LEARNING FROM IMPRECISE DATA

Eyke Hüllermeier
Intelligent Systems Group
Department of Computer Science
University of Paderborn, Germany

eyke@upb.de

ECSQARU/ISIPTA 2017, Lugano, 11-JUL-2017

OUTLINE

PART 1

Superset learning

PART 2

Optimistic loss minimization

PART 3

Data imprecisiation

SUPERSET LEARNING

... is a specific type of **weakly supervised learning**, studied under different names in machine learning:

- *learning from partial labels*
- *multiple label learning*
- *learning from ambiguously labeled examples*
- *...*

... also connected to learning from **coarse data** in statistics (Rubin, 1976; Heitjan and Rubin, 1991), missing values, **data augmentation** (Tanner and Wong, 2012),

... as well as data modeling based on **generalized sets and measures**, such as **fuzzy data** (Kwakernaak, 1978; Kruse and Meyer, 1987; Puri and Ralescu, 1986; Coppi et al., 2006; Bandemer and Näther, 2011; Viertl, 2011) and **belief functions** (Denoeux, 1995).

SUPERVISED LEARNING

Given a set of (i.i.d.) **training data**

$$\mathcal{D} = \Big\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\Big\} \subset \mathcal{X} \times \mathcal{Y}$$

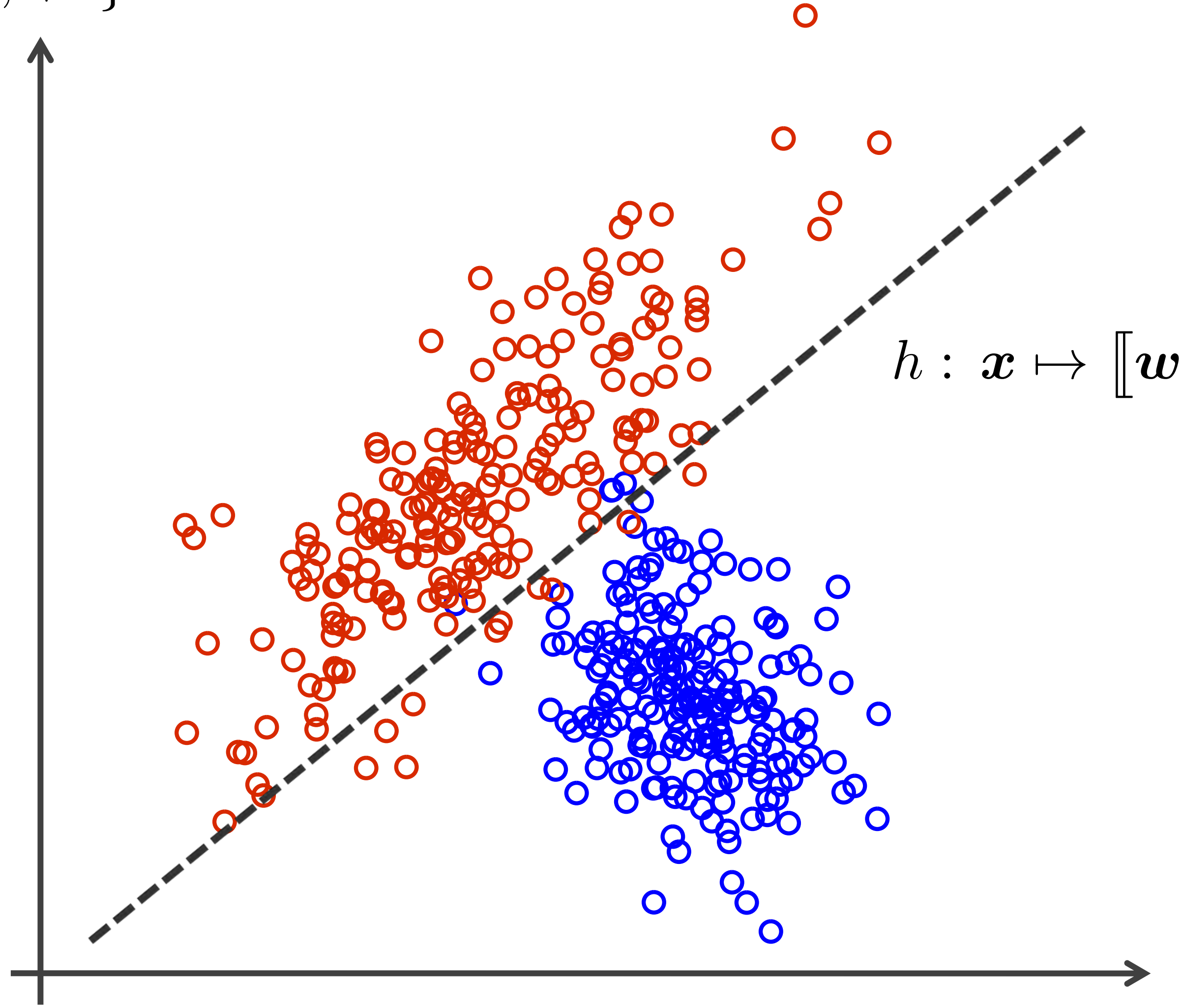
and a **hypothesis space** $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$, find a model with low **risk**

$$\mathcal{R}(h) = \int_{\mathcal{X}\times\mathcal{Y}} \underset{\substack{\uparrow \\ \textit{loss function}}}{L}\big(h(\boldsymbol{x}), y\big)\, d\,\underset{\substack{\uparrow \\ \textit{data generating} \\ \textit{process}}}{\mathbf{P}}(\boldsymbol{x}, y)\,.$$

EXAMPLE: BINARY CLASSIFICATION

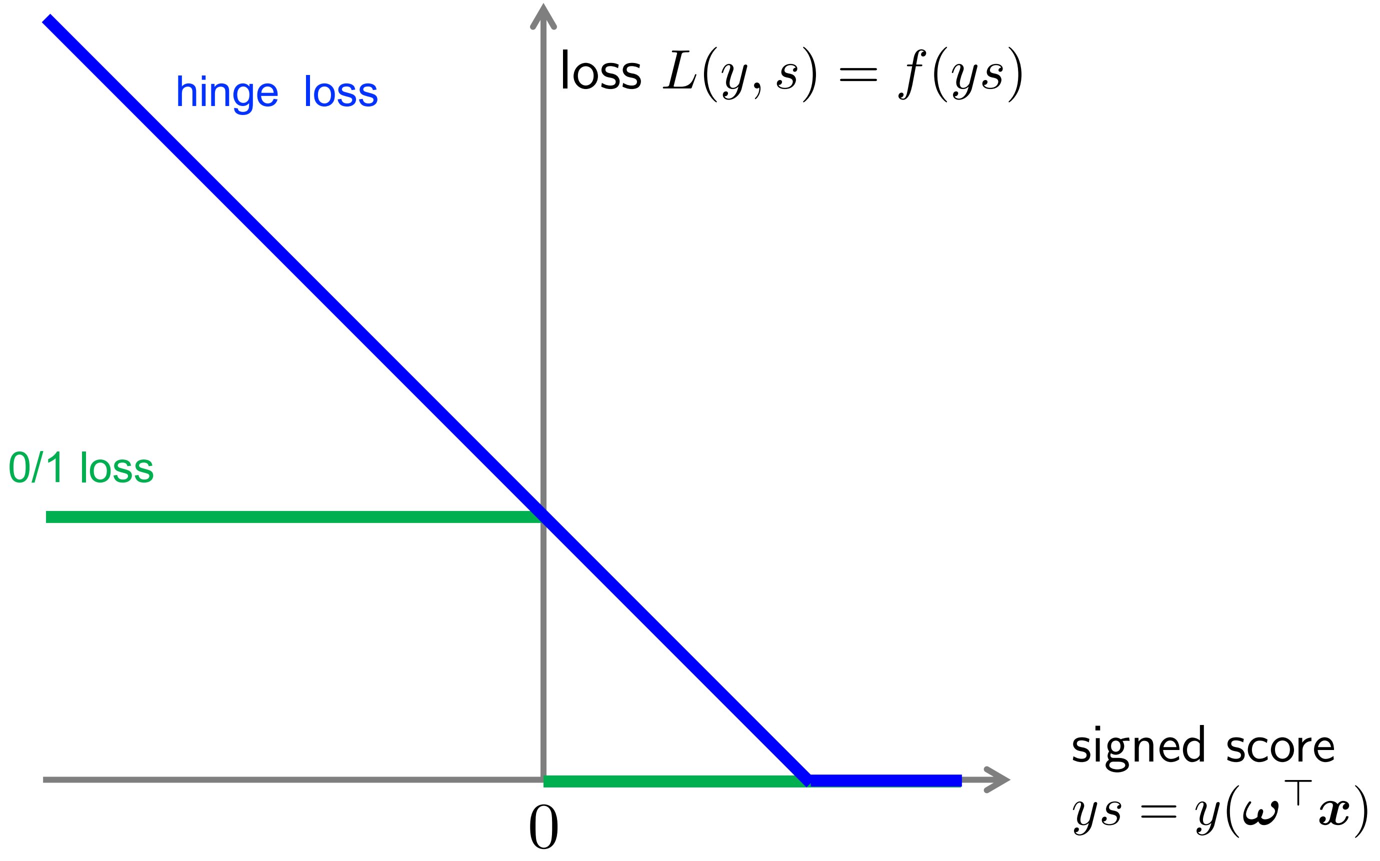
$\mathcal{X} = \mathbb{R}^d,$

$\mathcal{Y} = \{-1, +1\}$



$h : \boldsymbol{x} \mapsto [\![\boldsymbol{w}^\top \boldsymbol{x} > 0]\!]$

EXAMPLE: BINARY CLASSIFICATION



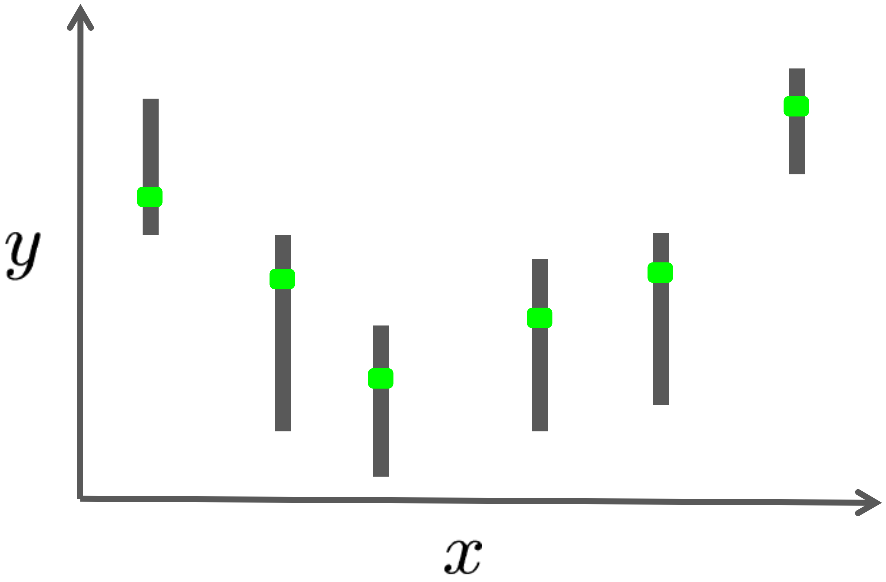
SUPERSET LEARNING

- Set of imprecise/ambiguous/coarse observations

$$\mathcal{O} = \big\{ (\boldsymbol{x}_1, Y_1), \ldots , (\boldsymbol{x}_N, Y_N) \big\}$$

with **supersets** $Y_n \ni y_n$.

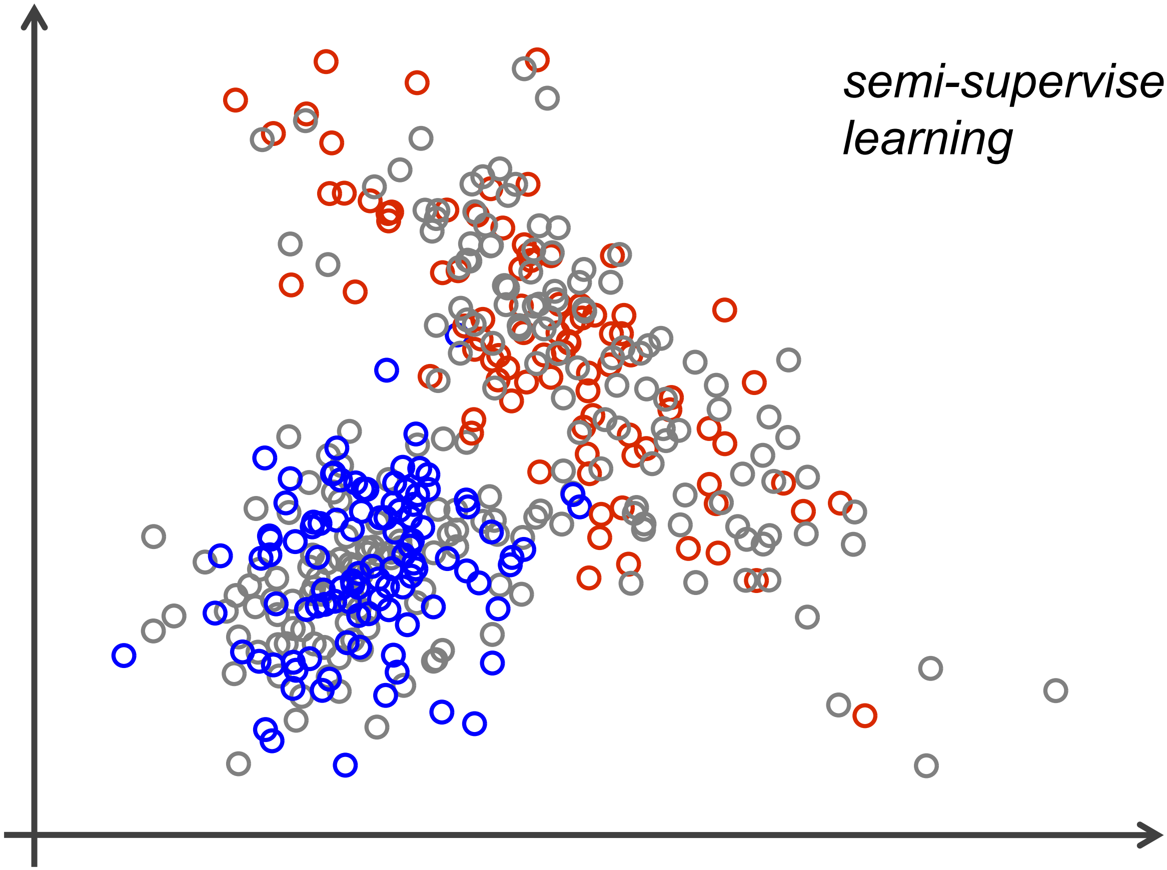
- An **instantiation** of $\mathcal{O}$, denoted $\mathcal{D}$, is obtained by replacing each Y_n with a candidate $y_n \in Y_n$.



