When the **learning** distribution differs from the **target** (true) distribution

Learning from **positive examples** only

Semi-supervised learning

Antoine Cornuéjols

AgroParisTech – INRAE MIA Paris-Saclay

EKINOCS research group

When $P_X(\text{train}) \neq P_X(\text{test})$

$P_X(\text{train}) \neq P_X(\text{test})$

- In which scenarios?

$P_X(\text{train}) \neq P_X(\text{test})$

In which scenarios?

1. Classes are severely **unbalanced**
2. Learning from **positive** examples **only**
3. **Semi-supervised** learning
4. **Active** learning

Outline

1. Classes severely imbalanced
2. Learning from positive examples only
3. Semi-supervised learning
4. Active learning

Illustrations

- Rare pathologies

- Anomaly detection

- Fraud

- Rare species
  - E.g. Pl@ntNet: **46,000** species, but only ~**1000** well represented

Remedies

Remedies

- If **enough** data
  - **undersample** the over-represented classes

Remedies

- If **enough** data
  - **undersample** the over-represented classes
- If **not enough** data

Remedies

- If **enough** data
  - **undersample** the over-represented classes
- If **not enough** data
  - **oversample** the under-represented classes
    - Create **noisy** clones of the data points
    - Create **new** data points generated by **well chosen transformations**
      - E.g. respecting **invariances** (E.g. translations, rotations, change of luminosity, …)

Remedies

- If **enough** data
  - **undersample** the over-represented classes

- If **not enough** data
  - **oversample** the under-represented classes
    - Create **noisy** clones of the data points
    - Create **new** data points generated by **well chosen transformations**
      - E.g. respecting **invariances** (E.g. translations, rotations, change of luminosity, …)

- Modify the **loss function**
  - **Penalize** more the errors on the under-represented class

$$\ell_{\hat{\mathrm{M}},\mathrm{m}} P_{\hat{\mathrm{M}},\mathrm{m}} + \ell_{\hat{\mathrm{m}},\mathrm{M}} P_{\hat{\mathrm{m}},\mathrm{M}} \qquad \text{with} \qquad \ell_{\hat{\mathrm{M}},\mathrm{m}} >> \ell_{\hat{\mathrm{m}},\mathrm{M}}$$

Proportion of all points where points of the minority class are misclassified as from the Majority one

Outline

1. Classes severely unbalanced
2. Learning from positive examples only
3. Semi-supervised learning
4. Active learning

Scenarios for learning from positive examples only

- ???

Scenarios for learning from positive examples only

- Collaborative science
  - Biodiversity
  - E.g. Pl@ntNet
    - The users take pictures of plants: positive examples
    - That does not say: “these other plants were not present”

- Medicine
  - Reports of subjects with some disease does not say how many and which ones do not have the disease

- Adds on web pages
  - Pages that have **not been visited** are not necessarily uninteresting

Scenarios for learning from positive examples only

- In general
  - Detecting absence can be more difficult than detecting presence

Possibly **lots** of **false negative**

The fully observable case

- We look for a **hypothesis** $h : \mathcal{X} \rightarrow [0,1]^L$ A **vector** of predictions
  where *L* is the number of possible classes (labels)

- We want to **minimize the risk** $R(h) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p(\mathbf{x},\mathbf{y})}\ \ell(h(\mathbf{x}),\mathbf{y})$
  with *loss function* $\ell : [0,1]^L \times \mathcal{Y} \rightarrow \mathbb{R}$
  (e.g. binary cross-entropy)

$$\ell_{\mathrm{BCE}}(h(\mathbf{x}_n),\mathbf{y}_n) = -\frac{1}{L}\sum_{i=1}^{L} P(\mathbf{y}_n^i = 1|\mathbf{x}_n)\log\big(h(\mathbf{x}_n^i)\big) + P(\mathbf{y}_n^i = 0|\mathbf{x}_n)\log\big(1 - h(\mathbf{x}_n^i)\big)$$

- Given a dataset $\mathcal{S} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{1\leq n\leq N}$
  we want to find a hypothesis that minimizes the empirical risk

$$\hat{h}_{\text{fully}} = \operatorname*{ArgMin}_{h\in\mathcal{H}} \frac{1}{N}\sum_{n=1}^{N} \ell\big(h(\mathbf{x}_n),\mathbf{y}_n\big)$$

The partially observable case

- We look for a **hypothesis** $h_{\text{partial}} : \mathcal{X} \rightarrow [0,1]^L$

- During training, we observe $\mathbf{z}_n \in \mathcal{Z} = \{0, 1, \oslash\}^L$

  where $\mathbf{z}_n^i = \oslash$ ← indicates that the i^{th} label is unobserved

  and only one $\mathbf{z}_n^i = 1$

- Given a dataset $\mathcal{S} = \{(\mathbf{x}_n, \mathbf{z}_n)\}_{1 \leq n \leq N}$

  we want to find a hypothesis that

  minimizes the empirical risk

$$\hat{h}_{\text{partial}} = \operatorname*{ArgMin}_{h \in \mathcal{H}} \frac{1}{N} \sum_{n=1}^{N} \ell\big(h(\mathbf{x}_n), \mathbf{z}_n\big)$$

