UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

SUPERSET LEARNING AND DATA IMPRECISIATION

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

eyke@upb.de

OUTLINE

INTELLIGENT SYSTEMS

PART 1	PART 2	PART 3
Superset learning	Optimistic loss minimization	Data imprecisiation
What it is about	*A general approach to superset learning ....*	*Using superset learning for weighted learning ...*

SUPERSET LEARNING

INTELLIGENT SYSTEMS

... 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).

- Consider a standard setting of **supervised learning** with instance space $\mathcal{X}$, output space $\mathcal{Y}$, and hypothesis space $\mathcal{H}$
- Output values $y_n \in \mathcal{Y}$ associated with training instances $\boldsymbol{x}_n$, $n = 1, \ldots, N$, are not necessarily observed precisely but only characterised in terms of **supersets**

$$Y_n \ni y_n \quad .$$

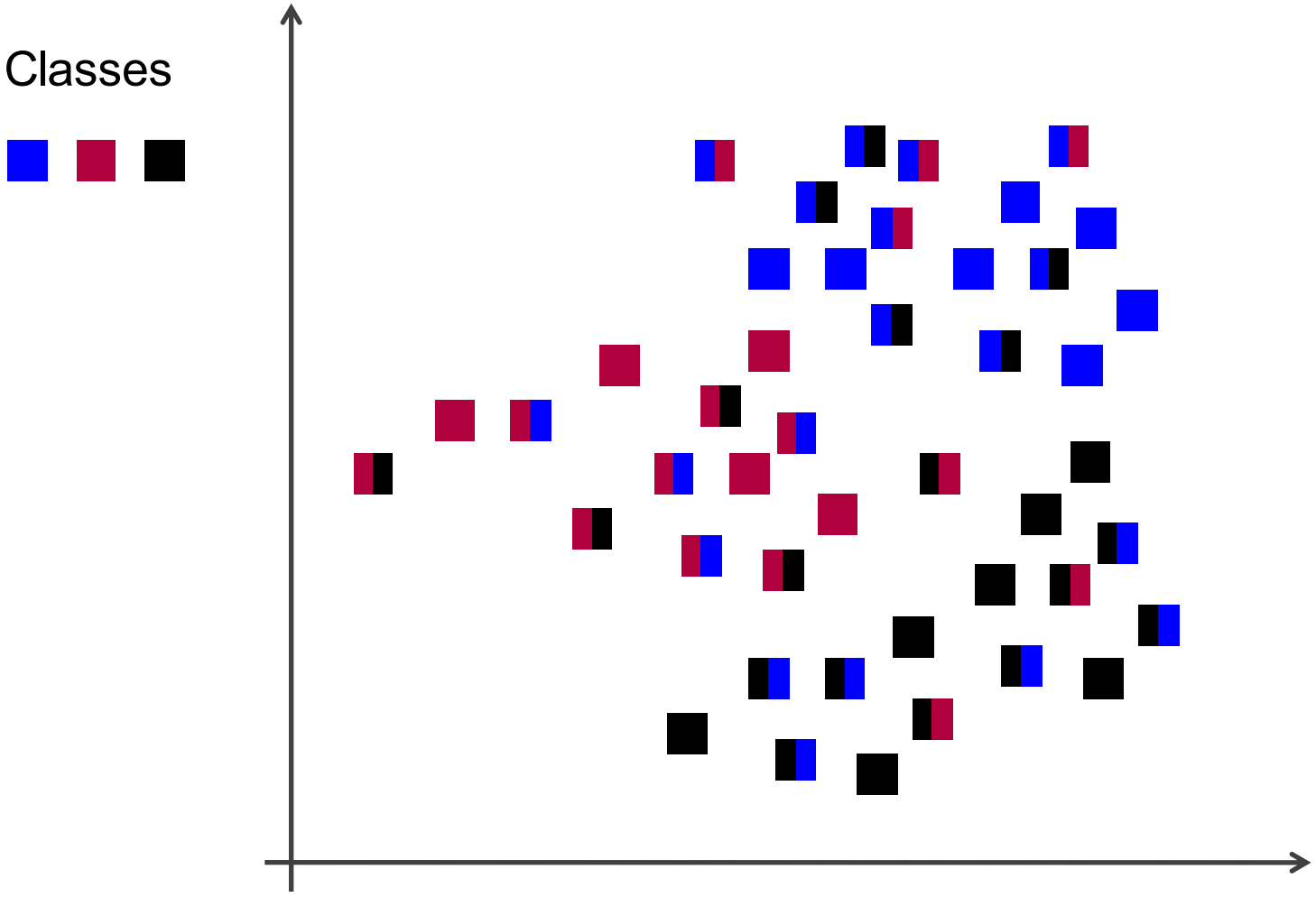
- Set of imprecise/ambiguous/coarse observations is denoted