one of infinitely many instantiations

EXAMPLE: BINARY CLASSIFICATION

○ = {○, ○}



semi-supervised learning

EXAMPLE: CLASSIFICATION

x_1	x_2	x_3	y_1	y_2	y_3	y_4
21.9	0	154.3	▨	▨		
43.2	1	133.2		▨		▨
53.3	1	163.5			▨	▨
…	…	…	…	…	…	…
42.7	0	142.8	▨	▨	▨	

EXAMPLE: CLASSIFICATION

x_1	x_2	x_3	y_1	y_2	y_3	y_4
21.9	0	154.3	🟥	▨		
43.2	1	133.2		▨		🟥
53.3	1	163.5			▨	🟥
…	…	…	…	…	…	…
42.7	0	142.8	▨	🟥	▨	

EXAMPLE: CLASSIFICATION

x_1	x_2	x_3	y_1	y_2	y_3	y_4
21.9	0	154.3	▨	🟥		
43.2	1	133.2		🟥		▨
53.3	1	163.5			🟥	▨
…	…	…	…	…	…	…
42.7	0	142.8	▨	▨	🟥	

EXAMPLE: COMPLEX DATA

In label ranking, we learn mappings from instances to rankings:

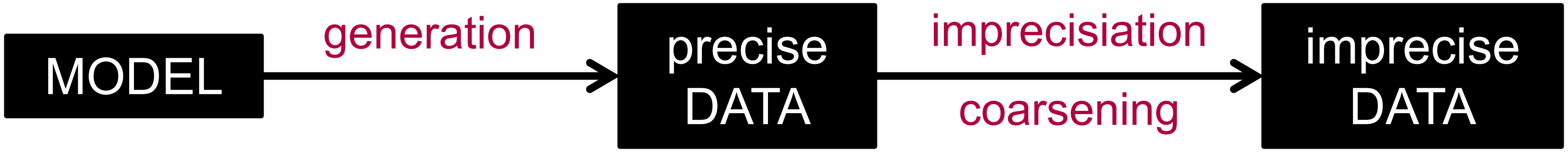
$$\boldsymbol{x} \quad \mapsto \quad \mathsf{A} \succ \mathsf{C} \succ \mathsf{D} \succ \mathsf{B}$$

$$\mathsf{A} \succ \mathsf{C} \quad \longleftrightarrow \quad \left\{ \begin{array}{c} \mathsf{A} \succ \mathsf{C} \succ \mathsf{B} \succ \mathsf{D} \\ \mathsf{A} \succ \mathsf{C} \succ \mathsf{D} \succ \mathsf{B} \\ \mathsf{A} \succ \mathsf{B} \succ \mathsf{C} \succ \mathsf{D} \\ \vdots \succ \vdots \succ \vdots \succ \vdots \\ \mathsf{D} \succ \mathsf{B} \succ \mathsf{A} \succ \mathsf{C} \end{array} \right\}$$

incomplete observation

set of consistent completions

LEARNING FROM SET-VALUED DATA



MODEL

generation

precise DATA

imprecisiation

coarsening

imprecise DATA

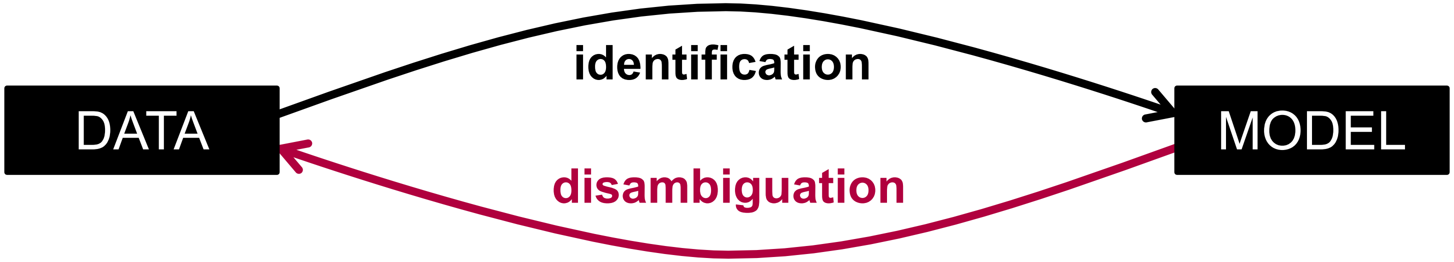
$\mathbf{P}_\theta(\mathcal{D})$ $\mathbf{P}_\lambda(\mathcal{O}\,|\,\mathcal{D})$

- We are interested in learning with **weak assumptions** about the coarsening process, and learning algorithms ought to be **robust** with respect to these assumptions.

- Similar to **epistemic random set setting** (Ω, P, Y), but with little knowledge about multi-valued mapping $Y : \Omega \rightarrow 2^{\mathcal{Y}}$.

- **Discriminative learning**, not generative.

LEARNING FROM SET-VALUED DATA

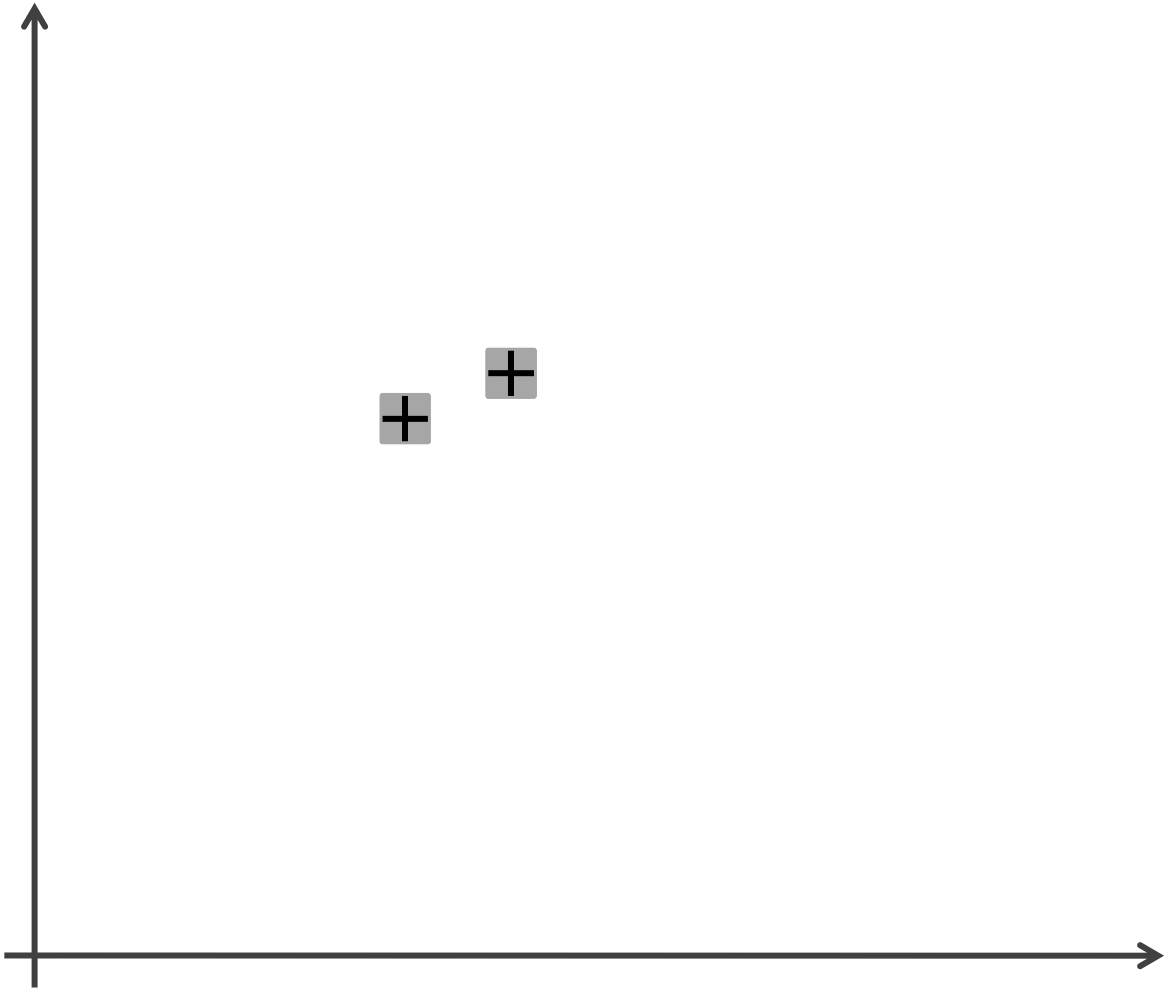


- In the setting of supervised learning with discriminative models, we suggest that model identification and data disambiguation can support each other, and should be performed simultaneously.
- Not only the data is telling us something about the model, but also the model (assumptions) about the data.