Approach “assume **unobserved** are **negative**”

- Assume that all **unobserved** labels are **negative**

$$P(\mathbf{y}_n^i = 1 | \mathbf{x}_n) = 0 \quad \text{if} \quad \mathbf{z}_n^i = \oslash$$

- The resulting loss is

$$\ell_{\mathrm{AN}}(h(\mathbf{x}_n), \mathbf{y}_n) = -\frac{1}{L} \sum_{i=1}^{L} \mathbb{1}_{[\mathbf{z}_n^i = 1]} \log\big(h(\mathbf{x}_n^i)\big) + \mathbb{1}_{[\mathbf{z}_n^i \neq 1]} \log\big(1 - h(\mathbf{x}_n^i)\big)$$

$$\mathbb{1}_{[\mathbf{z}_n^i = 1]} = 1 \quad \text{if } \mathbf{z}_n^i = 1 \quad \text{and 0, otherwise}$$

- We expect **false negatives**

Approach “assume **unobserved** are **negative**” + **smoothing**

- Assume that all **unobserved** labels are **negative**

$$P(\mathbf{y}_n^i = 1|\mathbf{x}_n) = 0 \quad \text{if} \;\; \mathbf{z}_n^i = \varnothing$$

- And give **more weight to the observed examples**. The resulting loss is

$$\ell_{\text{AN-LS}}(h(\mathbf{x}_n), \mathbf{y}_n) = -\frac{1}{L}\sum_{i=1}^{L} \mathbb{1}^{0.95}_{[\mathbf{z}_n^i=1]} \log\big(h(\mathbf{x}_n^i)\big) + \mathbb{1}^{0.05}_{[\mathbf{z}_n^i\neq 1]} \log\big(1 - h(\mathbf{x}_n^i)\big)$$

Observed as **positive**

No observation reported
Hence assumed as **negative**

Intuitively $R(\hat{h}_{\text{fully}}) \leq R(\hat{h}_{\text{partial}})$

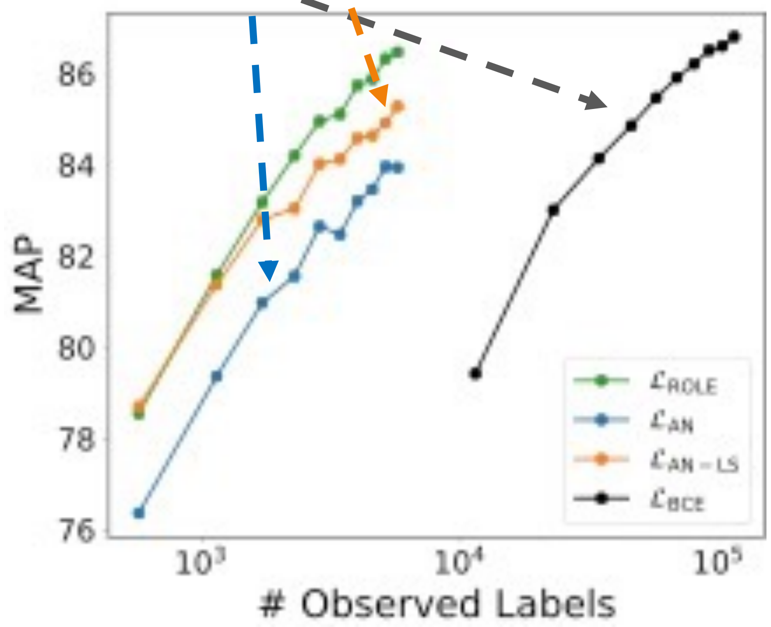
- But **by how much**?

- In the case of “assume unobserved = negative”

Intuitively $R(\hat{h}_{\text{fully}}) \leq R(\hat{h}_{\text{partial}})$

- But **by how much**?

- In the case of “assume unobserved = negative”



With 20 times fewer labeled examples, the performance is not that bad *on this dataset* compared to the fully observable case

COLE, Elijah, MAC AODHA, Oisin, LORIEUL, Titouan, *et al.* Multi-label learning from single positive labels. In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021. p. 933-942.