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

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

EXAMPLE: CLASSIFICATION

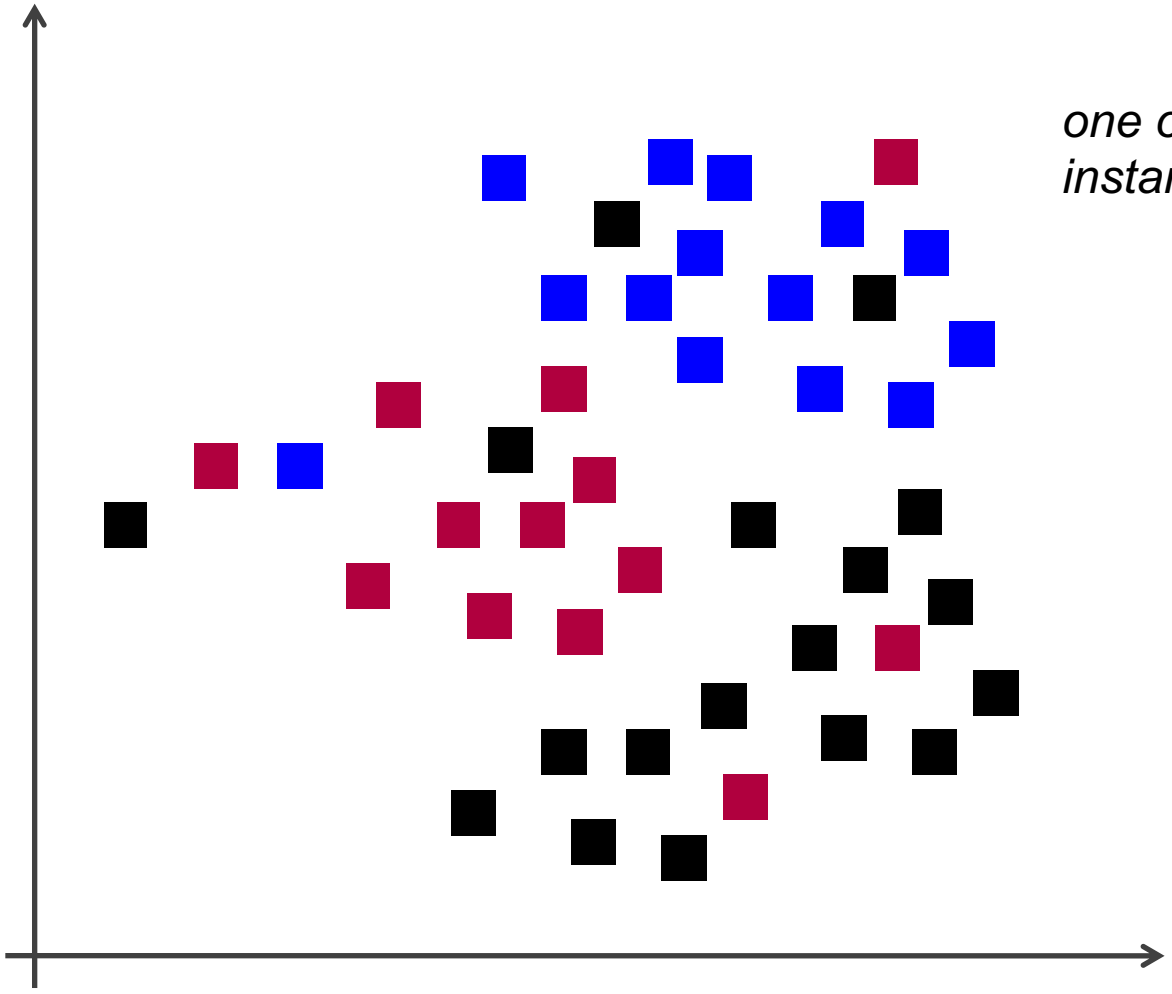
INTELLIGENT SYSTEMS



EXAMPLE: CLASSIFICATION

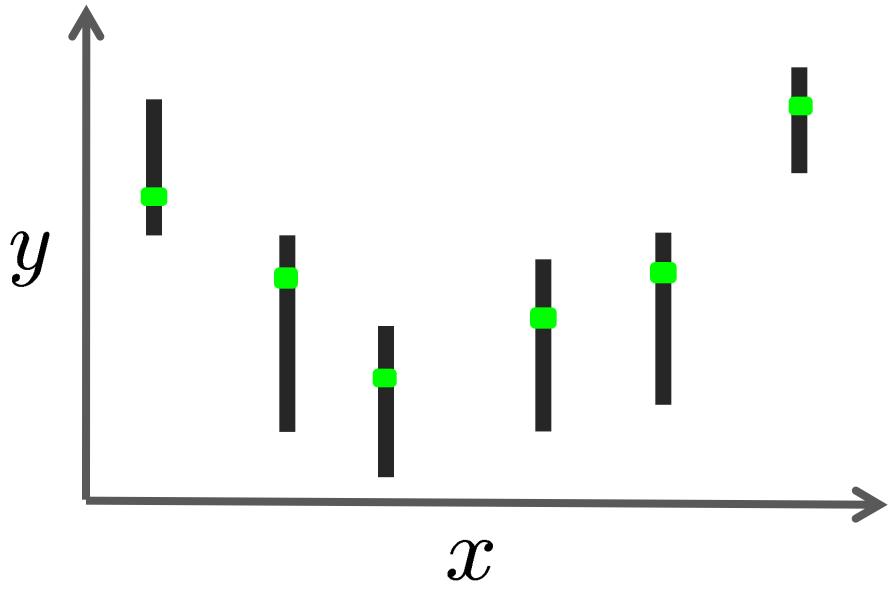