DATA DISAMBIGUATION

classes

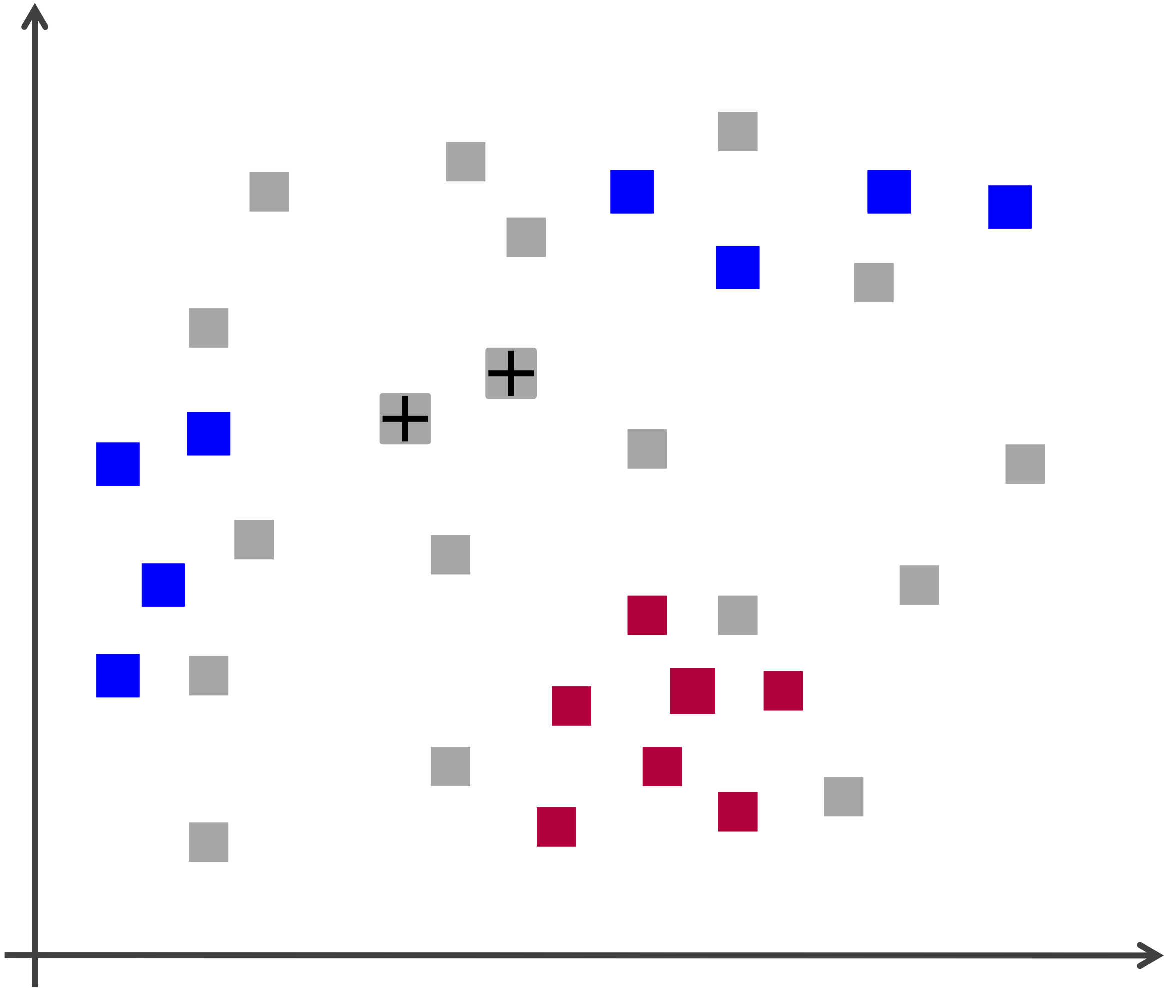
🟦 🟥



DATA DISAMBIGUATION

classes

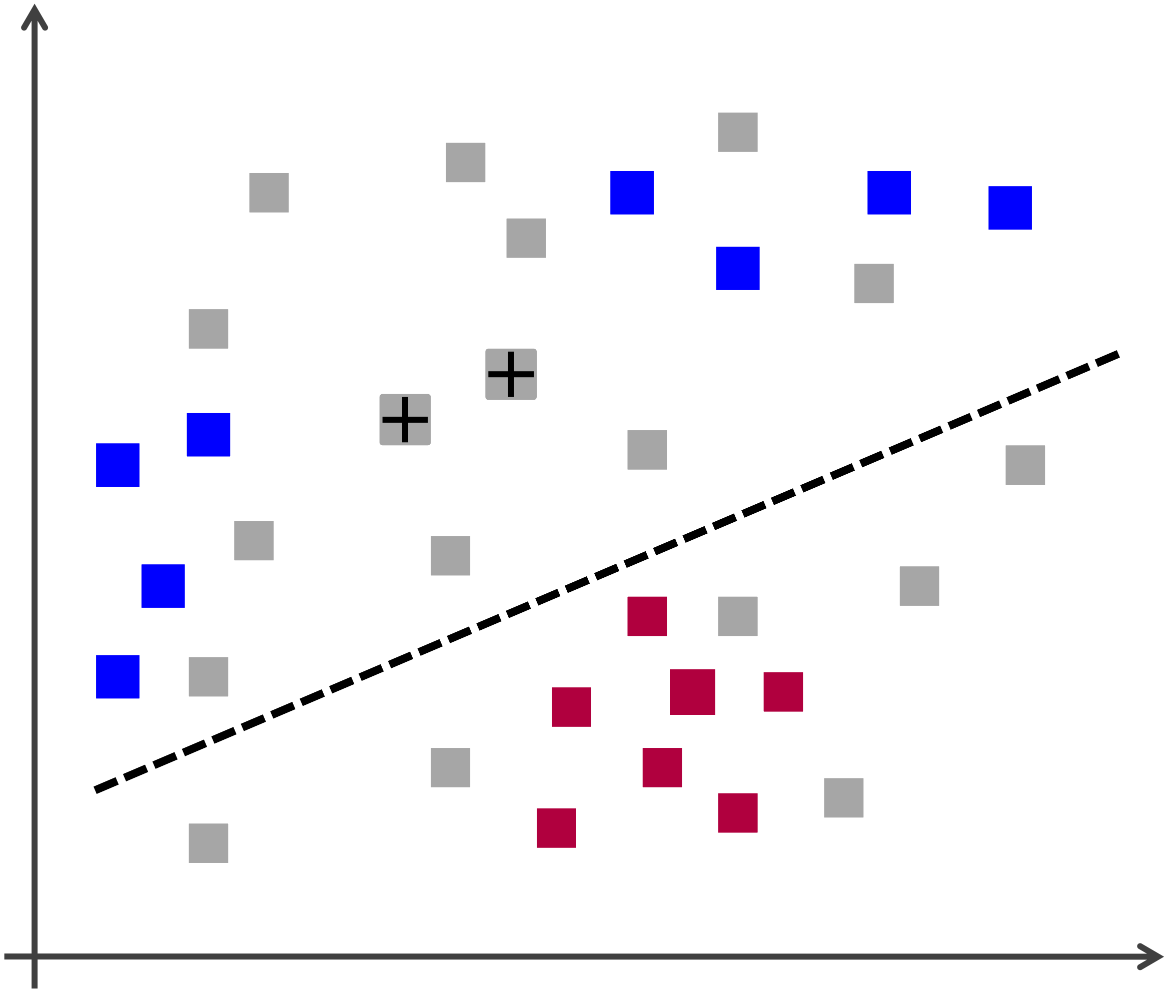
🟦 🟥



DATA DISAMBIGUATION

classes

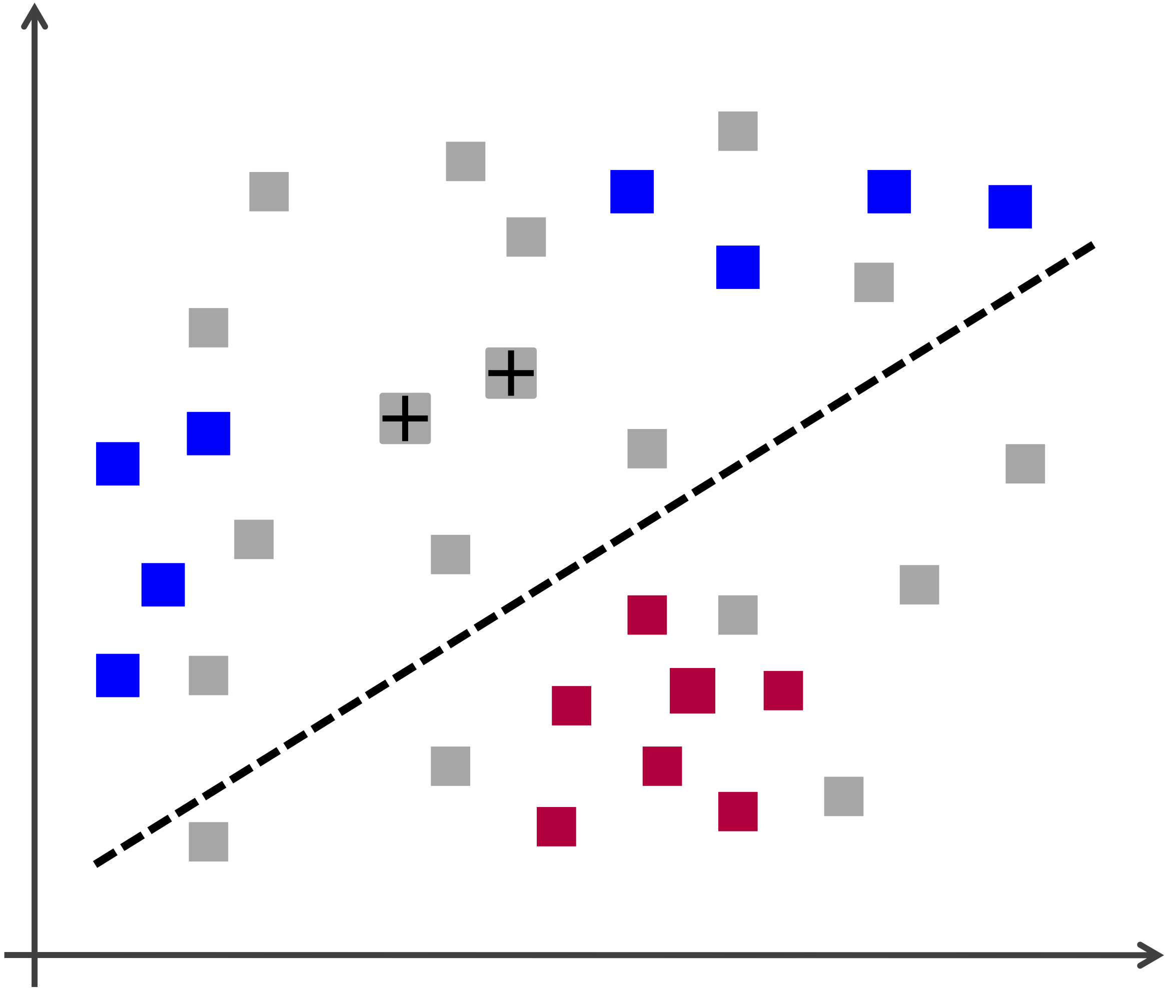
🟦 🟥



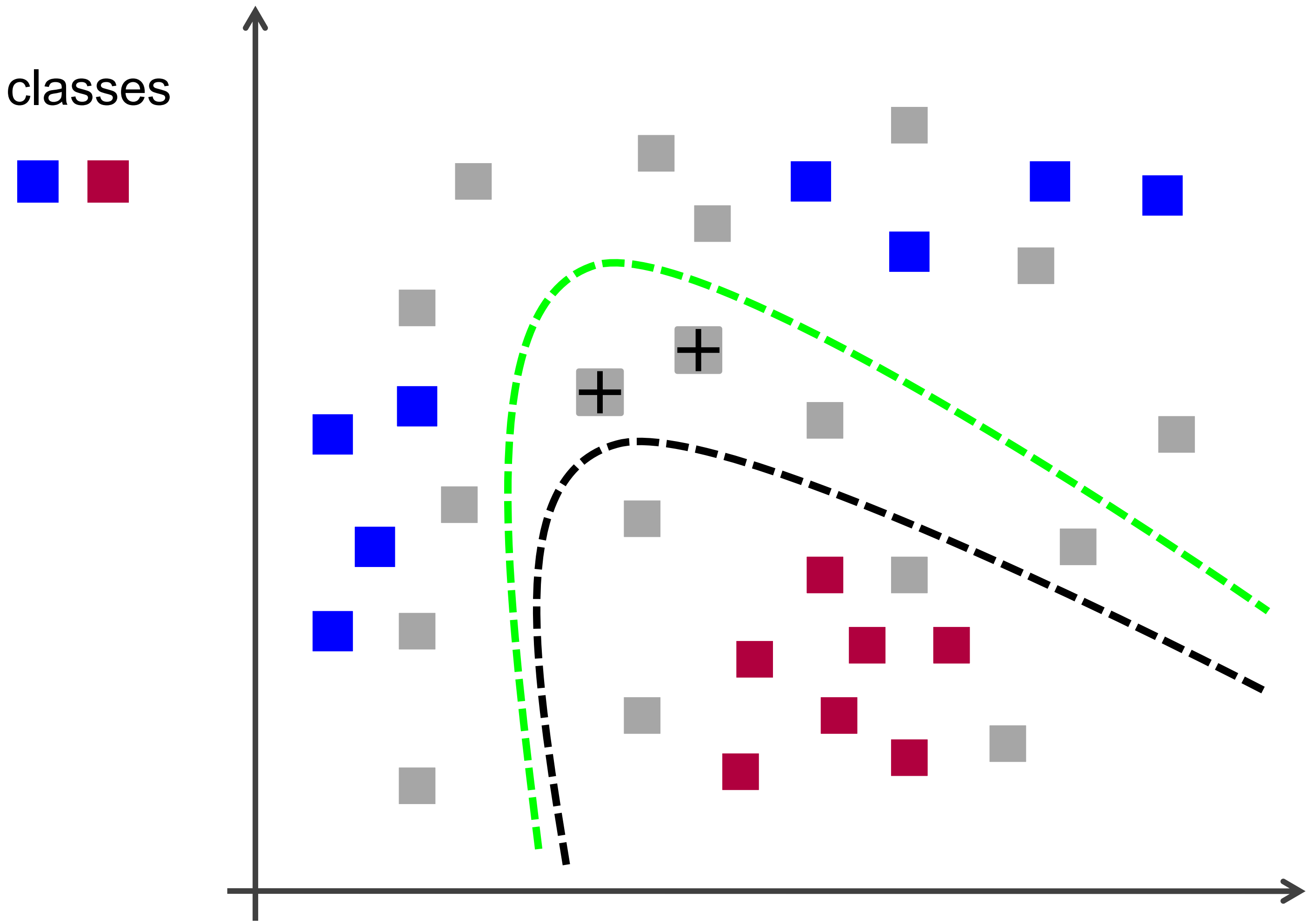
DATA DISAMBIGUATION

classes

🟦 🟥



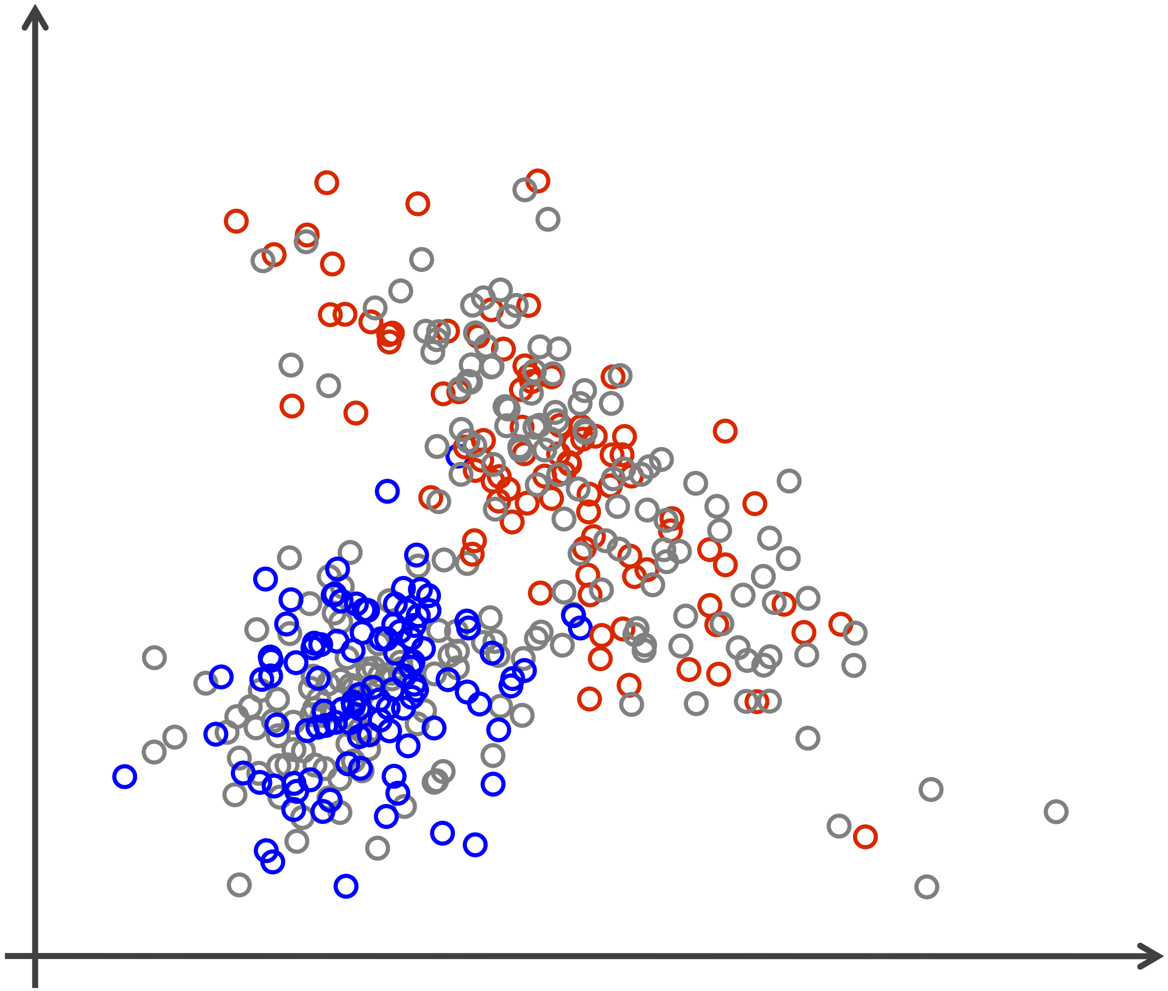
DATA DISAMBIGUATION



The more biased the view, the less ambiguous the data looks like.