Lessons

1. **Fomalize** the assumptions about your problem
   - The labelling process
   - The type of target (and hypothesis) function

2. Design a **loss function** appropriate for the problem
   - Able to **explore efficiently** the hypothesis space
     and to find a good minimum of the empirical risk

3. Design a good **evaluation scheme**

Learning from positive examples only: lots of approaches

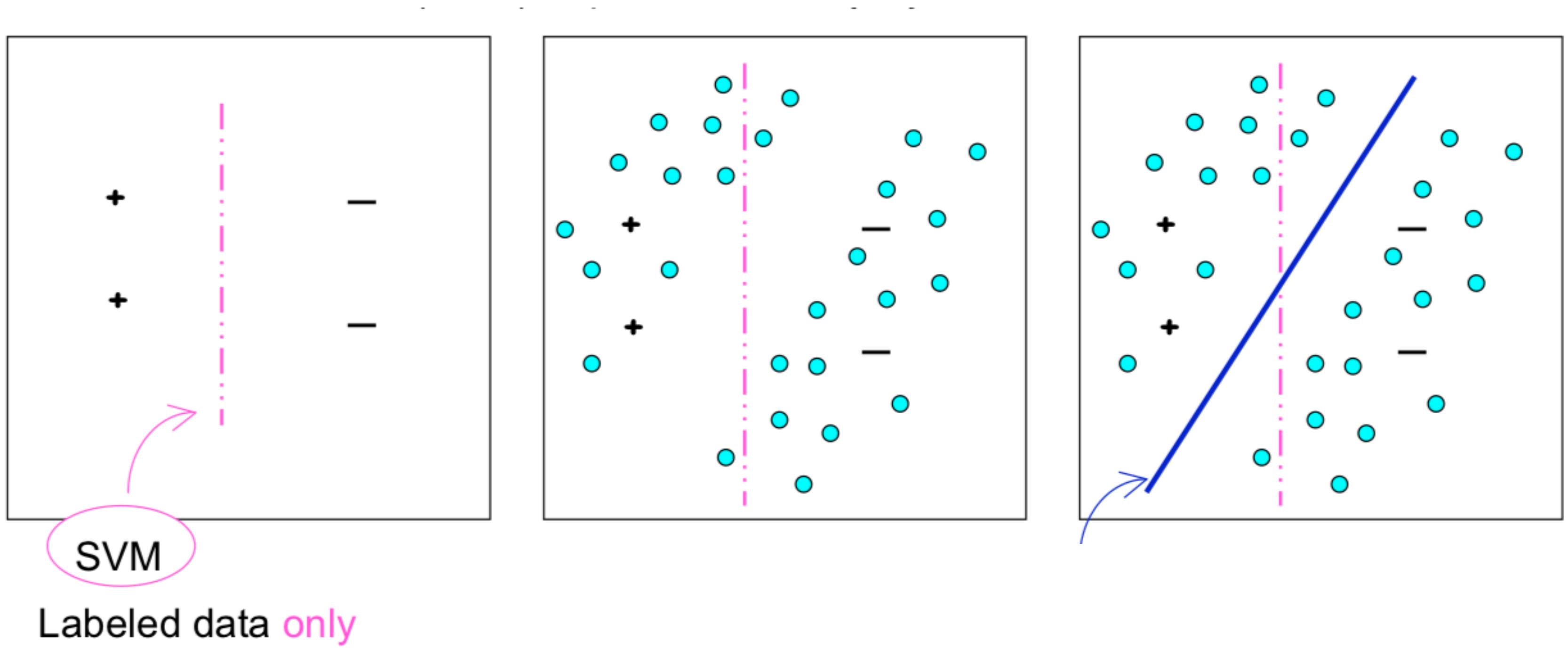
- Approaches
  - Assume that *the missing labels are negative*
  - *Ignore* the missing labels
  - Perform *label matrix reconstruction*
  - Learn *label correlations*
  - Learn *generative probabilistic models*
  - Train *label cleaning networks*

  - Related to **learning with label noise**
    - Here, some **unobserved labels** are incorrectly treated as being **absent**
  - Related to learning from a set of **positive examples** and a set of **unlabeled** ones (**PU** learning)

Outline

1. Classes severely unbalanced
2. Learning from positive examples only
3. Semi-supervised learning
4. Active learning

The idea



...

Semi-supervised learning

- **Unsupervised** learning $\mathbf{P}_{\mathcal{X}}$
- **Supervised** learning $\mathbf{P}_{\mathcal{Y}|\mathcal{X}}$

Semi-supervised learning

- **Unsupervised** learning $\mathbf{P}_{\mathcal{X}}$
- **Supervised** learning $\mathbf{P}_{\mathcal{Y}|\mathcal{X}}$

When can **unsupervised** learning help **supervised** learning?

