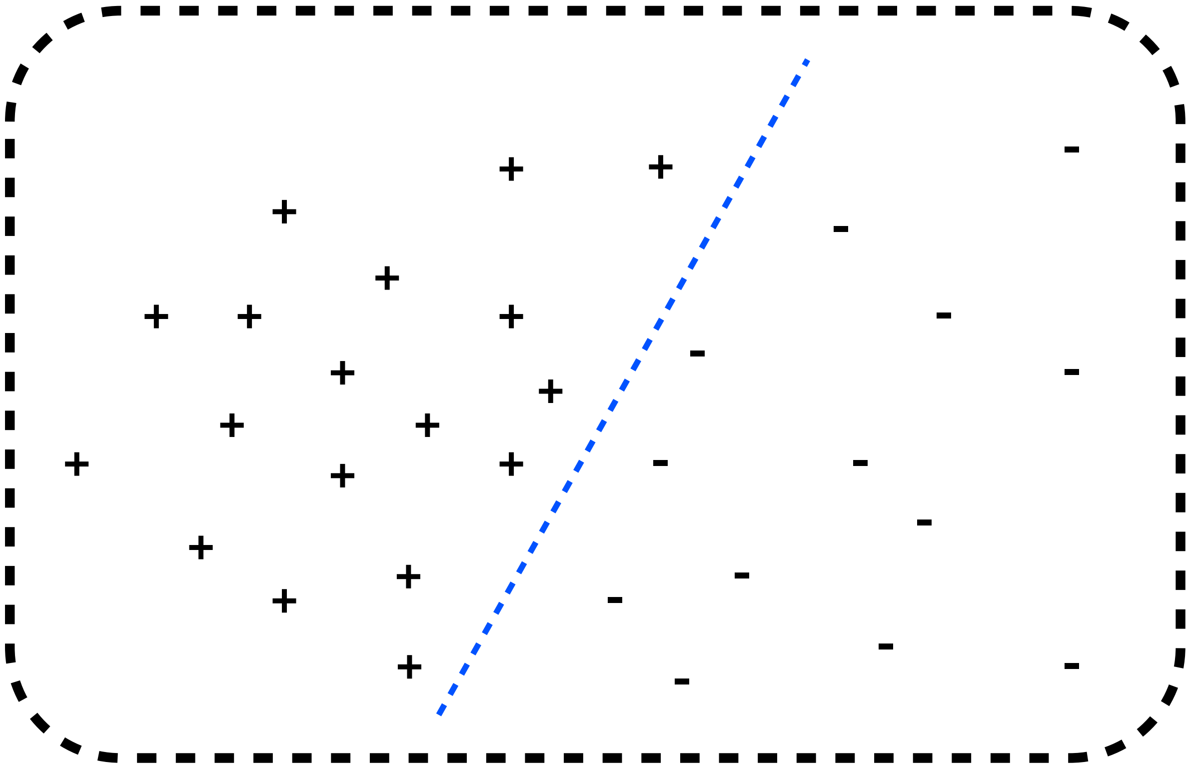
[$start, mike, is, kicking, the, ball, $stop]

[object(o1), sky(o1, sun), color(o1, yellow), size(o1,big), …, object(o3), human(o3,boy), pose(o3,pose2), expression(o3,happy), object(o4), human(o4,girl), pose(o4,pose3), expression(o4,surprised), …, object(o6), clothing(o6,glasses), color(o6,violet), object(o7), toy(o7,ball), sport(o7,soccer), act(o3,wear,o6), …]