DATA DISAMBIGUATION

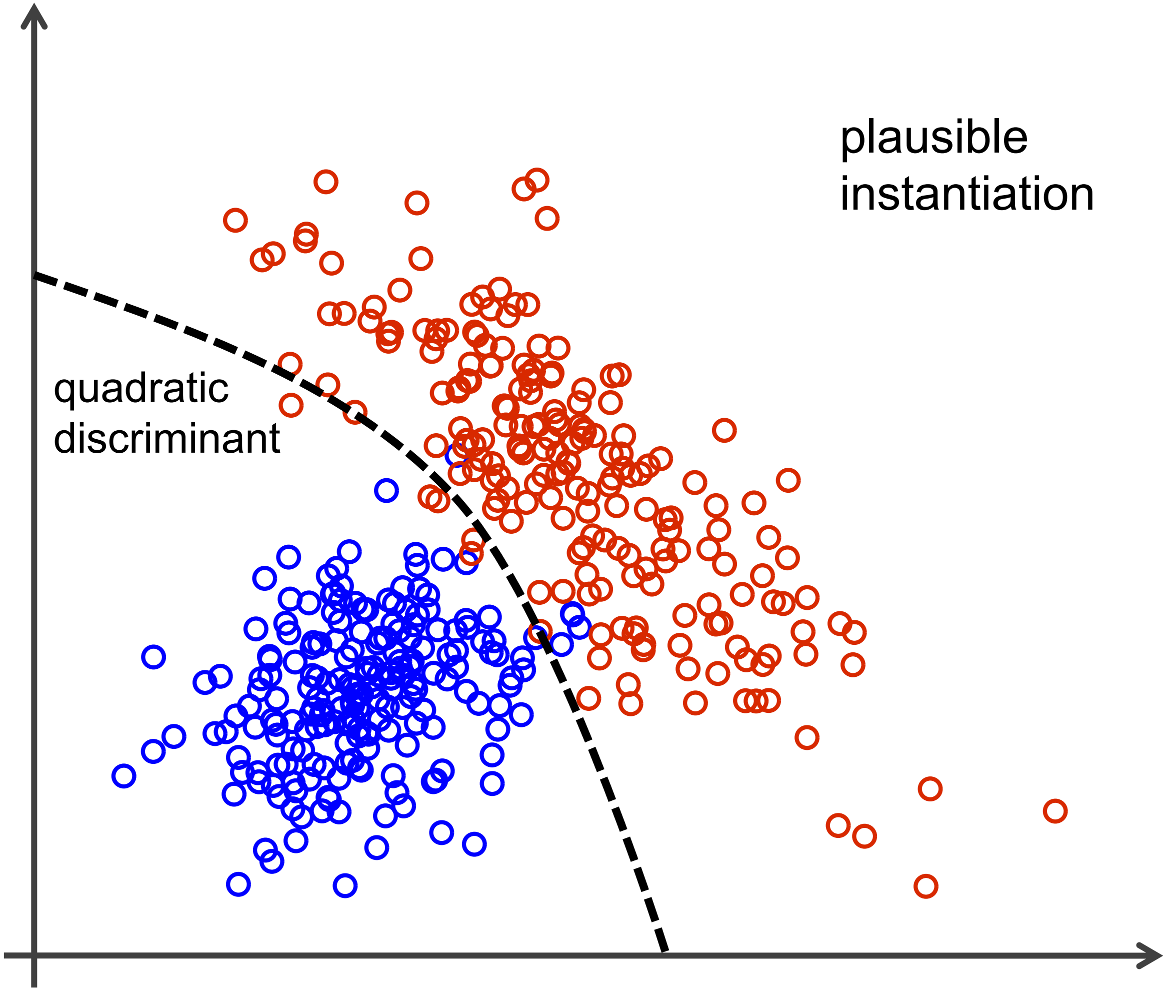
○ = { 🔴, 🔵 }



assume both class distributions to be Gaussian

DATA DISAMBIGUATION

⚪ = { 🔴 , 🔵 }

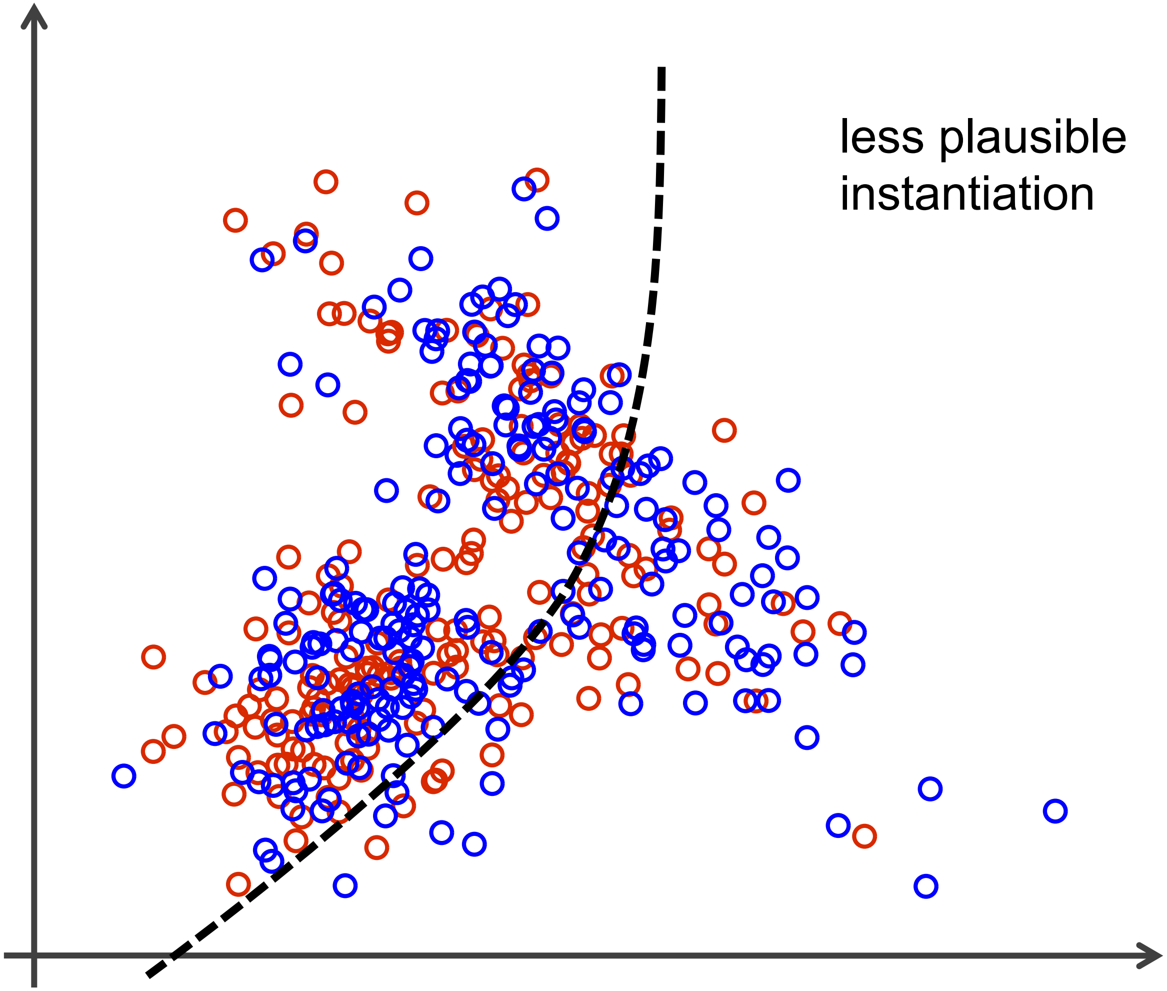


plausible
instantiation

assume both class distributions to be Gaussian

DATA DISAMBIGUATION

○ = { ○ , ○ }



less plausible instantiation

assume both class distributions to be Gaussian

OUTLINE

PART 1

Superset learning

PART 2

Optimistic loss minimization

PART 3

Data imprecisiation

EMPIRICAL RISK MINIMIZATION

Given a set of (i.i.d.) training data and a **hypothesis space** $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$, find a model with minimal **empirical risk**

$$\mathcal{R}_{emp}(h) = \frac{1}{N} \sum_{i=1}^{N} L\big(h(\boldsymbol{x}_i), y_i\big) \,.$$

In general, ERM won't work well (unless N is large)…

GENERALIZED ERM

We propose a principle of **generalized empirical risk minimization** with the empirical risk

$$\mathcal{R}^*_{emp}(h) = \frac{1}{N}\sum_{n=1}^{N} L^*\big(Y_n, h(\boldsymbol{x}_n)\big)$$

and the **optimistic superset loss** (OSL) function

$$L^*(Y, \hat{y}) = \min\big\{L(y, \hat{y}) \,|\, y \in Y\big\}\,.$$

how well the (precise) model fits the imprecise data

GENERALIZED ERM

We propose a principle of **generalized empirical risk minimization** with the empirical risk

$$\mathcal{R}^{**}_{emp}(h) = \frac{1}{N} \sum_{n=1}^{N} L^{**}\big(Y_n, h(\boldsymbol{x}_n)\big)$$

and the **optimistic fuzzy superset loss** (OFSL) function

$$L^{**}(Y, \hat{y}) = \int_0^1 L^{*}\Big([Y]_\alpha, \hat{y}\Big)\, d\alpha$$

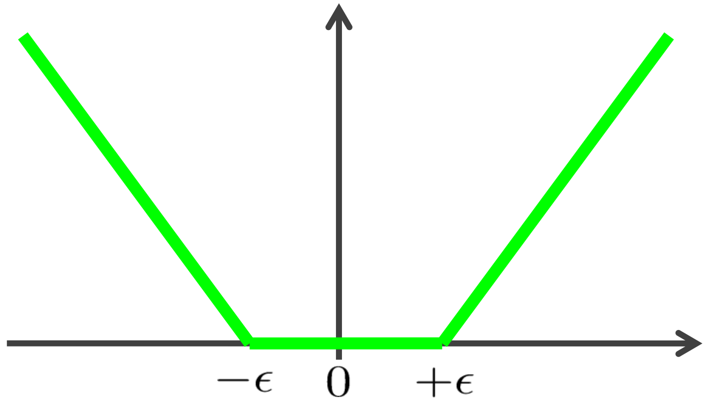
.

GENERALIZED ERM

- Generalized ERM derives from a likelihood-based approach, which proceeds from $\mathbf{P}(\mathcal{D}, \mathcal{O} \,|\, h)$,
- and makes (weak) assumptions about the coarsening $\mathbf{P}(\mathcal{O} \,|\, \mathcal{D}, h)$.
- Further, it exploits additivity of the loss.
- Finally, the logistic loss is replaced by any other loss function.

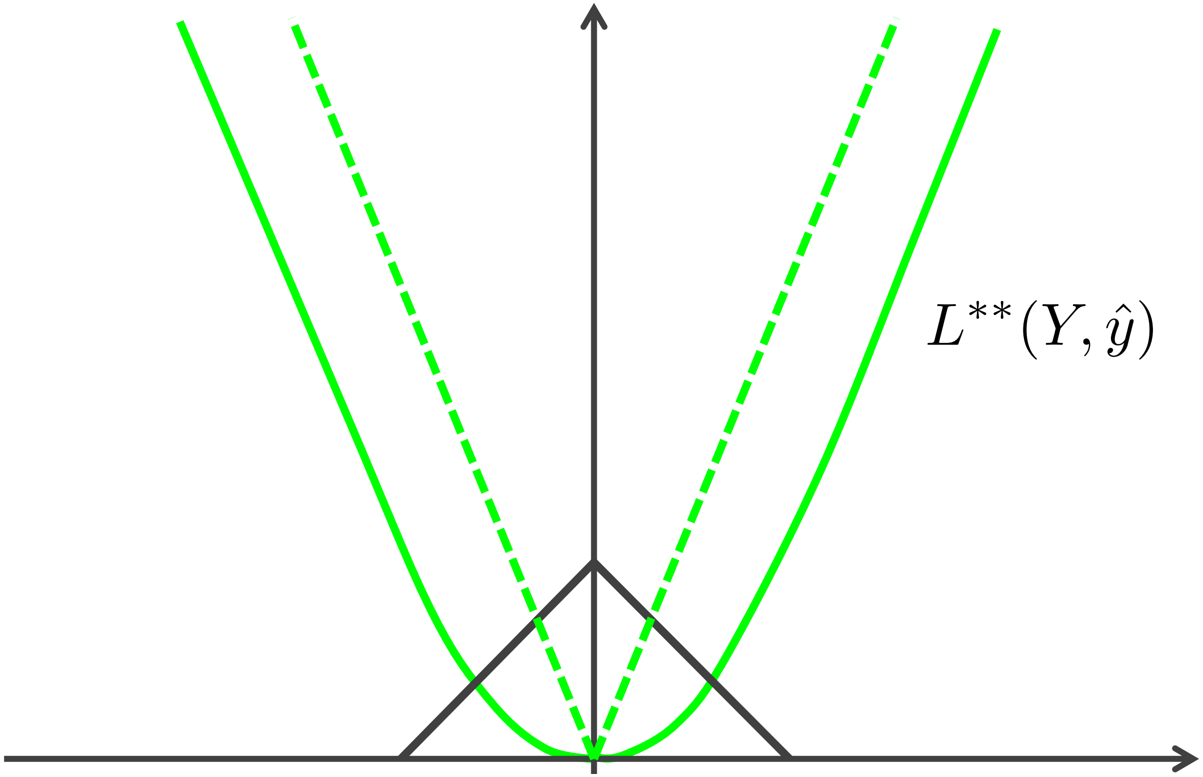
Why should generalized ERM actually work?

SPECIAL CASES



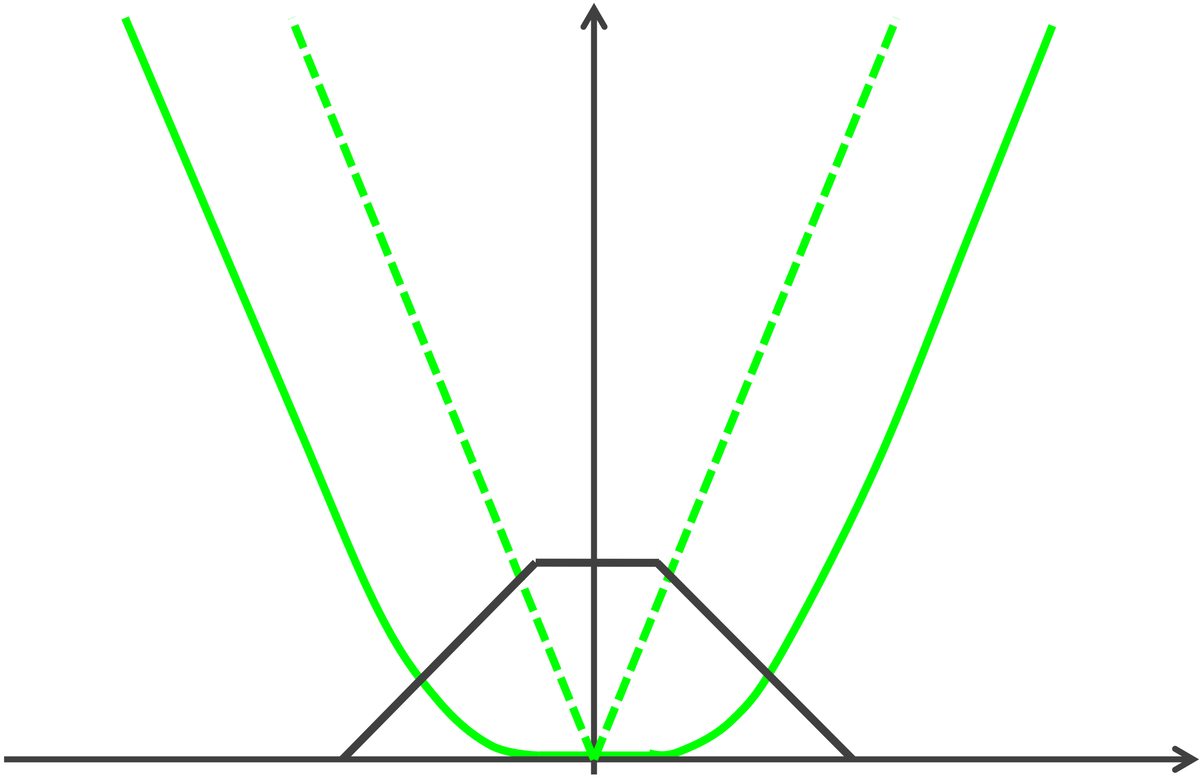
The ϵ-insensitive loss $L(y, \hat{y}) = \max(|y - \hat{y}| - \epsilon, 0)$ used in support vector regression corresponds to L^* with L the standard L_1 loss $L(y, \hat{y}) = |y - \hat{y}|$ and precise data y_n being replaced by interval-valued data $Y_n = [y_n - \epsilon, y_n + \epsilon]$.