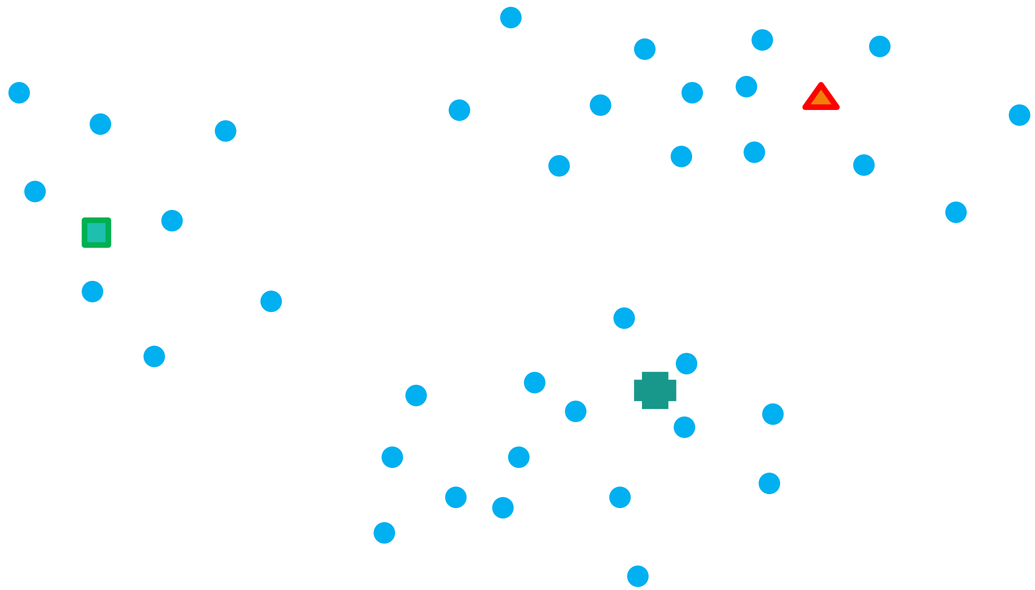
The underlying main idea:

The decision function (hypothesis *h*) **should not cut** through **high density** regions

Semi-supervised learning

Simplest approach

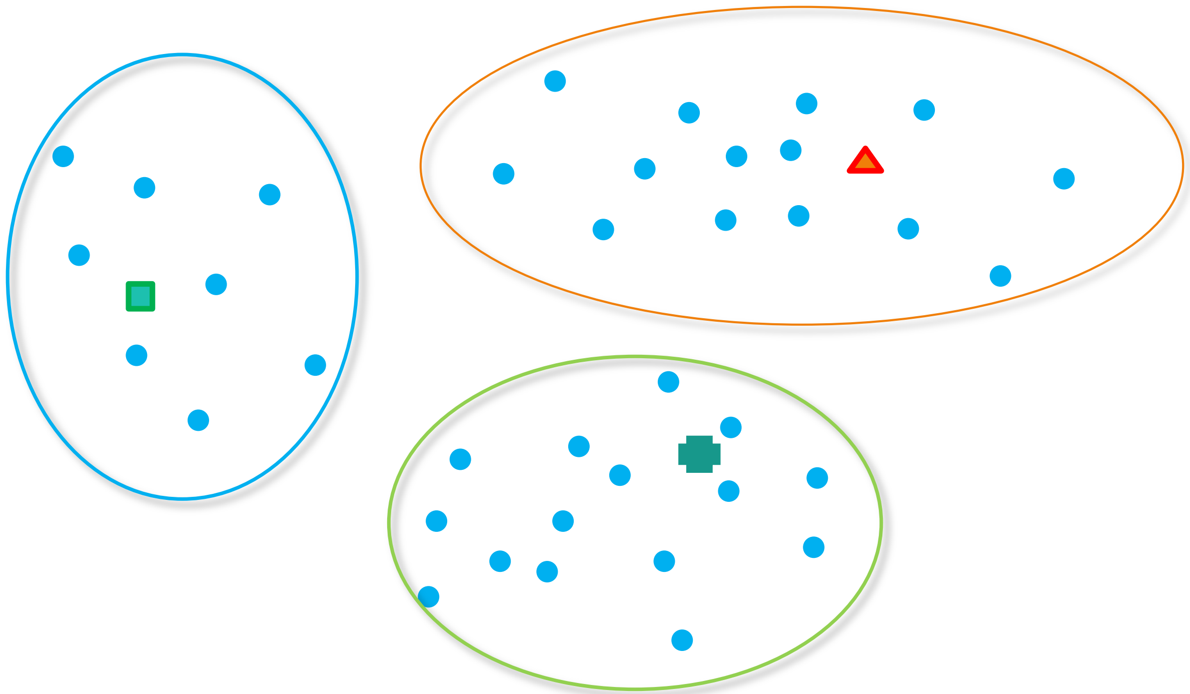
1. Compute a **clustering** of the all data (labeled and unlabeled)
2. For each cluster, **assign its class** to the majority vote of the labeled examples that belong to it



Semi-supervised learning

Simplest approach

1. Compute a **clustering** of the all data (labeled and unlabeled)
2. For each cluster, **assign its class** to the majority vote of the labeled examples that belong to it



Self-training approach

1. Given $\mathcal{S}_L = \{(\mathbf{x}_i, y_i)\}_{1 \leq i \leq l}$ and $\mathcal{S}_U = \{(\mathbf{x}_j)\}_{1 \leq j \leq u}$
2. Train on S_L to obtain h_1
3. Apply h_1 to S_U
4. Remove a set of unlabeled data from S_U and add them to S_L (the one where $h(\mathbf{x})$ is the more confident) with the label $h(\mathbf{x})$
5. Go to 2 and **repeat** until **convergence**

