

## Motivation and outline

- Many real problems: Available labeled data is much more limited or expensive than unlabeled data
- Utilizing unlabeled data may lead to advantage over (standard) supervised learning
- Learning Models: Active, Semi-supervised, and transductive

2

## Incorporating unlabeled data in the learning process

(preliminary version)

1

## Active learning

4

## A bit on each model

- Active Learning: The learner chooses its training set
- Semi-supervised Learning: Training set consists of both labeled and unlabeled examples
- Transductive Learning: The learner is given the relevant test set (before beginning the learning phase)

3

## Active learning (pool based)

- Learning is modeled by  $P_{X,Y}$  probability on (example, label) pairs:  $(X, Y) \in \mathcal{X} \times \mathcal{Y}$
- Pool of unlabeled point is selected i.i.d. according to (marginal)  $P_X$
- Active learner queries points (only) from the pool

6

## Introduction

- **Active Learner controls the selection of its training set**
- **Iterative learning process:**
  1. Learner queries for some example's label
  2. Teacher retrieve the (true) label
  3. The process repeats (until satisfying some termination rule)
- **Types of queries:**
  - Online (*stream based active learning*)
  - Restricted to a subset of unlabeled points (*pool based active learning*)
  - Unrestricted (*membership queries*)

5

### Hypotheses class: reminder + assumption (for the sake of this tutorial)

- Assume (here) the learner chooses  $h$  from hypotheses class  $\mathcal{H}$  (with  $VC < \infty$ ) and the perfect hypothesis is in  $\mathcal{H}$
- In such case the number of labels needed for passively learn "good" hypothesis (with high prob.) is bounded by  $\Omega(\frac{1}{\epsilon})$

8

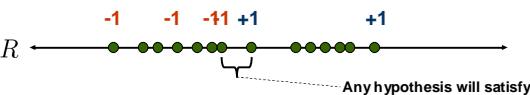
### Active learning goal

- Goal:**
  - Find hypothesis  $h: \mathcal{X} \rightarrow \mathcal{Y}$  s.t.,  $P_{X,Y}\{h(X) \neq Y\} \leq \epsilon$
  - Do so with minimal labeling effort
- Hope:**
  - Number of queries is substantially smaller than needed for passively learning  $h$

7

### Example (cont.)

- Active Strategy:**
  - Draw a pool (of size  $\Omega(\frac{1}{\epsilon})$ )
  - Perform binary search to find the threshold  $x$

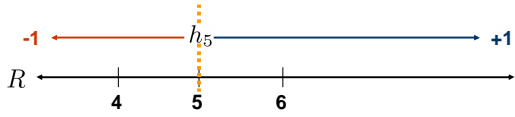


- How many queries?

10

### Example: Threshold functions on the line

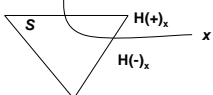
- $\mathcal{H} = \{h_x : x \in R, h_x(x') = 2I[x' \geq x] - 1\}$
- For example:



9

### Searching the hypotheses space

- Version Space:** The set of hypotheses that are consistent with the (currently) aggregated training set (labels)
- Illustration:  $x$  splits the version space  $S$



- Good strategy: Query  $x$  that halves  $S$

12

### Active learning as search in hypotheses space

- Let  $h^*$  be the perfect hypothesis (recall that here  $h^* \in \mathcal{H}$ )
- For every  $x$  (in the pool) define:
 
$$\mathcal{H}_x^+ = \{h \in \mathcal{H} : h(x) = +1\}; \text{ similarly } \mathcal{H}_x^-$$
- After each query either  $\mathcal{H}_x^+$  or  $\mathcal{H}_x^-$  contains  $h^*$

11

### Does active learning always improve passive learning?

- No. Consider the following  $\mathcal{H}$

The diagram shows a circle representing a hypothesis space. Several lines representing hyperplanes are drawn across it, labeled  $h_0, h_k, h_3, h_2, h_1$ . A small distance  $\epsilon$  is indicated between the circle and one of the hyperplanes.