SPECIAL CASES



Huber loss

SPECIAL CASES



(generalized) Huber loss

LABEL RANKING

The Kendall loss used in label ranking:

$$L(\pi, \hat{\pi}) = \sum_{i<j} \left[\!\left[\operatorname{sign}(\pi(i) - \pi(j)) \neq \operatorname{sign}(\hat{\pi}(i) - \hat{\pi}(j)) \right]\!\right]$$

- Cheng and H. (2015) compare an approach to label ranking based on superset learning with state-of-the-art approaches.
- Very strong performance, more robust toward incompleteness.

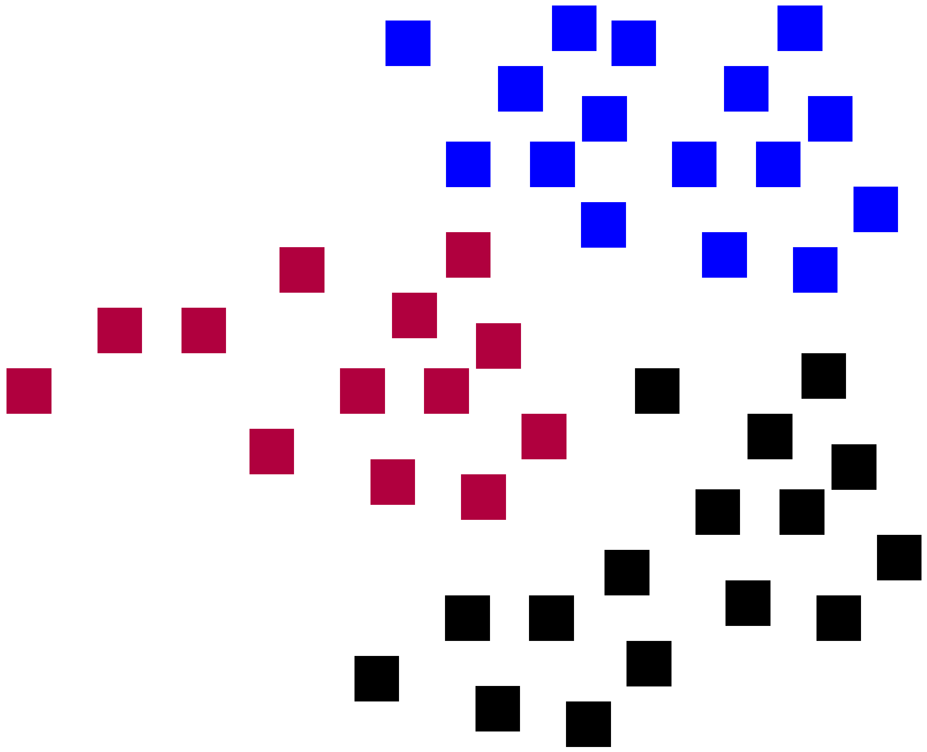
New methods as natural instantiations of the generalized ERM framework!

THEORETICAL FOUNDATIONS

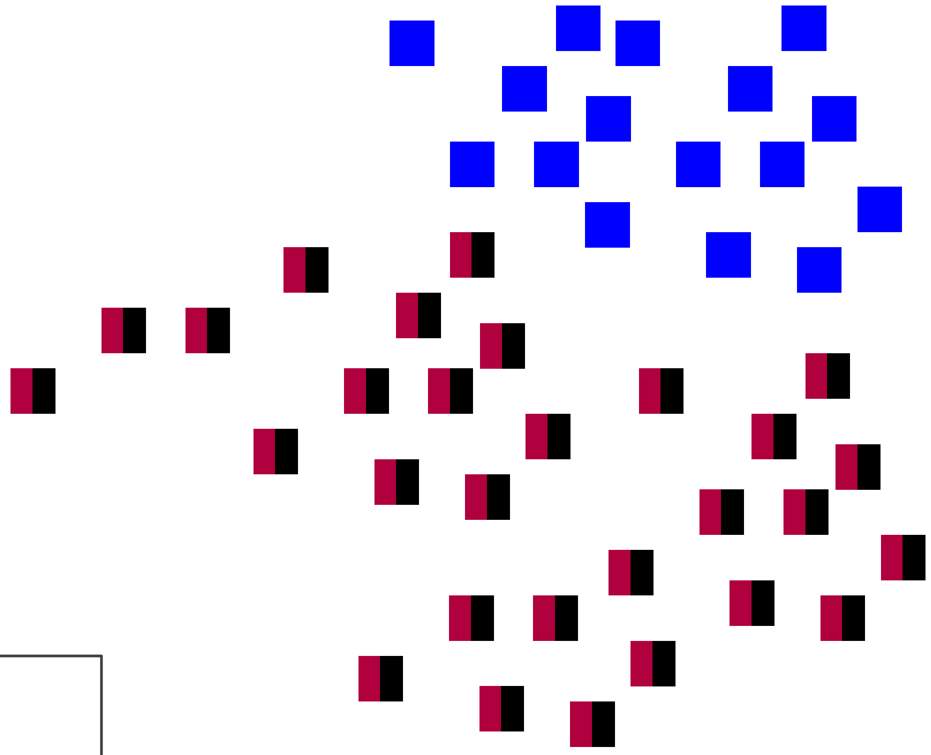
- Under what conditions is (successful) learning in the superset setting actually possible?
- Specifically, under what conditions does generalized ERM work?
- Couldn't the optimism induce a strong bias?
- Might other principles (pessimism, agnosticism) be better?

$$L^*(Y,\hat{y}) = \min\left\{L(y,\hat{y}) \,|\, y \in Y\right\}$$
$$L^*(Y,\hat{y}) = \operatorname{avg}\left\{L(y,\hat{y}) \,|\, y \in Y\right\}$$
$$L^*(Y,\hat{y}) = \max\left\{L(y,\hat{y}) \,|\, y \in Y\right\}$$