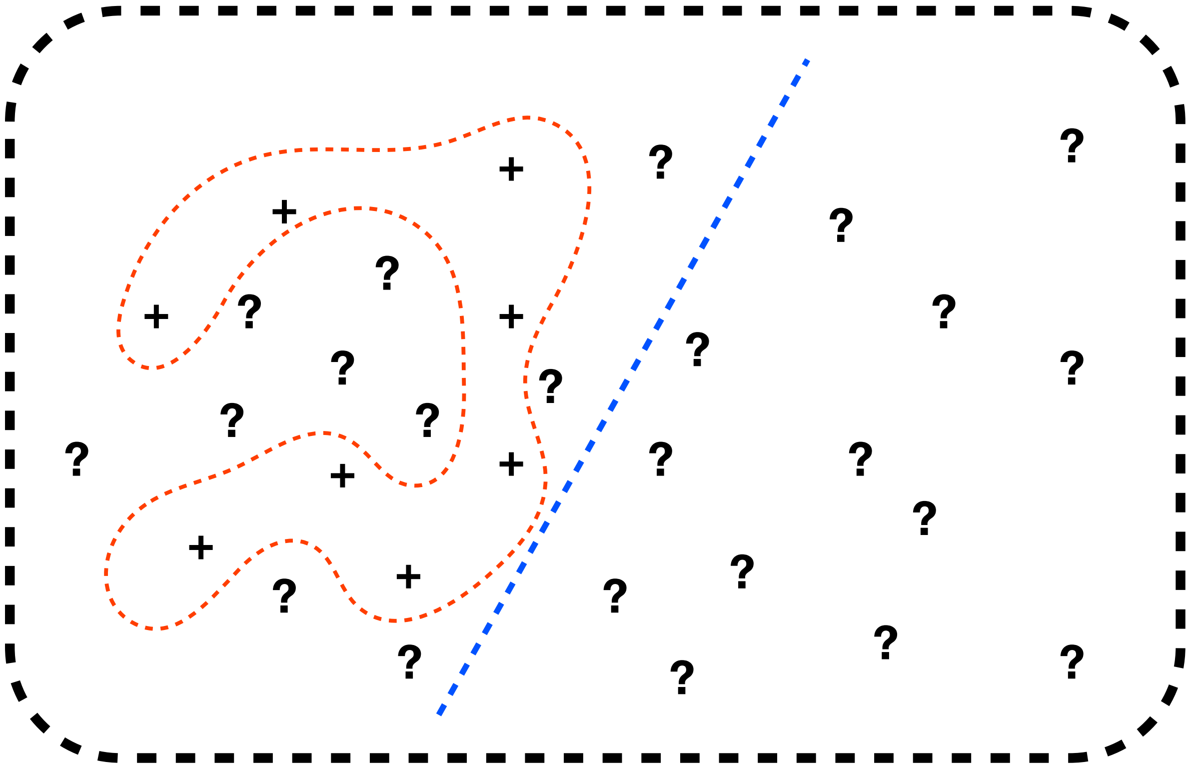
PU-learning

- “Normal” supervised concept learning: learn from positive and negative data
  - Ideal rule set covers **all** positives and **no** negatives
- PU-learning: Learn from positive and unlabeled data
  - Ideal rule set covers **all** positives and **some** unlabeled (we don’t know which, nor how many)

Supervised learning



PU learning



PU-learning

- Recall: **Mentioned ⇒ occurs**, but **occurs** ⇏ **mentioned**
- Viewing mentions as the labels, this makes our setting *PU-learning*
    - “Cat” in sentence labels the context as positive for cat
    - No “cat” in sentence does *not* label the context as negative, just unlabeled for cat

Elkan & Noto's result

- Elkan & Noto (2008) observed:
  - Let L = "x is labeled as positive", +/- = "x is positive/negative"