14

### Example with threshold func.

- $\mathcal{H}$  is the set of all threshold func's on line
- Example:
 

$S_0 = \mathcal{H}, S_1 = (3, 10], S_2 = (6, 10] \dots$

A horizontal number line labeled  $R$  with points from 3 to 10. Above the line, there are green dots at each integer. Labels  $-1$  are placed above the dots at 3 and 6, and a label  $+1$  is placed above the dot at 10.

13

### SVM hypotheses space

- Recall hypothesis is defined by  $w$
- Version space:
 
$$S = \{w \in W : \|w\| = 1, y_i(w \cdot \Phi(x_i)) > 0, i \in [n]\}$$
- Version/Feature spaces Duality:
  - (definition) points in  $W$  are hyper-planes in  $F$
  - (intuition) point  $\Phi(x_i) \in F$  restricts hyper-planes to ones that classify  $x_i$  correctly

16

### SVM active learner

- SVM active learner:  $(f, q, (L, U))$

```

    graph TD
      A["(f, q, (L, U))"] --> B[Classifier]
      A --> C[Querying]
      A --> D["Pool = (L, U)"]
      B --> C
      C --> D
    
```

- Operation mode:
  - Use  $L$  to create  $f : X \rightarrow \{-1, +1\}$
  - Query next point  $x = q(U, f)$
  - Update the pool and repeat

15

### Duality idea: by illustration

The diagram shows a 2D coordinate system with axes  $x_1$  and  $x_2$ . A shaded region represents the 'Version space'. Several lines represent hyperplanes in the parameter space  $W$ . A dashed line indicates a 'Hyper-plane corresponding to example x'. A label 'Defines SVM's solution' points to the version space.

18

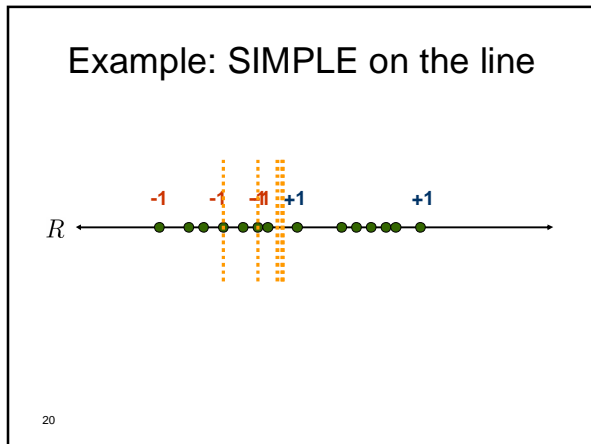
### Version/Feature space duality

- (Recall) SVM objective:  $\max_{w \in F} \min_i \{y_i(w \cdot \Phi(x_i))\}$   
subject to  $\|w\| = 1, \forall i y_i(w \cdot \Phi(x_i) + b) > 0$
- Solution: point  $w$  in version space that maximize the distance  $\min_i \{y_i(w \cdot \Phi(x_i))\}$

---

- Duality: each  $\frac{\Phi(x_i)}{\|\Phi(x_i)\|}$  is hyper-plane in  $W$
- Because constraints they delimit  $S$
- $y_i(w \cdot \Phi(x_i))$  is  $\sim$ distance between  $w$  and  $\Phi(x_i)$

17



### SIMPLE querying scheme

- Solution  $w^*$  in “center” of version space
- SIMPLE strategy: query  $x \in U$  that is closets to  $w^*$
- Such  $x \in U$  will “halve” version space
- Always “halves”?

19

### Practical issues

- Empirical evaluation – protocol:

Training

Test

1. Run active learner for [Training] rounds
2. At each round evaluate resulted classifier on Test

↻ Pool

22

### Example: SIMPLE on XOR

- What happens here?

21

### Practical issues (cont.)

- Discuss (practical) open issues
- Show live demonstration: image classification

ONE -vs- SEVEN

Tap object on:

Stop querying and start: [Advanced Classification](#)

24