THEORETICAL FOUNDATIONS



THEORETICAL FOUNDATIONS



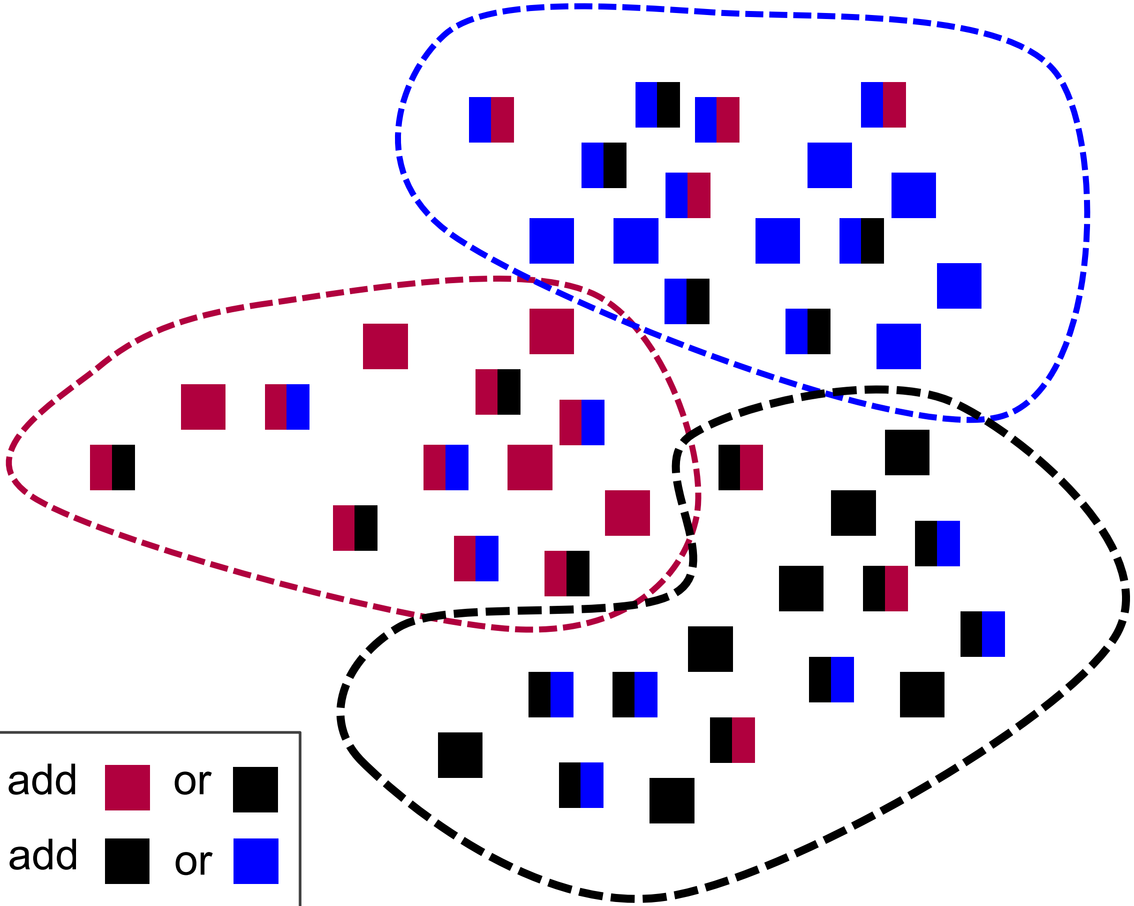
🟦 →

🟥 → add ⬛

⬛ → add 🟥

systematic (adversarial) coarsening

THEORETICAL FOUNDATIONS



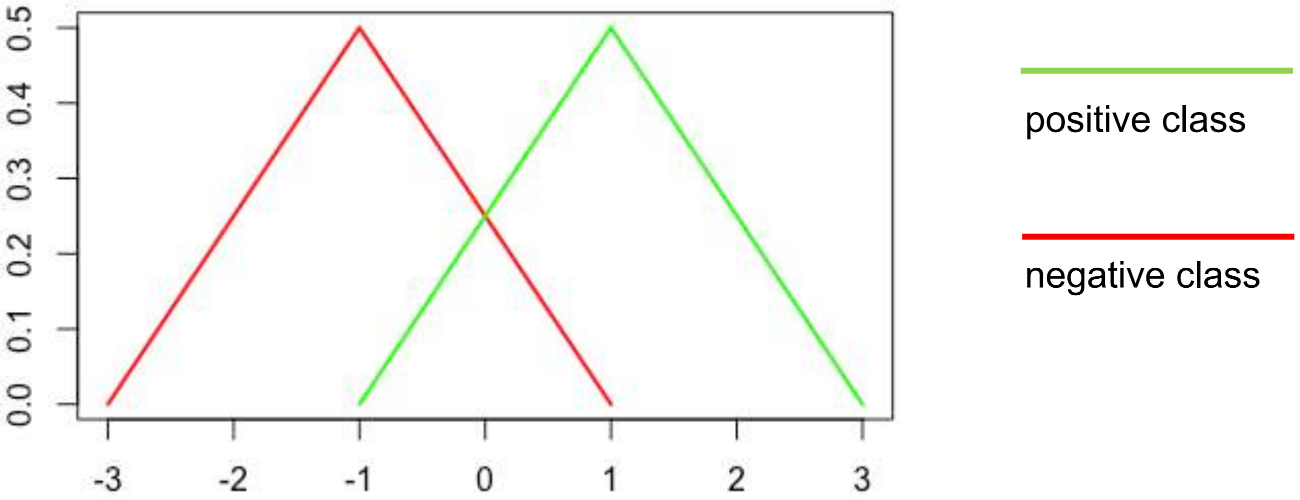
🟦 → add 🟥 or ⬛

🟥 → add ⬛ or 🟦

⬛ → add 🟥 or 🟦

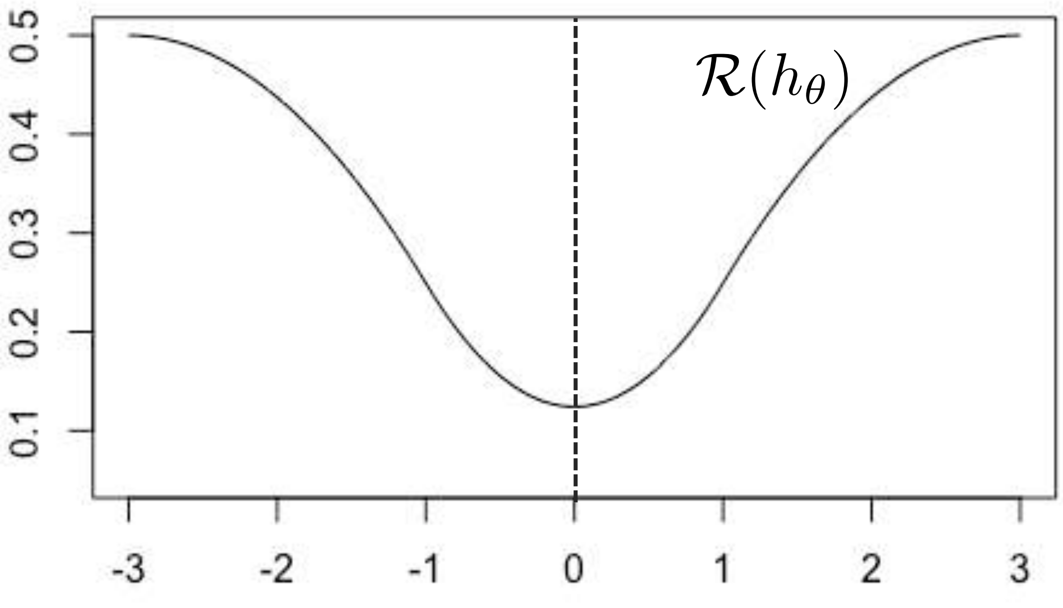
non-systematic (random) coarsening

AN EXAMPLE



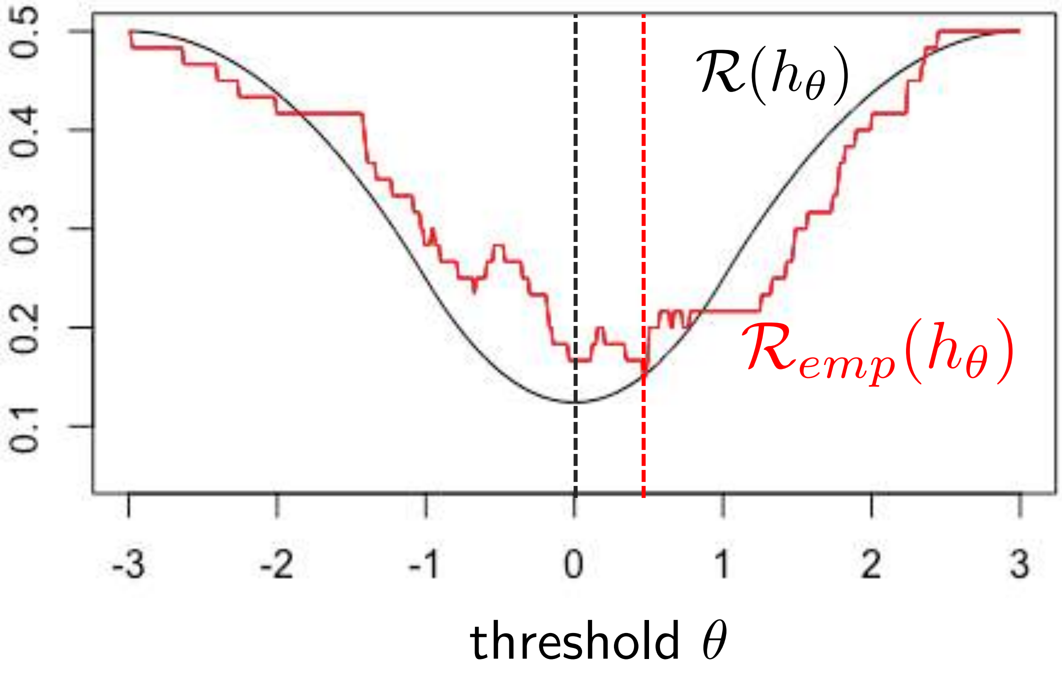
$$h_\theta(x) = \begin{cases} +1\,, & x \geq \theta \\ -1\,, & x < \theta \end{cases}$$

AN EXAMPLE



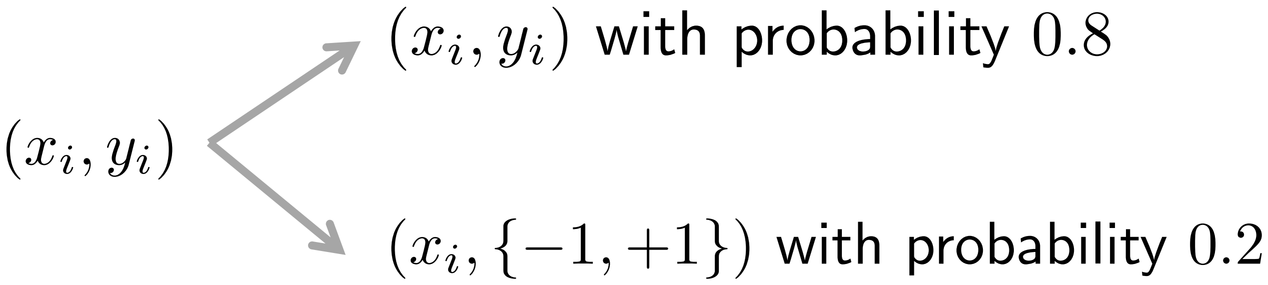
threshold θ

AN EXAMPLE


0.1
0.2
0.3
0.4
0.5
-3
-2
-1
0
1
2
3
$\mathcal{R}(h_\theta)$
$\mathcal{R}_{emp}(h_\theta)$
threshold θ

AN EXAMPLE

All examples are coarsened with probability 0.2.



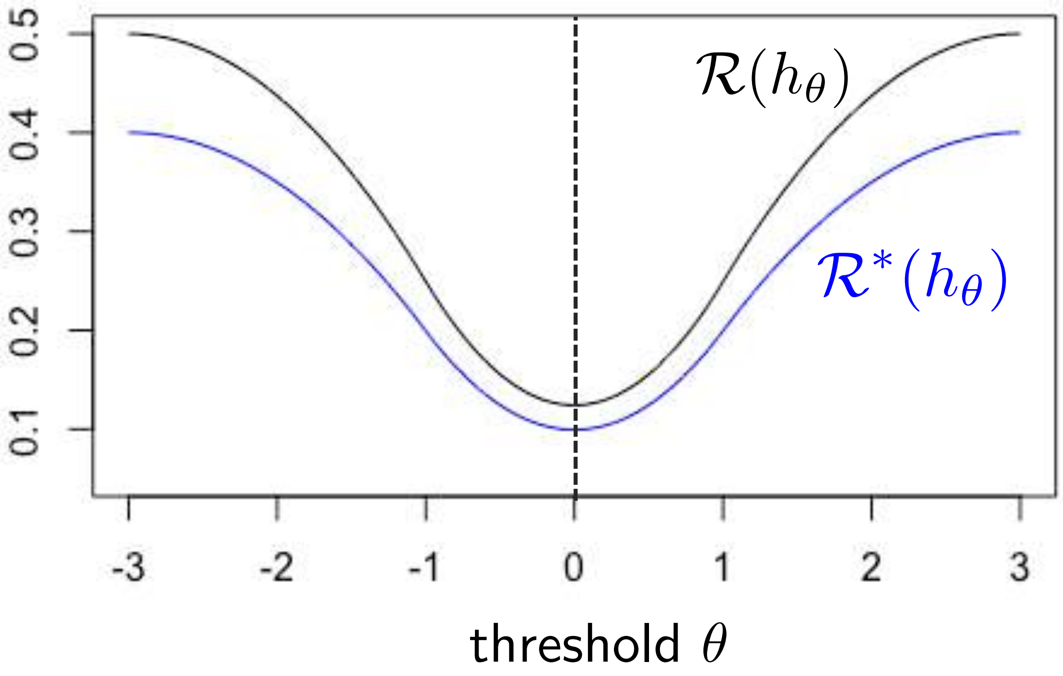
(x_i, y_i)

(x_i, y_i) with probability 0.8

$(x_i, \{-1, +1\})$ with probability 0.2

AN EXAMPLE

All examples are coarsened with probability 0.2.



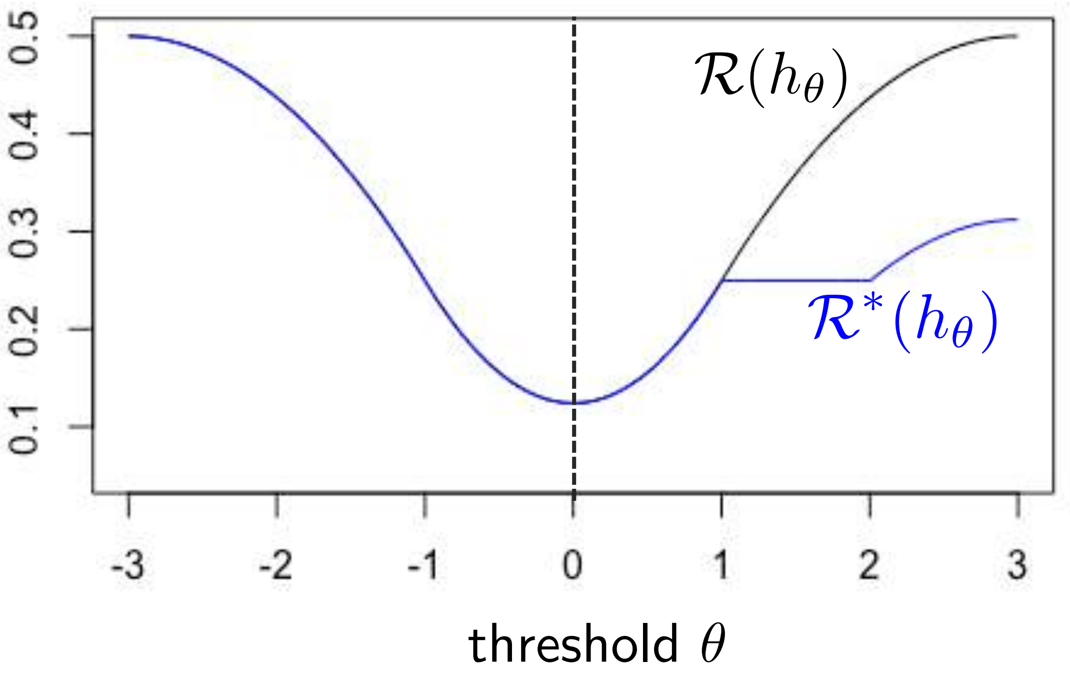
AN EXAMPLE

Examples with x between 1 and 2 are coarsened.

$$\begin{tikzcd}
 & (x_i, \{-1, +1\}) \text{ if } x_i \in [1, 2] \\
(x_i, y_i) \arrow[ur] \arrow[dr] & \\
 & (x_i, y_i) \text{ otherwise}
\end{tikzcd}$$

AN EXAMPLE

Examples with x between 1 and 2 are coarsened.



AN EXAMPLE

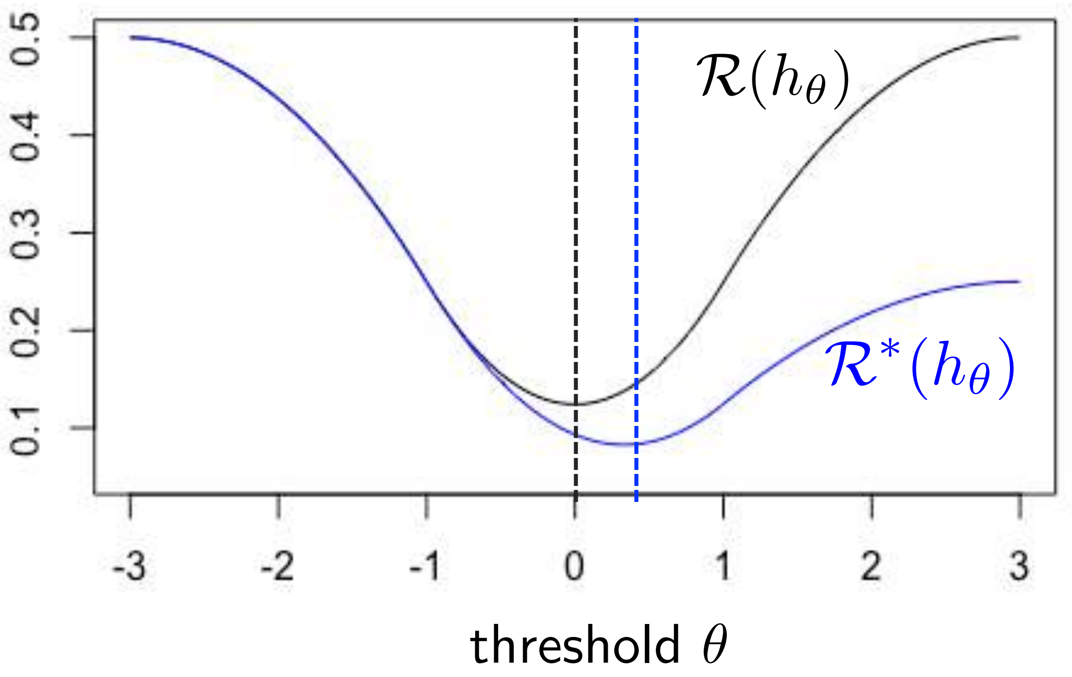
Positive examples are coarsened with probability $1/2$.

$$\begin{tikzcd}
 & (x_i, +1) \text{ with probability } 0.5 \\
(x_i, +1) \arrow[ur] \arrow[dr] & \\
 & (x_i, \{-1, +1\}) \text{ with probability } 0.5
\end{tikzcd}$$

$$\begin{tikzcd}
(x_i, -1) \arrow[r] & (x_i, -1)
\end{tikzcd}$$

AN EXAMPLE

Positive examples are coarsened with probability $1/2$.



THEORETICAL FOUNDATIONS

The **balanced benefit condition**:

$$0 \leq \eta_1 \leq \inf_{h \in \mathcal{H}} \frac{\mathcal{R}^*(h)}{\mathcal{R}(h)} \leq \sup_{h \in \mathcal{H}} \frac{\mathcal{R}^*(h)}{\mathcal{R}(h)} \leq \eta_2 \leq 1 \ ,$$

where $\mathcal{R}^*(h)$ is the expected superset loss of h.

For sufficiently large sample size,

$$\mathcal{R}(\hat{h}) \leq \mathcal{R}(h^*) + \Delta(d_{\mathcal{H}}, \epsilon, \delta, \eta_1, \eta_2) \ ,$$

with probability $1 - \delta$, where $d_{\mathcal{H}}$ is the Natarajan dimension of $\mathcal{H}$, h^* the Bayes predictor and $\hat{h}$ the minimizer of $\mathcal{R}^*_{emp}$.