  - **Assume** positives are labeled completely at random: P(L| +, x) = P(L | +) = k
  - **Then** P(L|x) = P(L|+) P(+|x) + P(L|-) P(-|x) = c P(+|x) + 0 P(-|x) = k P(+|x)
  - From a probabilistic classifier that predicts L, we can derive a probabilistic classifier that predicts +, *if we know k*
  - The former can be learned in a supervised manner
- PU-learning reduces to supervised learning under the mentioned assumptions
- Ways to **estimate k** have been proposed (e.g. Bekker & Davis, ILP 2017)

Weakening Elkan & Noto's "constant c" assumption

- Learning disjunctive concepts with constant c = P(label|pos) : see Bekker & Davis, ILP 2017 (previous talk)
- But "constant c" is not realistic in our setting
- e.g. P("dog" | dog occurs) ≠ P("dog" | hot dog occurs)

  538/1706=0.315

  134/769=0.174

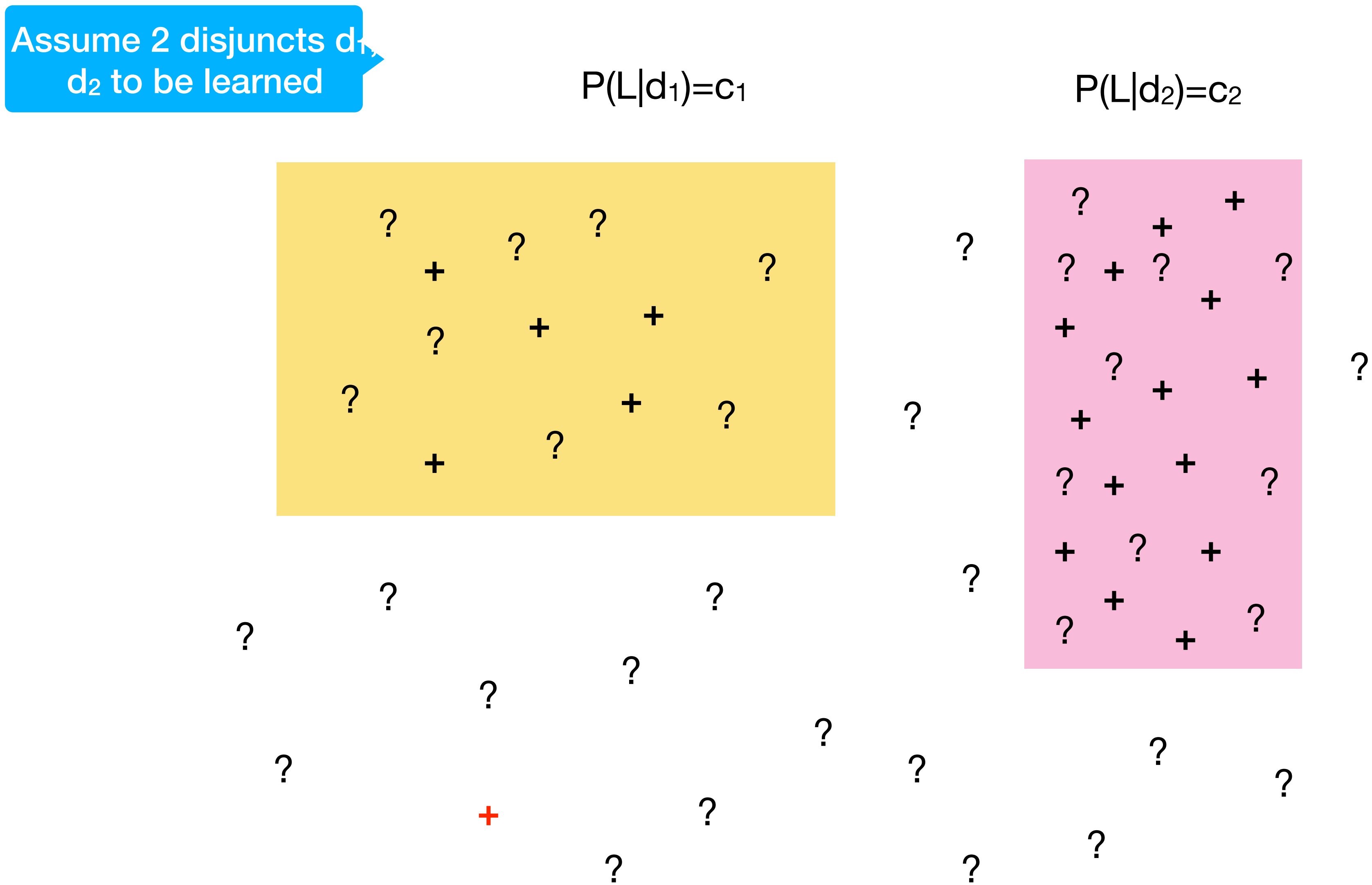
- More "noteworthy" things are more likely to be mentioned