Semi-supervised learning

- Idea: endow unlabeled data with pseudo-labels
  (the likeliest class at time *t*)

$$y_i \;=\; \begin{cases} 1 & \text{if } i = \text{argmax}_{i \in \{1,\ldots,C\}}\, h_i^t(\mathbf{x}) \\ 0 & \text{otherwise} \end{cases}$$

Output of the i^th output neuron

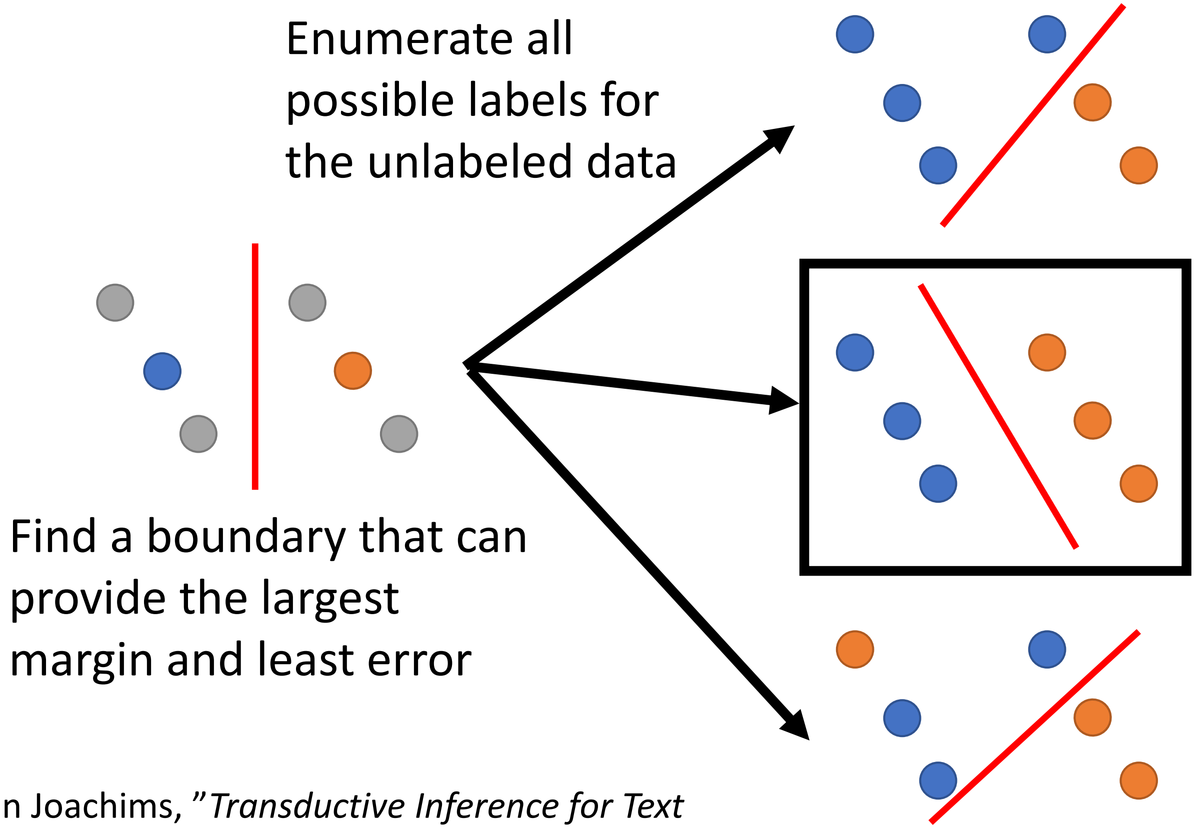
- Train with the empirical risk:

$$R_{\text{emp}}(h) \;=\; \frac{1}{m_l} \sum_{i=1}^{m_l} \sum_{j=1}^{C} \ell(h_j(\mathbf{x}_i), y_j^i) \;+\; \alpha(t)\, \frac{1}{m_u} \sum_{i=1}^{m_u} \sum_{j=1}^{C} \ell(h_j(\mathbf{x}_i), \underbrace{y_j^i}_{\text{pseudo-label}})$$

Crucial to set $\alpha(t)$ with great care

[Dong-Hyun Lee (2013) "*Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks*", ICML-2013]

Transductive SVM approach

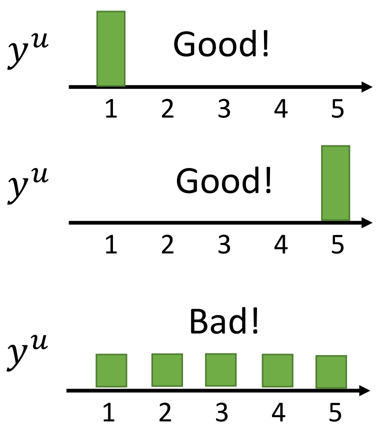


Thorsten Joachims, *"Transductive Inference for Text Classification using Support Vector Machines",* ICML, 1999