THEORETICAL FOUNDATIONS

Liu and Dietterich (2014) consider the **ambiguity degree**, which is defined as the largest probability that a particular **distractor** label co-occurs with the true label in multi-class classification:

$$\gamma = \sup\left\{\mathbf{P}_{Y\sim\mathcal{D}^s(\boldsymbol{x},y)}(\ell \in Y) \,|\, (\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}, \ell \in \mathcal{Y}, p(\boldsymbol{x}, y) > 0, \ell \neq y\right\}$$

Let $\theta = \log(2/(1+\gamma))$ and $d_{\mathcal{H}}$ the Natarajan dimension of $\mathcal{H}$. Define

$$n_0(\mathcal{H}, \epsilon, \delta) = \frac{4}{\theta\epsilon}\Bigg(d_{\mathcal{H}}\Bigg(\log(4d_{\mathcal{H}} + 2\log L + \log\left(\frac{1}{\theta\epsilon}\right)\Bigg)\Bigg) + \log\left(\frac{1}{\delta}\right) + 1\Bigg).$$

Then, in the realizable case, with probability at least $1-\delta$, the model with the smallest **empirical superset loss** on a set of training data of size $n > n_0(\mathcal{H}, \epsilon, \delta)$ has a **generalisation error** of at most ϵ.

OUTLINE

INTELLIGENT SYSTEMS

PART 1

Superset learning

PART 2

Optimistic loss minimization

PART 3

Data imprecisiation

DATA IMPRECISIATION

So far: Imprecision as a necessary evil

Observations are imprecise/incomplete, and we have to deal with that!

Now: Imprecision as a means for modeling

Deliberately turn precise into imprecise data, so as to modulate the influence of an observation on the learning process!

Motivated by the following monotonicity property:

$$Y \subset Y' \quad \Rightarrow \quad L^*(Y, \cdot) \geq L^*(Y', \cdot)$$

EXAMPLE WEIGHING

We suggest an alternative way of **weighing examples**, namely, via **„data imprecisiation“** ...

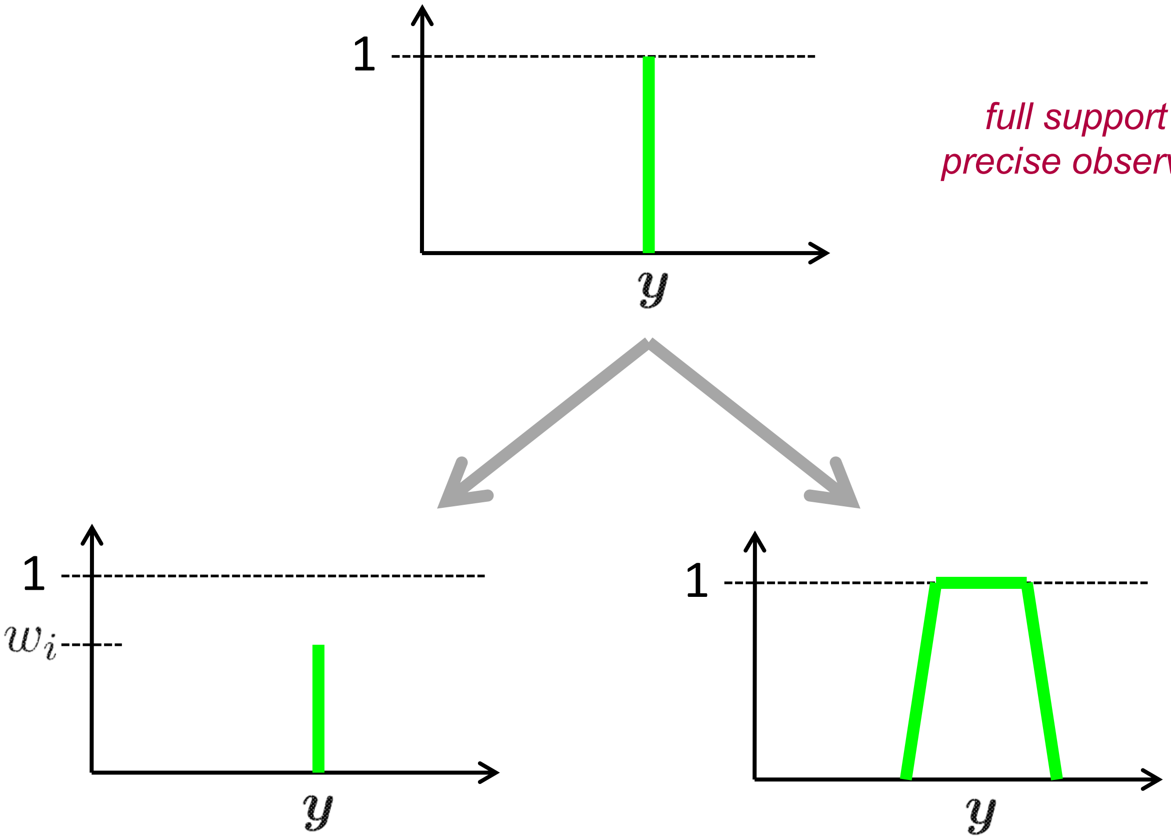
$$\begin{tikzcd}
& L(y_i, \cdot) \arrow[dl] \arrow[dr] & \\
w_i \, L(y_i, \cdot) & & L^*(Y_i, \cdot)
\end{tikzcd}$$

modulating the influence of a training example $(\boldsymbol{x}_i, y_i)$ by multiplying the loss with a constant w_i.

modulating the influence of a training example $(\boldsymbol{x}_i, y_i)$ by coarsening the observation y_i.

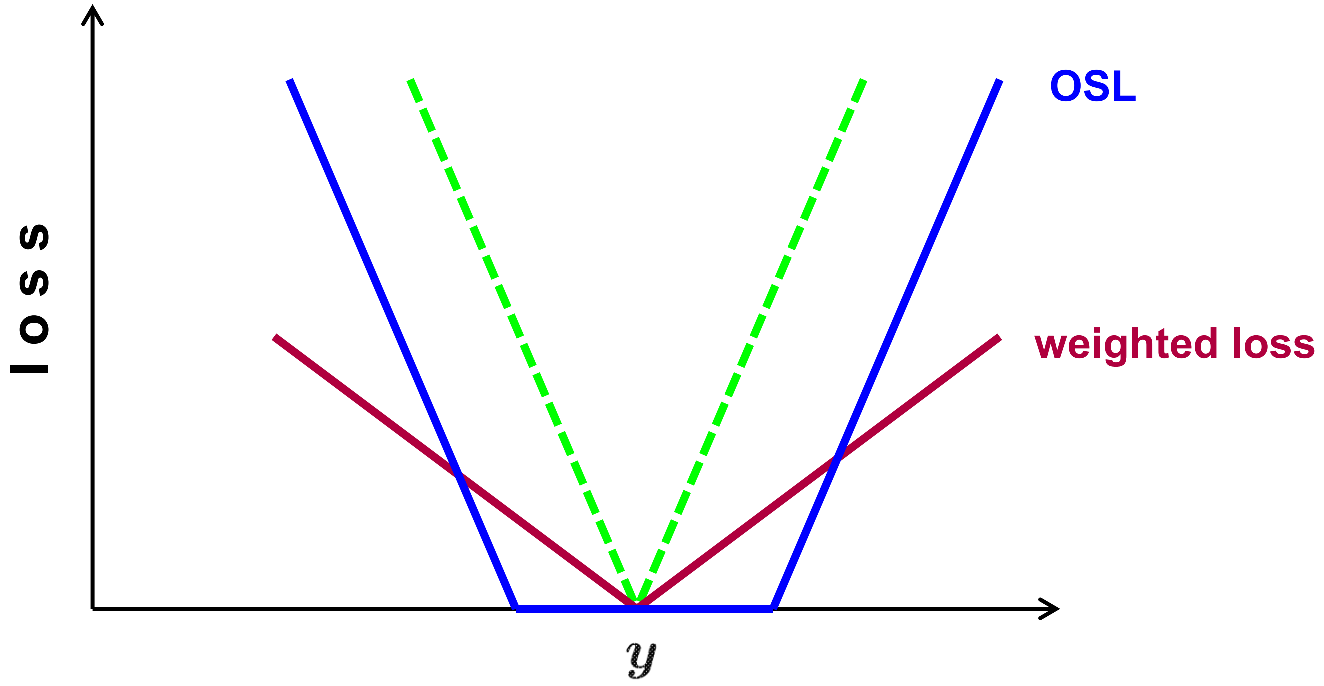
EXAMPLE WEIGHING

We suggest an alternative way of **weighing examples**, namely, via **„data imprecisiation“** ...



full support for precise observation

EXAMPLE WEIGHING

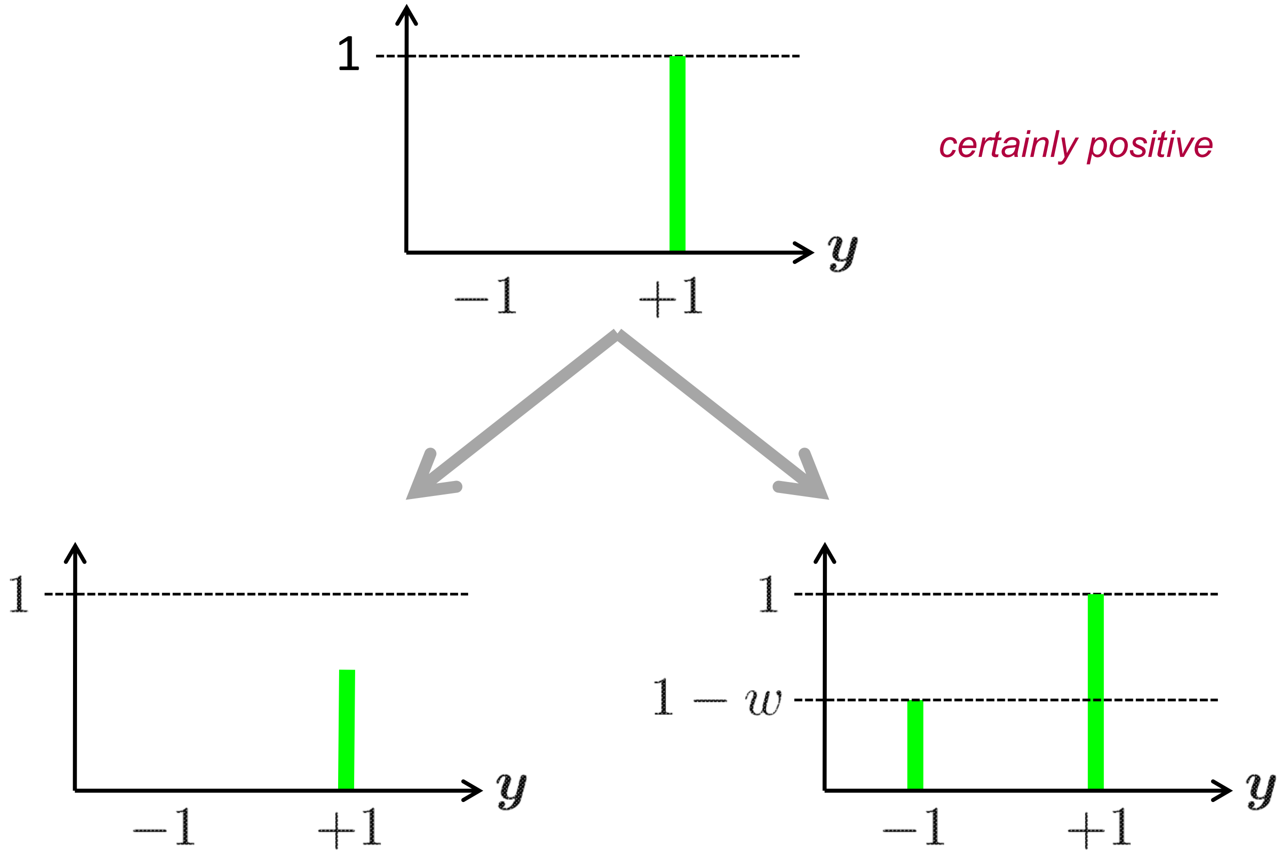


Different ways of (individually) discounting the loss function.

In (Lu and H., 2015), we empirically compared standard **locally weighted linear regression** with this approach and essentially found no difference.

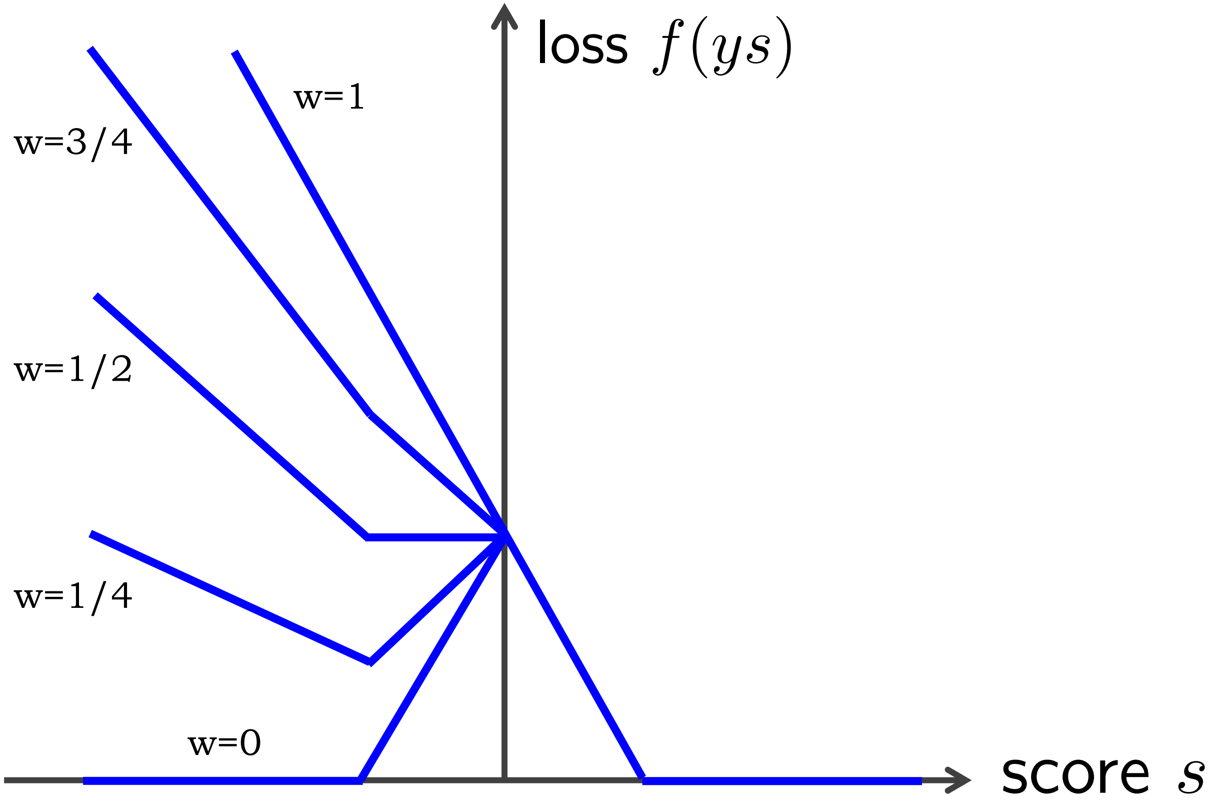
EXAMPLE WEIGHING

We suggest an alternative way of weighing examples, namely, via **„data imprecisiation“** ...