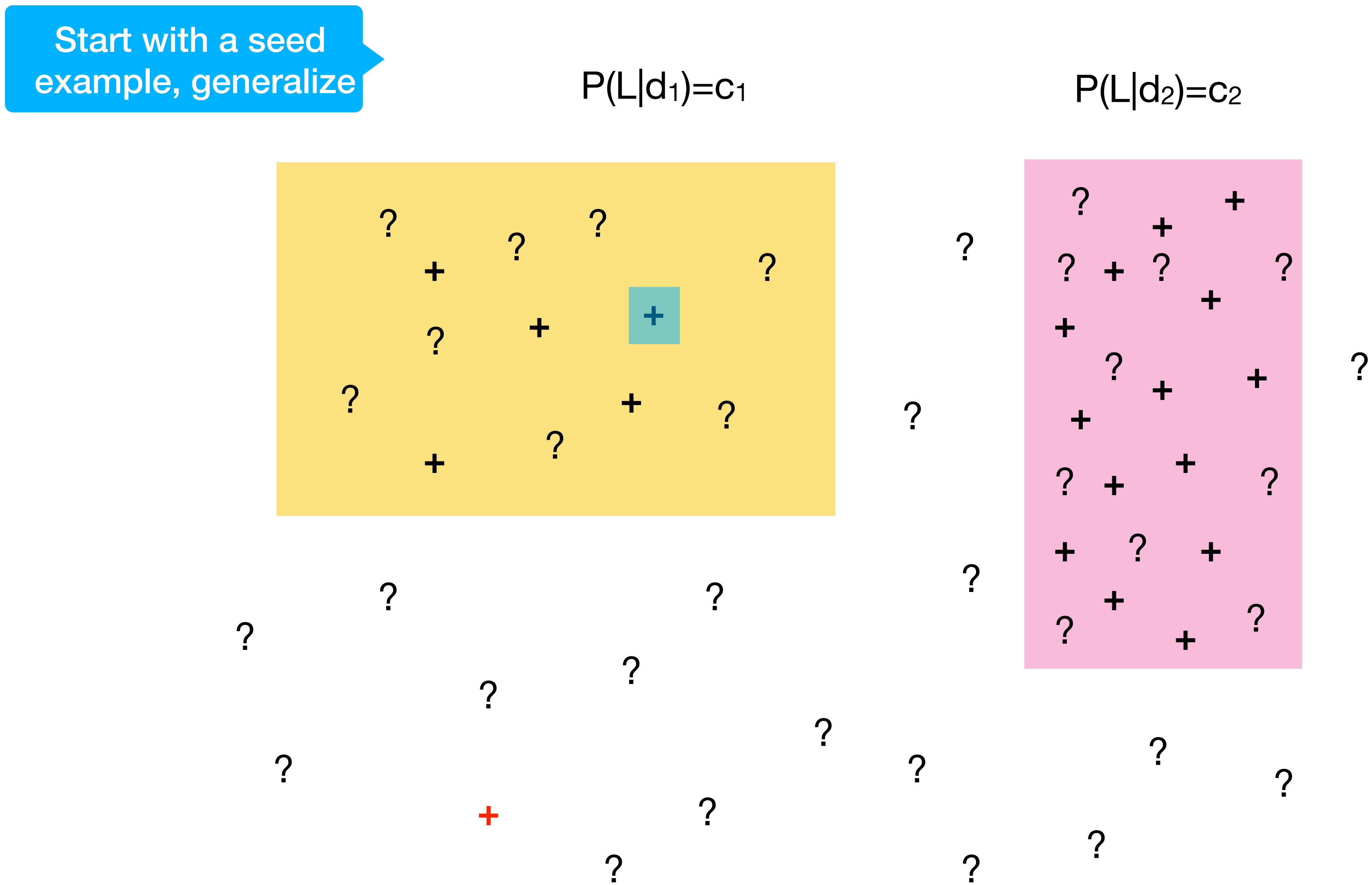
PU-learning disjunctive concepts

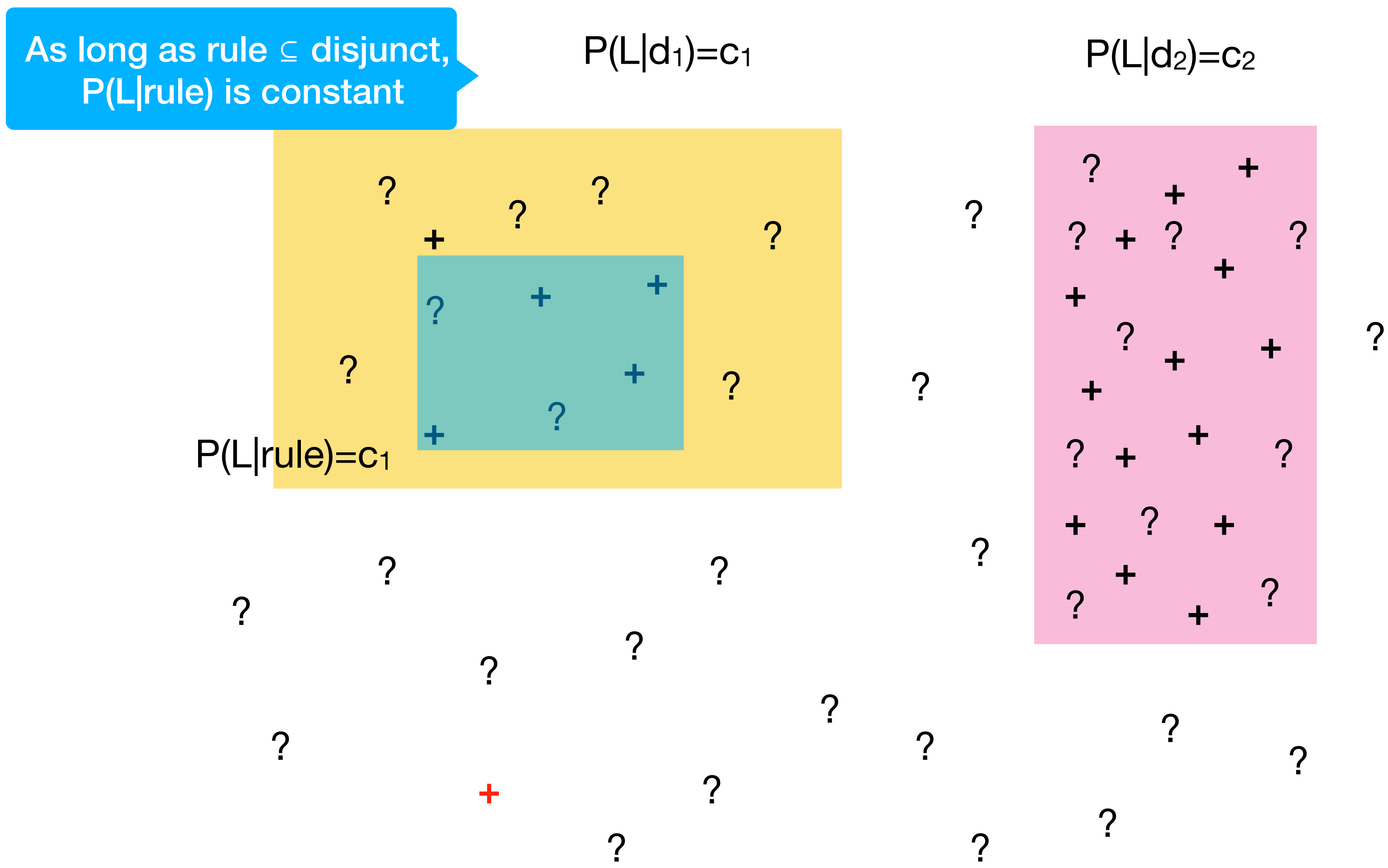
- Consider the following setting:
  - The concept is disjunctive : $d_1 \vee d_2 \vee \ldots \vee d_k$
  - **Assumption**: Within each disjunct d_i, each x has the same probability c_i of being labeled, but it is possible that $c_i \neq c_j$ for $i \neq j$
- How to PU-learn in this setting?
- Estimate c_i for a given disjunct d_i? Then we first need to know d_i! Catch-22. Need to learn d_i and c_i simultaneously.