Semi-supervised learning

Entropy regularization approach

$$\hat{h} = \underset{h \in \mathcal{H}}{\text{ArgMin}} \left[\underbrace{\frac{1}{l} \sum_{i=1}^{l} \ell(h(\mathbf{x}_i), y_i)}_{\text{Empirical risk on labeled data}} + \lambda \underbrace{\sum_{j=1}^{u} -h(\mathbf{x}_j) \log h(\mathbf{x}_j)}_{\text{Entropy of the predictions}} \right]$$



$E(y^u) = 0$

$E(y^u) = 0$

$$E(y^u) = -ln\left(\frac{1}{5}\right) = ln5$$

- You have to make assumptions about what you think is reasonable as a bias
  - E.g. that classes are separated by low density regions

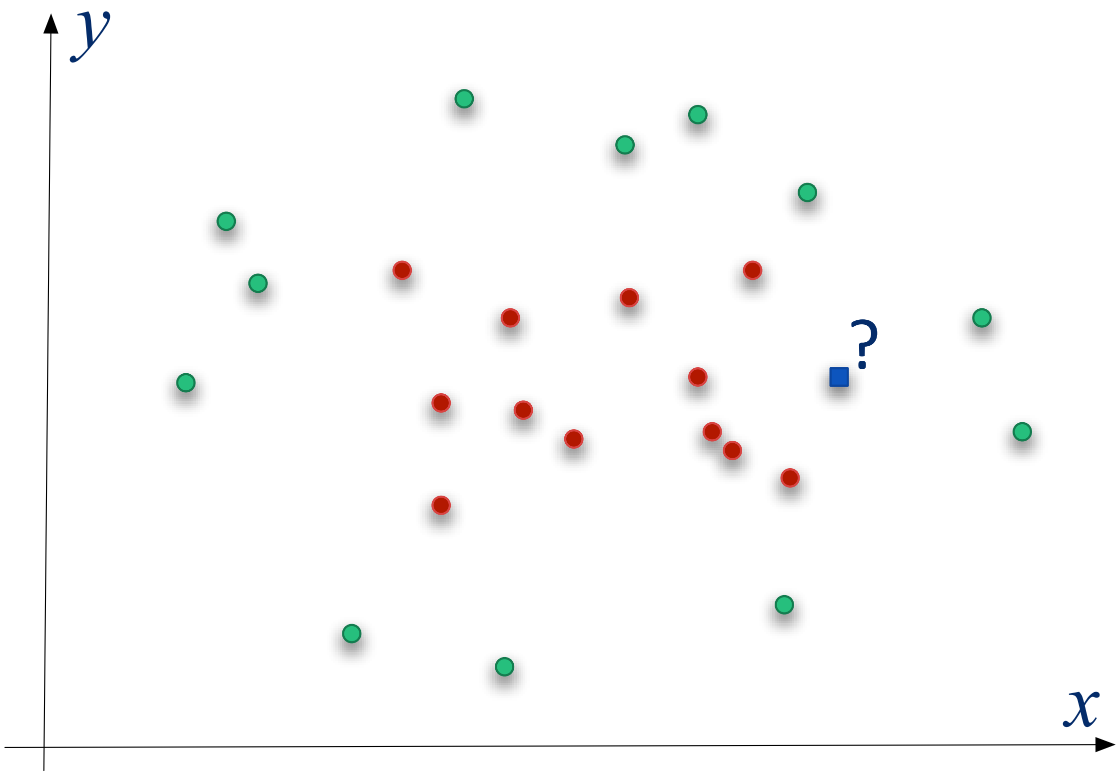
- Then, you show that if the assumption is met by Nature, then you find a correct hypothesis

A **remark** on semi-supervised learning

- Could be regarded as **transductive learning** where one wants to label unlabeled training instances

Transductive learning

- I know in advance where I will be queried



Transductive learning

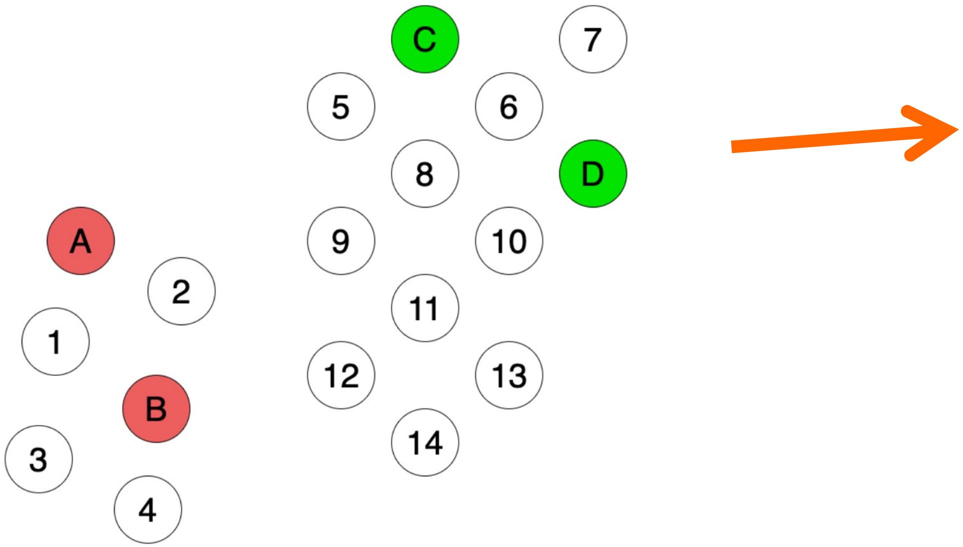
- "When solving a problem of interest, do not solve a more general problem as **an intermediate step**.