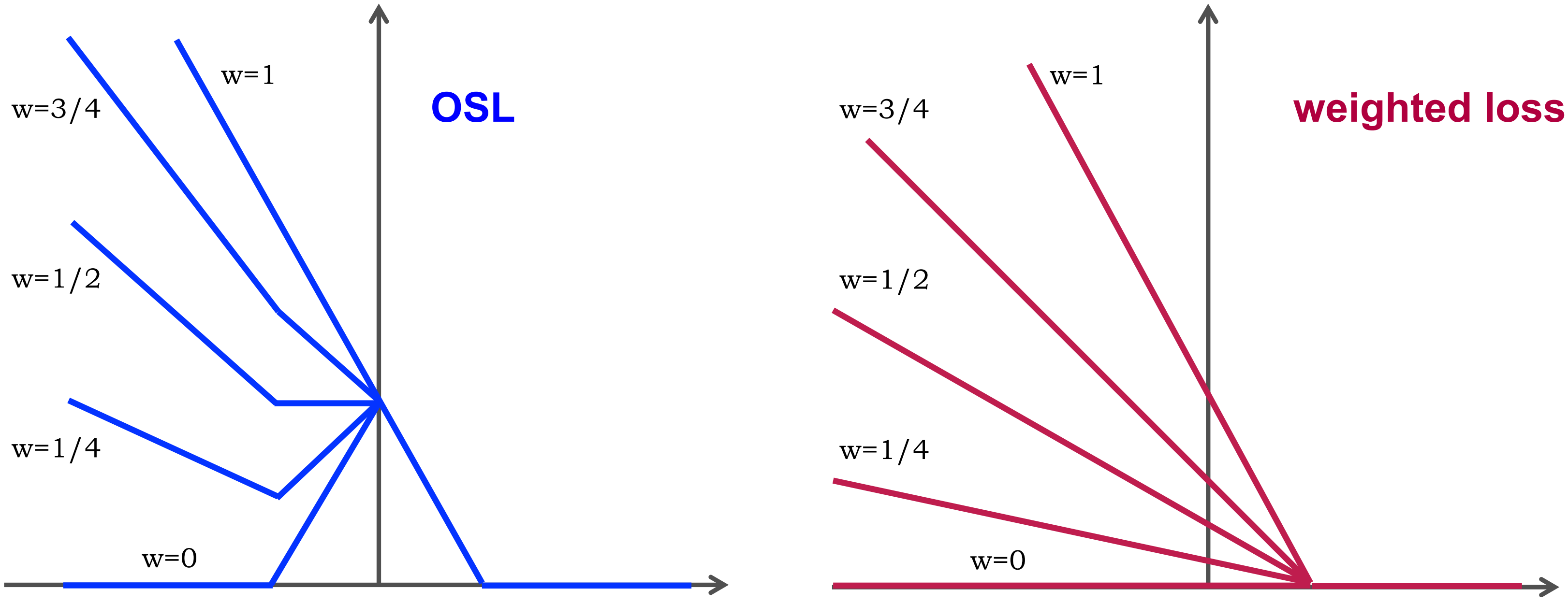
certainly positive

FUZZY MARGIN LOSSES



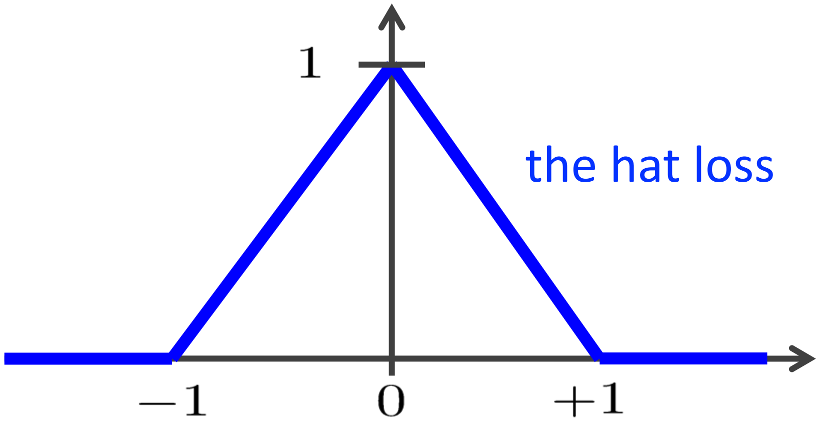
GENERALIZED HINGE LOSS

FUZZY MARGIN LOSSES



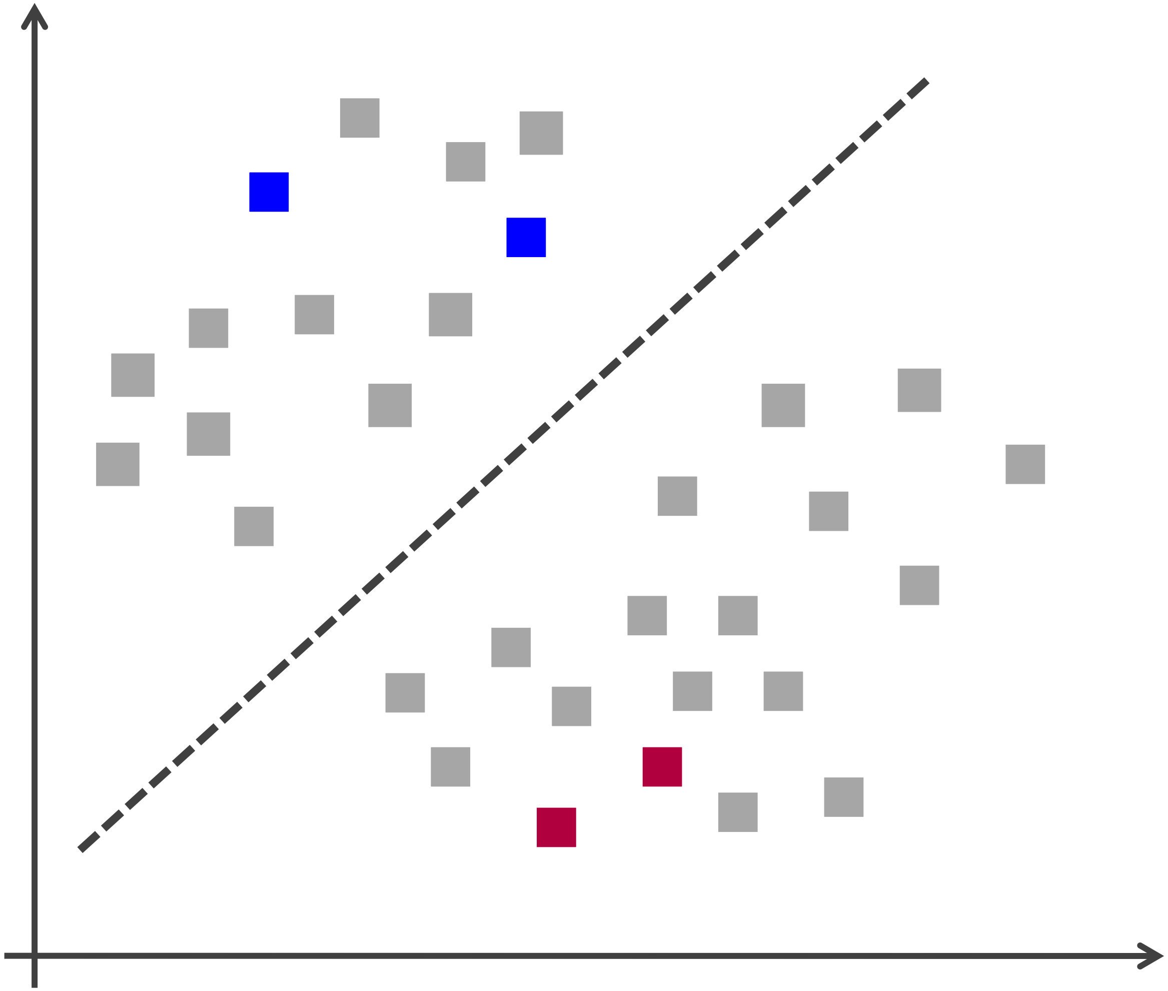
Different ways of (individually) discounting the loss function.

THE HAT LOSS

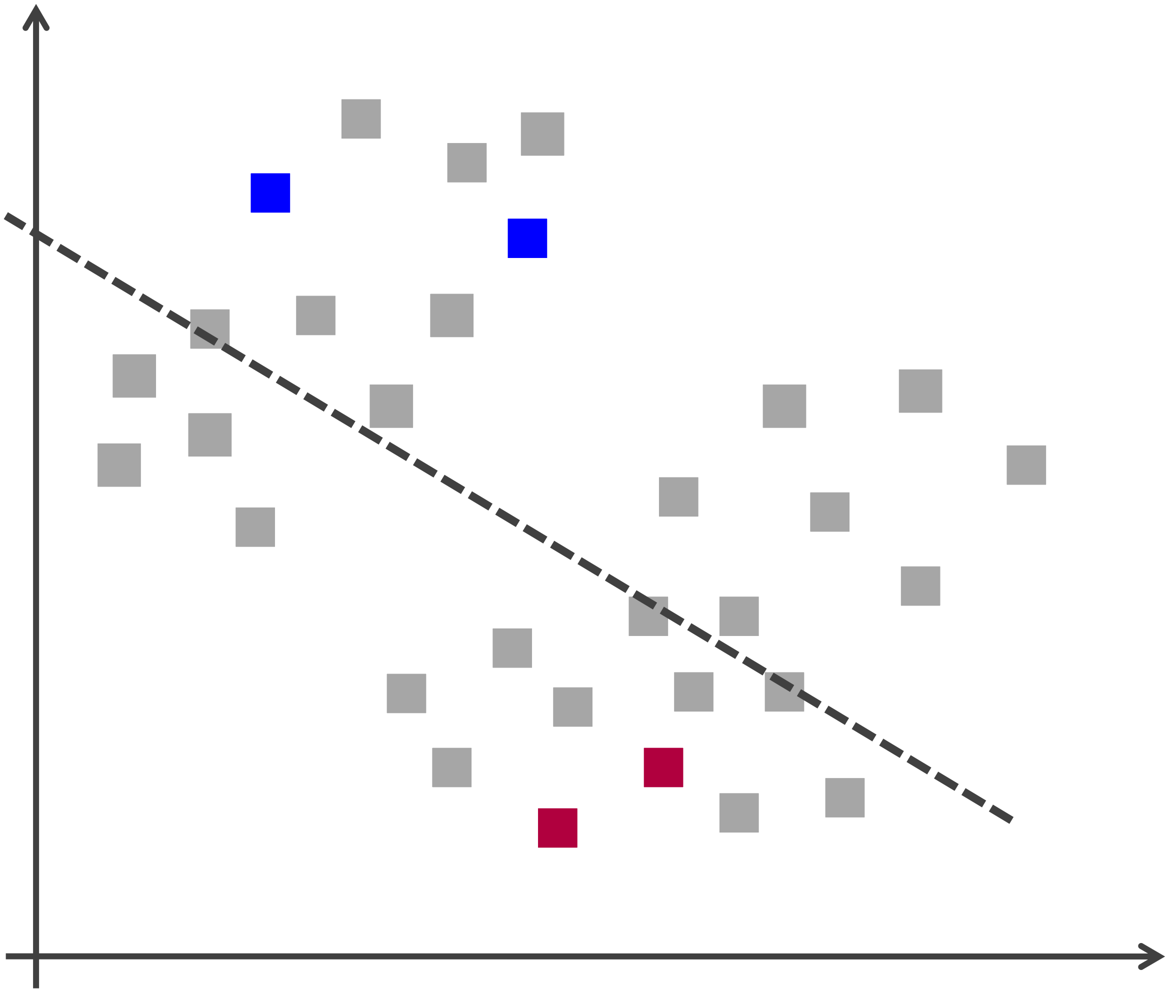


Semi-supervised learning with SVMs: Consider unlabeled data as instances labeled with the superset $\{-1, +1\}$. The generalized loss L^* with L the standard hinge loss then corresponds to the (non-convex) “hat loss”.

DATA DISAMBIGUATION



DATA DISAMBIGUATION



Robust loss minimization for SVM:

- **Robust truncated-hinge-loss support vector machines** (RSVM) trains SVMs with the a truncated version of the hinge loss in order to be more robust toward outliers and noisy data (Wu and Liu, 2007).
- **One-step weighted SVM** (OWSVM) first trains a standard SVM. Then, it weighs each training example based on its distance to the decision boundary and retrains using the weighted hinge loss (Wu and Liu, 2013).
- **Our approach** (FLSVM) is the same as OWSVM, except for the weighted loss: instead of using a simple weighting of the hinge loss, we use the OSL.

Promising first results, especially competitive in the high-noise regime.

SUMMARY AND OUTLOOK

- Method for **superset learning** based on **optimistic loss minimization**, performing simultaneous model identification and data disambiguation.
- Our framework covers several **existing methods** as special cases but also supports the systematic development of **new methods**.
- **Completely generic principle** (classification, regression, structured output prediction, ...)

- Example weighing via **data imprecisiation** (→ „modeling data“)
- Works for regression and classification, but seems to be even more interesting for other problems, including ranking, transfer learning, ...
- **More future work**: Algorithmic solutions for specific instantiations of our framework, theoretical foundations, non-additive losses, ...

REFERENCES

E. Hüllermeier (2014). **Learning from Imprecise and Fuzzy Observations: Data Disambiguation through Generalized Loss Minimization.** International Journal of Approximate Reasoning, 55(7):1519-1534, 2014.

first paper introducing the general framework

E. Hüllermeier and W. Cheng (2015). **Superset Learning Based on Generalized Loss Minimization**. Proc. ECML/PKDD 2015.

instantiation for label ranking

S. Lu and E. Hüllermeier. **Locally Weighted Regression through Data Imprecisiation.** Workshop Computational Intelligence, Dortmund, 2015.

instantiation for locally weighted regression

S. Lu and E. Hüllermeier. **Support Vector Classification on Noisy Data using Fuzzy Superset Losses**. Workshop Computational Intelligence, Dortmund, 2016.

instantiation for noise-tolerant classification