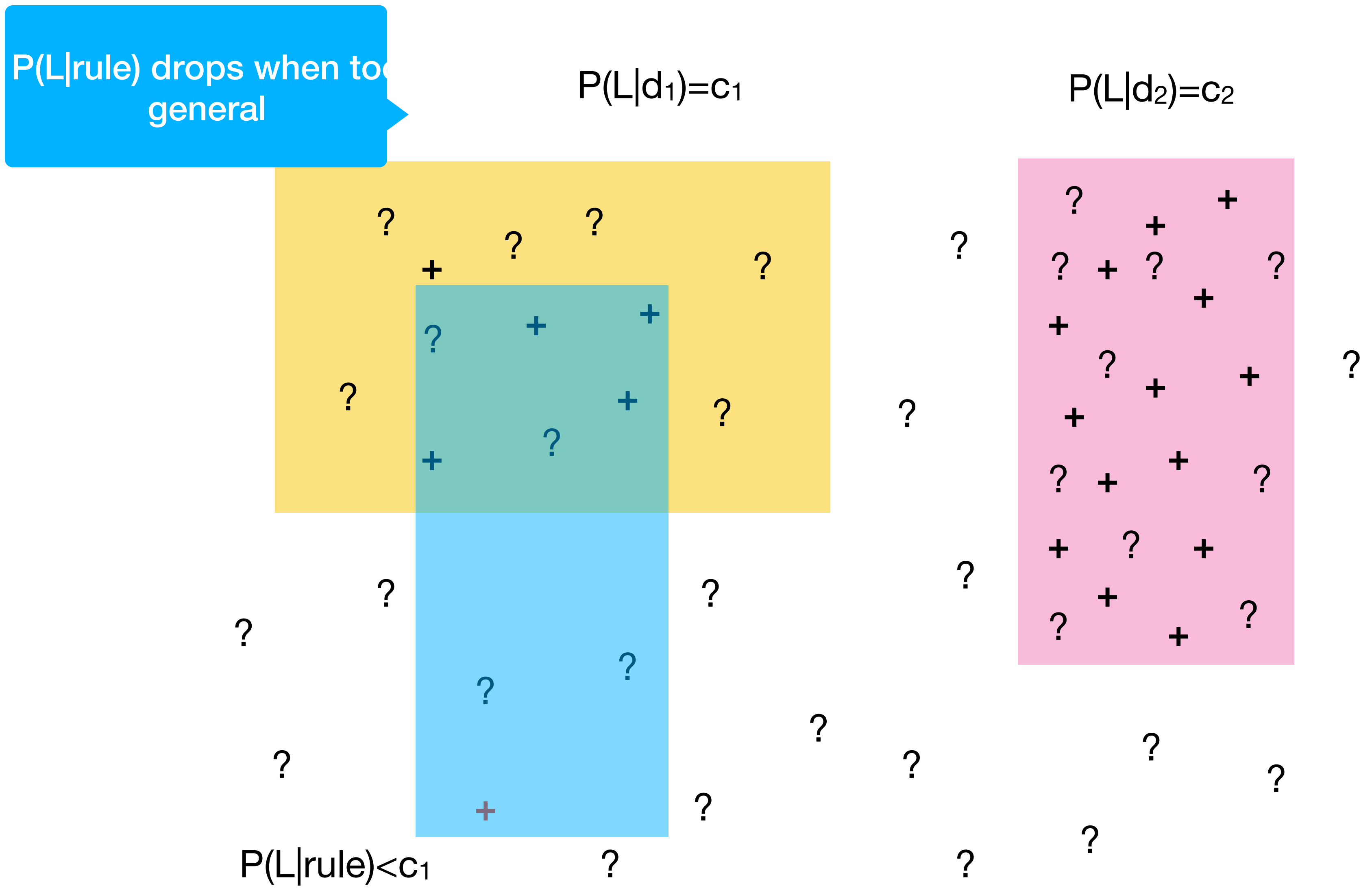
Our proposal

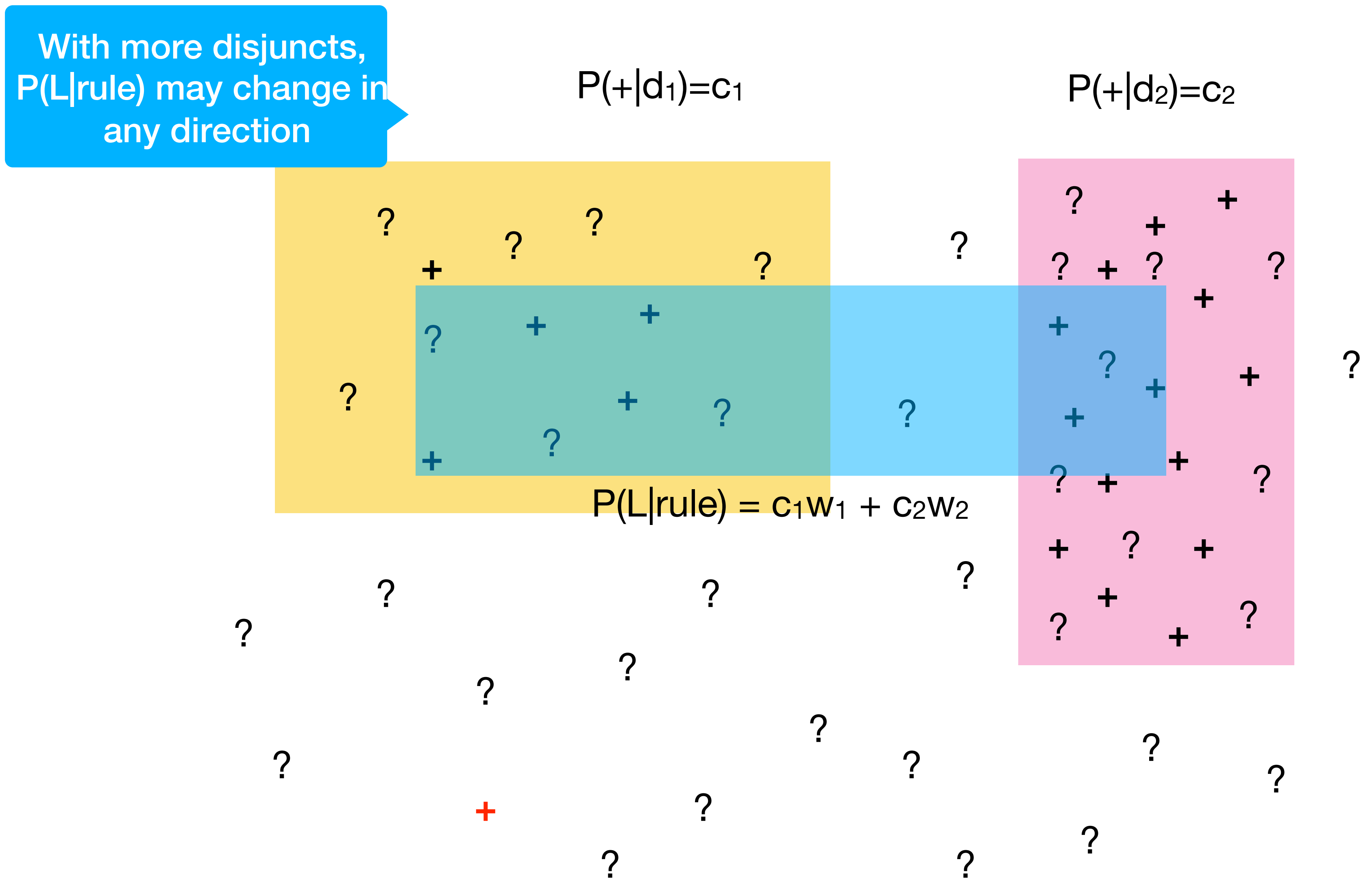
- Learn one rule (= disjunct) at a time, using a bottom-up rule learner (BURL)
    - In our case: Golem-like learner (Muggleton & Feng 1992), generalizing via lgg
- Normally, BURL starts with a specific rule (covering 1 example), and generalizes until it can't generalize anymore without including negatives
- In PU-learning, the stopping criterion is less obvious: generalization *should* include unlabeled - but not too many. *What's too many?*