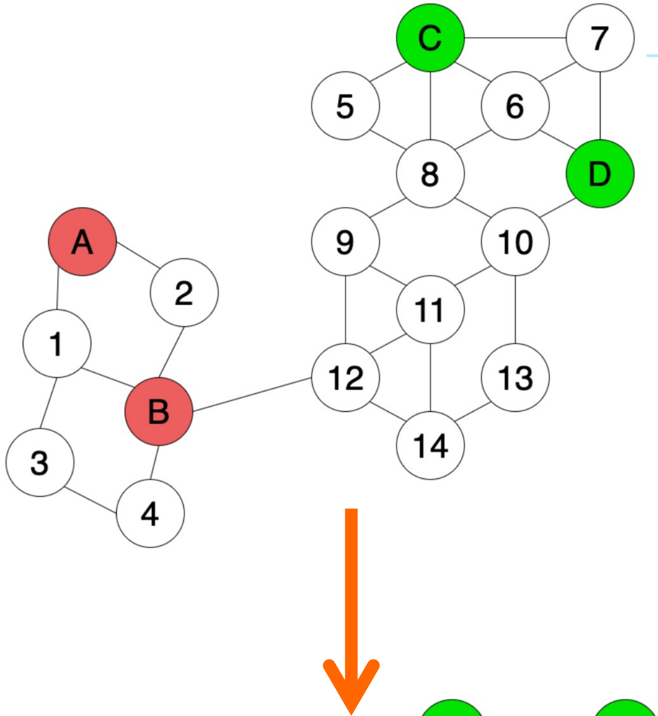
  Try to get the answer that you really need but not a more general one.”

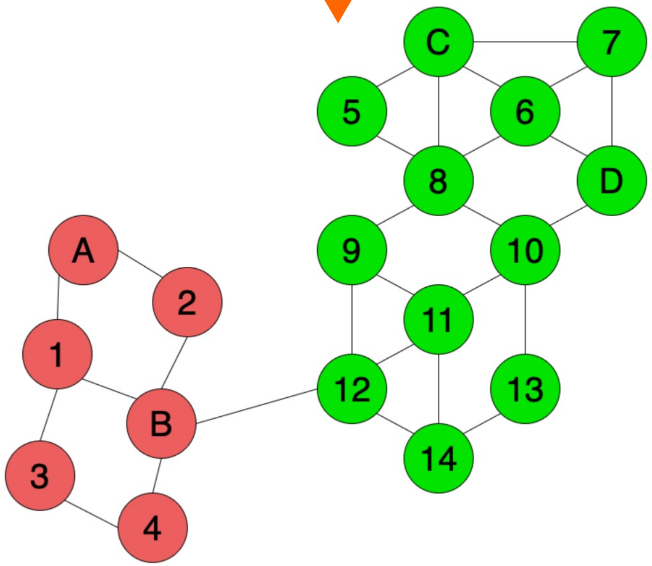
(Vapnik, 1995)

Semi supervised learning with **transductive** learning

- Graph-Based labelling







Then **learn** a hypothesis on the new training set

Outline

1. Classes severely unbalanced
2. Learning from positive examples only
3. Semi-supervised learning
4. Active learning

Active learning

- When the learner can **actively ask** for pieces of information
  - Labels of selected **examples**
  - Values of some selected **descriptors**
    - E.g. ask for a medical examination

- Examples
  - MasterMind
  - Scientific activity

Active learning

- When the learner can **actively ask** for pieces of information
  - Labels of selected **examples**
  - Values of some selected **descriptors**
    - E.g. ask for a medical examination

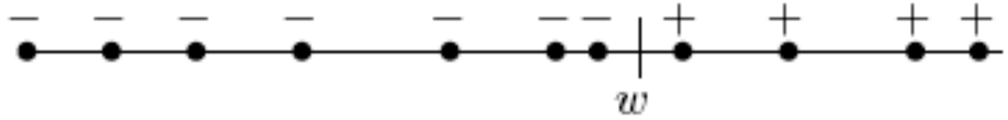
- The **hope**
  - Need of **less** (costly) examples
  - Having a **faster** convergence rate

$$\forall h \in \mathcal{H}, \forall \delta \leq 1: \quad P^m \left[R_{\text{Réel}}(h) \; \leq \; R_{\text{Emp}}(h) \; + \; \boxed{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{m}} \right] \; > \; 1 - \delta$$

$$\forall h \in \mathcal{H}, \forall \delta \leq 1: \quad P^m \left[R_{\text{Réel}}(h) \; \leq \; R_{\text{Emp}}(h) \; + \; \boxed{\sqrt{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2\, m}}} \right] \; > \; 1 - \delta$$

Active learning

$$h_w(x) = \begin{cases} 1 & \text{if } x \geq w \\ 0 & \text{if } x < w \end{cases}$$



How to find the **best** threshold from querying points?