INTELLIGENT SYSTEMS

Classes



one of many instantiations

EXAMPLE: REGRESSION



one of infinitely many instantiations

DATA DISAMBIGUATION

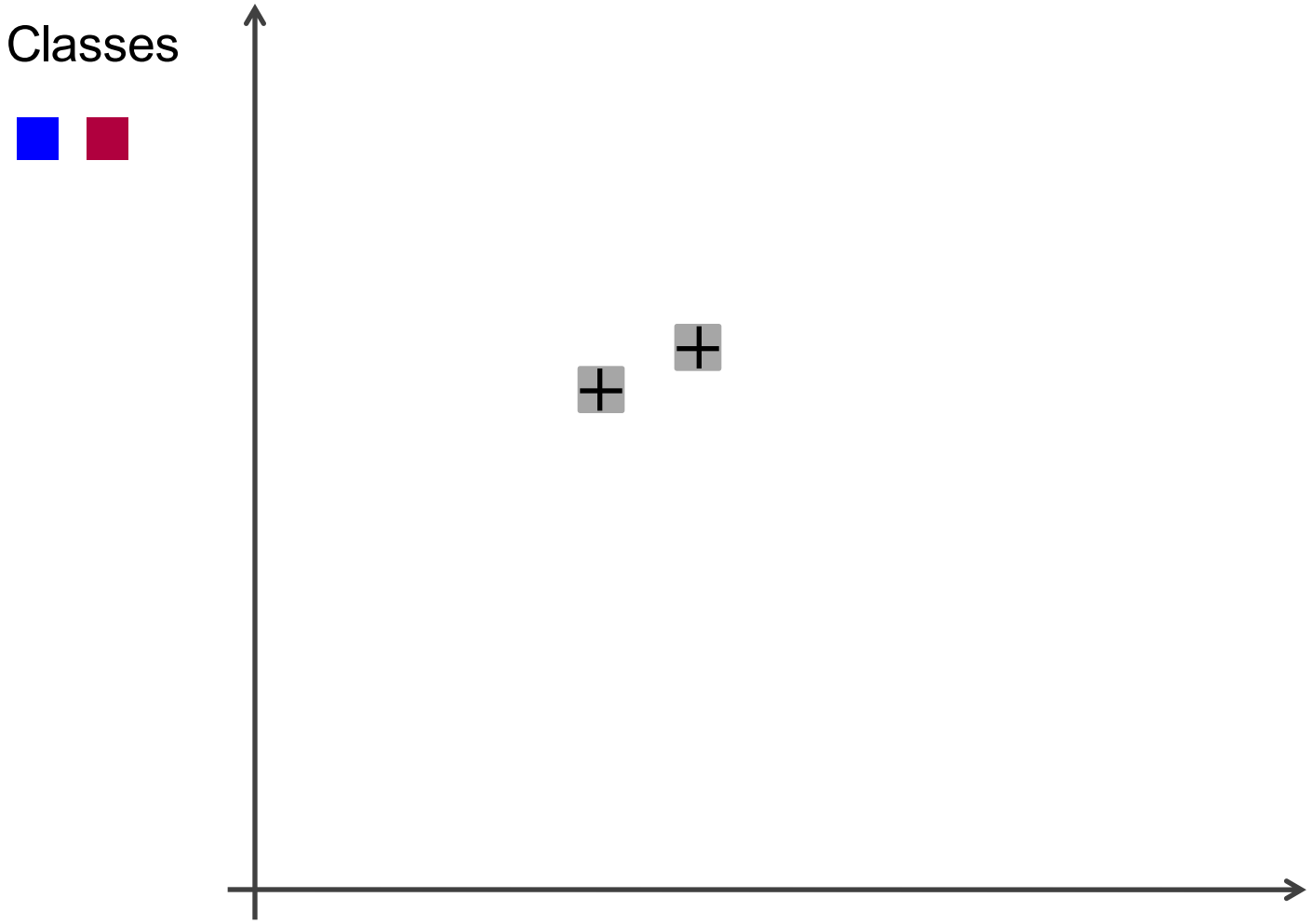
INTELLIGENT SYSTEMS

How to learn from (super)set-valued data?