Assume 2 disjuncts d_1, d_2 to be learned
$P(L|d_1)=c_1$
$P(L|d_2)=c_2$


Start with a seed example, generalize
P(L|d1)=c1
P(L|d2)=c2


As long as rule ⊆ disjunct,
P(L|rule) is constant
P(L|d1)=c1
P(L|d2)=c2
P(L|rule)=c1


P(L|rule) drops when to
general
$P(L|d_1)=c_1$
$P(L|d_2)=c_2$
$P(L|rule)<c_1$

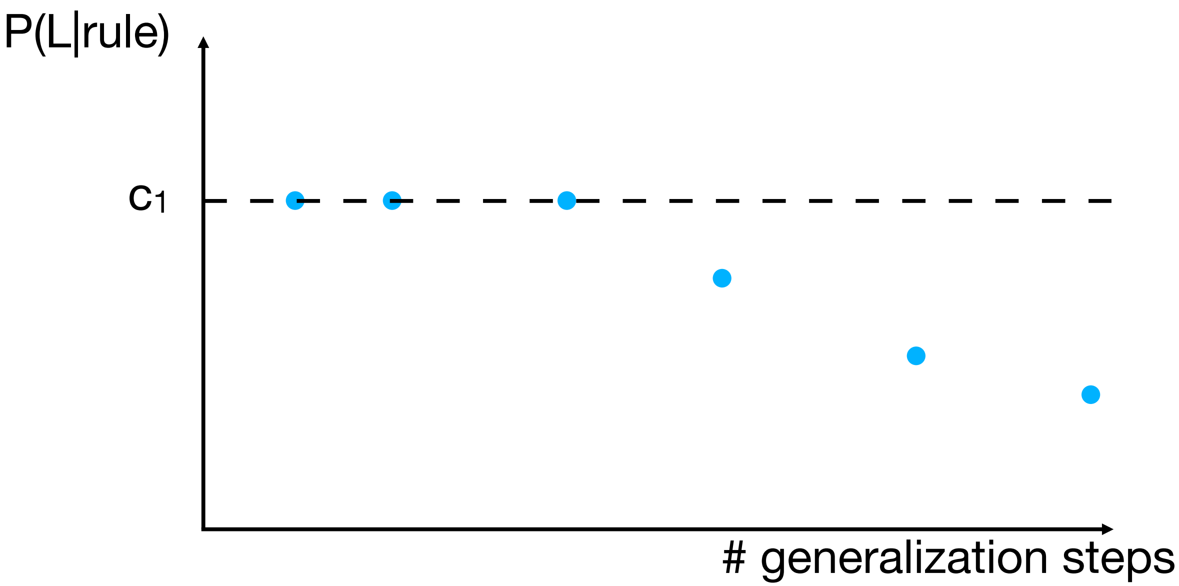


With more disjuncts, P(L|rule) may change in any direction
P(+|d1)=c1
P(+|d2)=c2
P(L|rule) = c1w1 + c2w2