- By **random** selection of points $\quad m = \mathcal{O}(\frac{1}{\varepsilon} \log \frac{1}{\varepsilon})$

- By **active** selection $\quad m = \mathcal{O}(\log \frac{1}{\varepsilon})$

Much faster!

Active learning

- Two main approaches
  - "**Constructive**" approach
    - The learner **constructs** queries
  - "**Selective**" (pool-based) approach
    - The learner **selects** points among the **unsupervised** ones

Why is the **constructive** approach sometimes **not** applicable?

How to **select** the examples? (some ideas)

- The more **informative** examples

1. The ones where the **confidence** of the current hypothesis is the **lowest**
   - Measured by a **probability**

$$\mathbf{x}^\star = \underset{\mathbf{x} \in \mathcal{S}_U}{\text{ArgMax}} \; \text{Uncertain}(\mathbf{x}) \qquad \text{Uncertain}(\mathbf{x}) = \frac{1}{\text{ArgMax}_{y \in \mathcal{Y}} \, p\big(h_t(\mathbf{x}) = y\big)}$$

$$\mathbf{x}^\star = \underset{\mathbf{x} \in \mathcal{S}_U}{\text{ArgMax}} \left\{ -\sum_i p\big(h_t(\mathbf{x}) = y_i\big) \, \log p\big(h_t(\mathbf{x}) = y_i\big) \right\}$$

Entropy criyeria

   - Measured by **distance** to the decision function

2. Learn an **ensemble** of hypotheses and select the examples where they **disagree** the most

Illustration

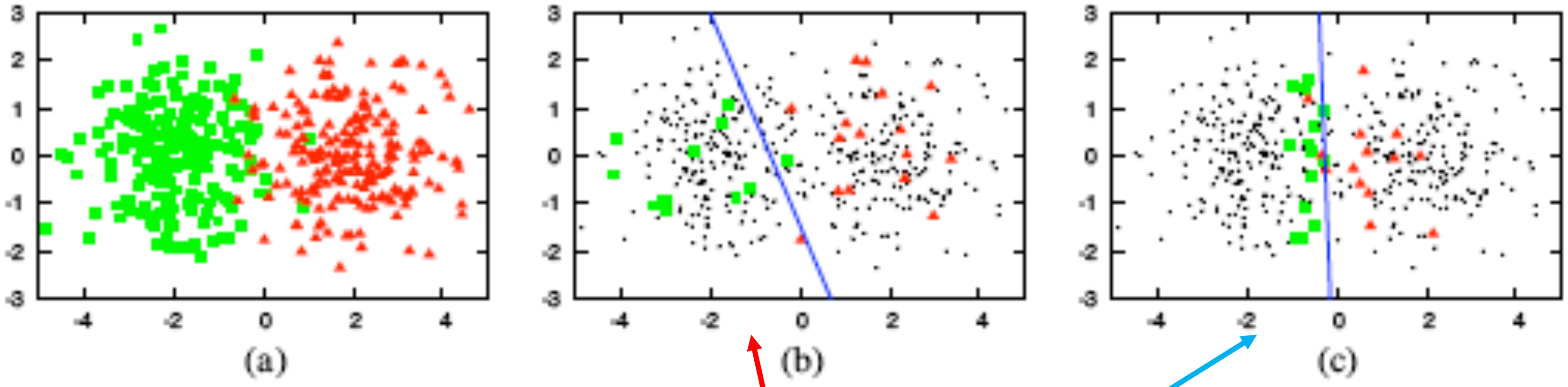


Figure 2: An illustrative example of pool-based active learning. (a) A toy data set of 400 instances, evenly sampled from two class Gaussians. The instances are represented as points in a 2D feature space. (b) A logistic regression model trained with 30 labeled instances randomly drawn from the problem domain. The line represents the decision boundary of the classifier (accuracy = 0.7). (c) A logistic regression model trained with 30 actively queried instances using uncertainty sampling (accuracy = 0.9).

...

- What is the danger?

Active Learning

- What is the **danger**?

  - **No** more **theoretical** guarantees

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m \left[R_{\text{Réel}}(h) \;\leq\; R_{\text{Emp}}(h) \;+\; \boxed{\sqrt{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2\, m}}} \right] \;>\; 1 - \delta$$

Does not make sense anymore!!

  - Why?

Active learning: **lessons**

- Active learning is **not much used** in practice

  1. **Costly** to identify informative examples
  2. **Risk** of ignoring important regions of X

- Interesting: learning under **budget constraints**

  – What measurements should I made under some budget constraints?