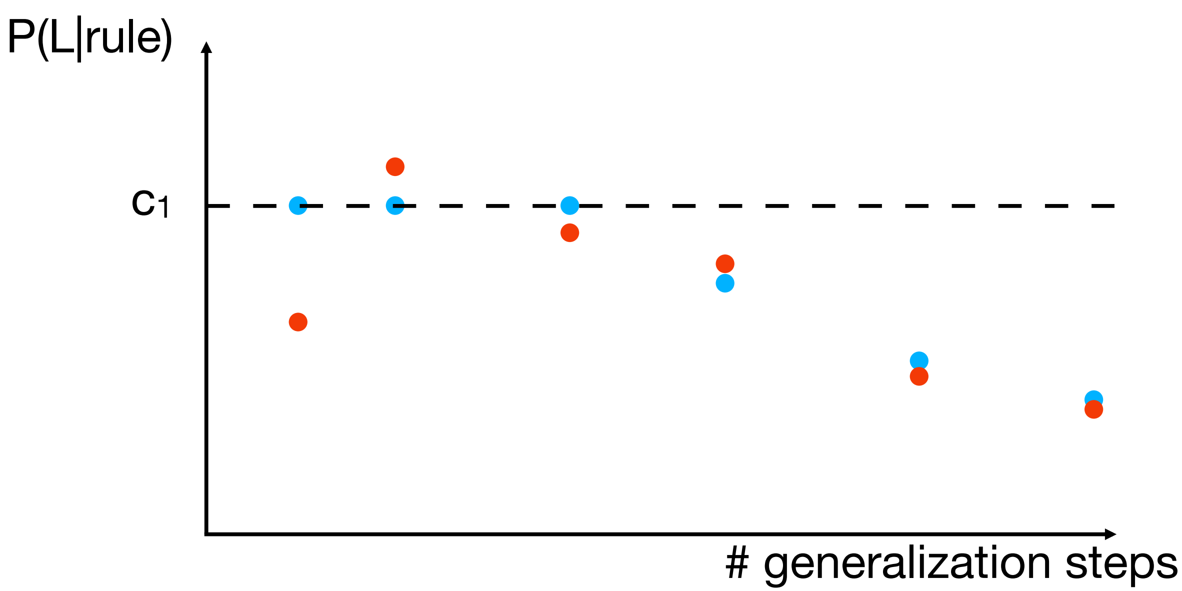
Assumption

- Our method implicitly assumes that P(L|rule) starts going down when rule $\not\subseteq c_1$

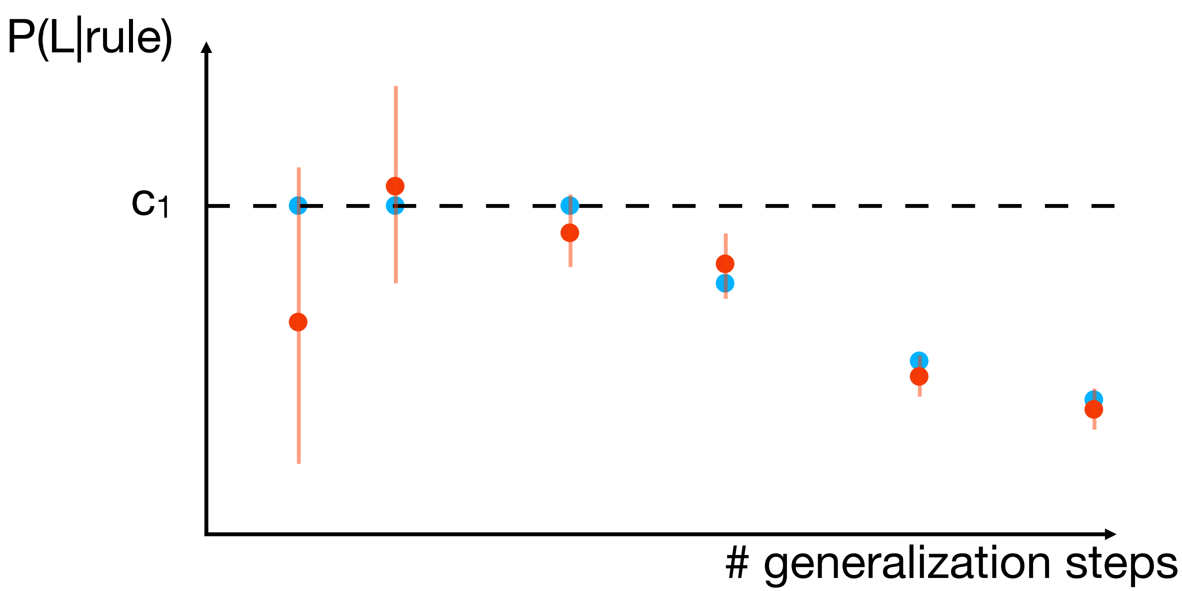
- This assumption holds when $c_1 > c_2$, or when disjuncts are "small" and "faraway"



While rule $\subseteq d_1$: $P(L|rule) = c_1$
As soon as rule $\not\subseteq d_1$: $P(L|rule) < c_1$



… But we have estimates for P, not P itself!
Estimates are closer to P as coverage increases



We can construct confidence intervals.

To find the rule with largest coverage ⊆ d_1, choose the point with maximal lower bound.

This heuristic makes sense because

- Very small coverage => wide interval => low LB
- Rule ⊈ d_1 => low expectation => low LB

Algorithm: PULOR (PU-Learn One Rule)

- Choose a random positive e
- d_0 = e, i=0
- While examples left:
  - Choose ≤s random examples e_j, not covered by d_i
  - d_{i+1} = argmax_j quality(lgg(d_i,e_j))
  - i++
- Return argmax_i quality(d_i)

With quality(d) = LB(P(pos|d))

Algorithm: PULSE

- Repeat
  - choose a random uncovered positive example, e
  - R = PULOR(e)
  - Add R to RuleSet
  - Mark positive examples covered by R as covered
- Until no good rule R can be found
- Remove redundant rules (subsumed by other rules)

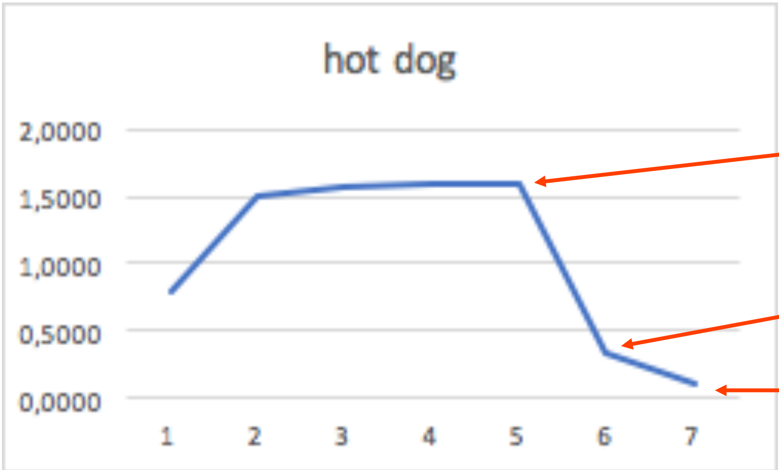
Standard covering approach

Results

- Experiments on a dataset with 10k (sentence, context) pairs (Becerra-Bonache et al., ECAI 2016, derived from Zitnick et al.'s (2013) dataset) : some representative results

Cat	`[object(A),animal(A,c_cat),size(A,c_small)]`
Dog	`[object(A),animal(A,c_dog),color(A,c_brown),size(A,c_small)]` `[object(A),color(A,B),object(C),food(C,c_hot_dog)]`
Table	`[object(A),large(A,c_table),color(A,c_yellow)]`
Sitting	`[object(A)]`
The dog	`[object(A),animal(A,c_dog),color(A,c_brown),size(A,c_small)]`
Hot dog	`[object(A),food(A,c_hot_dog)]`
Hot air balloon	`[object(A),sky(A,c_hot_air_balloon),size(A,c_big),act(A,c_fly)]`
Mike	`[object(A),human(A,c_boy),pose(A,B),expression(A,C)]`
Angry	`[object(A),color(A,B),object(C),human(C,c_boy),pose(C,D),expression(C,c_angry)]` `[object(A),color(A,B),object(C),human(C,c_girl),pose(C,D),expression(C,c_angry)]`

Quality curves



[object(A),food(A,c_hot_dog)]

[object(A),food(A,_)]

[object(A)]

"Hot dog"

Quality curves

Quality

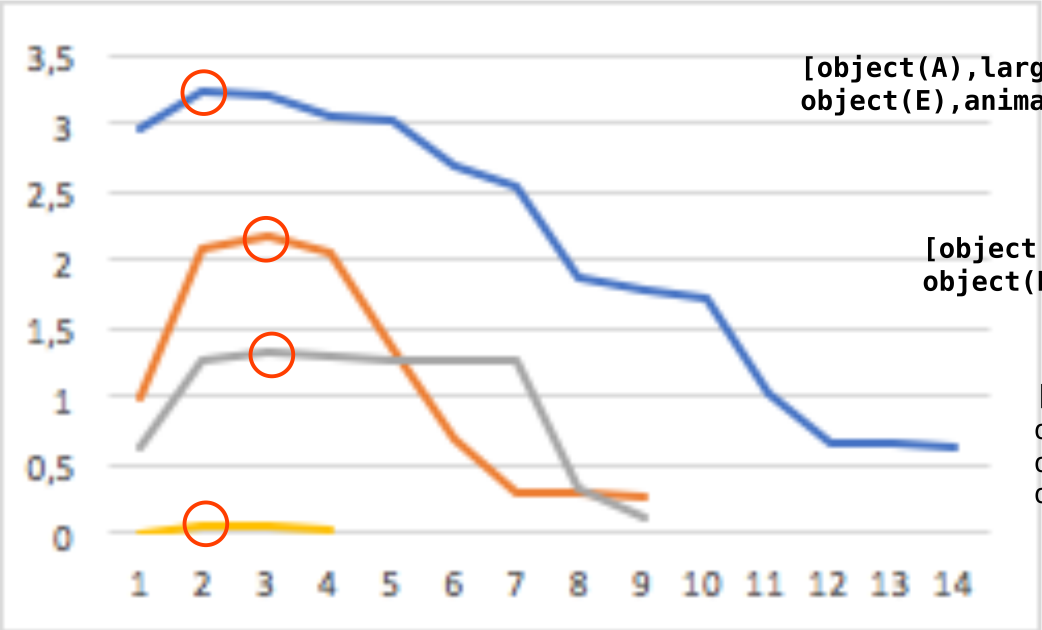
`[object(A),large(A,B),object(C),object(D),`
`sky(C,E),color(D,F),size(C,c_big),`
`object(G),animal(G,c_cat),size(G,c_small)]`

`[object(A),animal(A,c_cat),size(A,c_small)]`

generalization steps

“Cat”

Quality curves



[object(A),large(A,B),object(C),color(C,D),size(C,c_big), object(E),animal(E,c_dog),color(E,c_brown),size(E,c_small)]

[object(A),human(A,c_boy),pose(A,B),expression(A,C), object(D),animal(D,c_dog),color(D,c_brown),size(D,c_small)]

[object(A),size(A,B),object(C),color(C,D), object(E),human(E,c_boy),pose(E,F),expression(E,G), object(H),human(H,c_girl),pose(H,I),expression(H,J), object(K),food(K,c_hot_dog)]

[object(A),food(A,B)]

"Dog"

Conclusions

- Main contribution: PULSE: A new method for PU-learning rule sets in ILP, does not assume constant P(L|+)
- Applied to relational grounded language learning:
  - Finds similar conjunctive concepts as before, but can also identify disjunctive concepts
  - Interpretable results obtained
  - "Quality curves" confirm soundness of method's principle
  - (Bonus: ad-hoc noise handling method, used in earlier work, now no longer needed)