We suggest that successful learning should go hand in hand with **data disambiguation**, i.e., finding out about the (precise) y_n underlying the imprecise observations Y_n ...

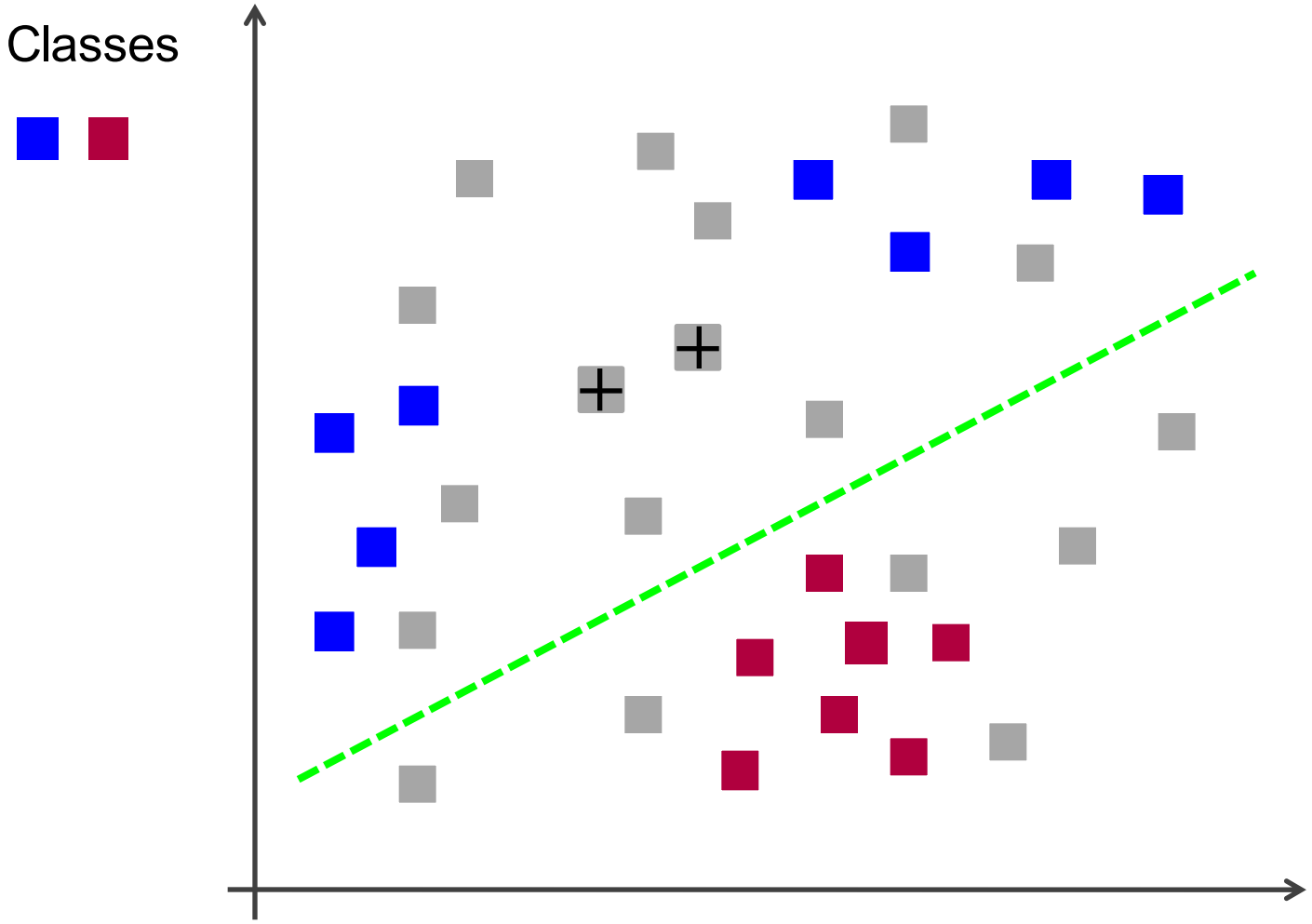
DATA DISAMBIGUATION

INTELLIGENT SYSTEMS



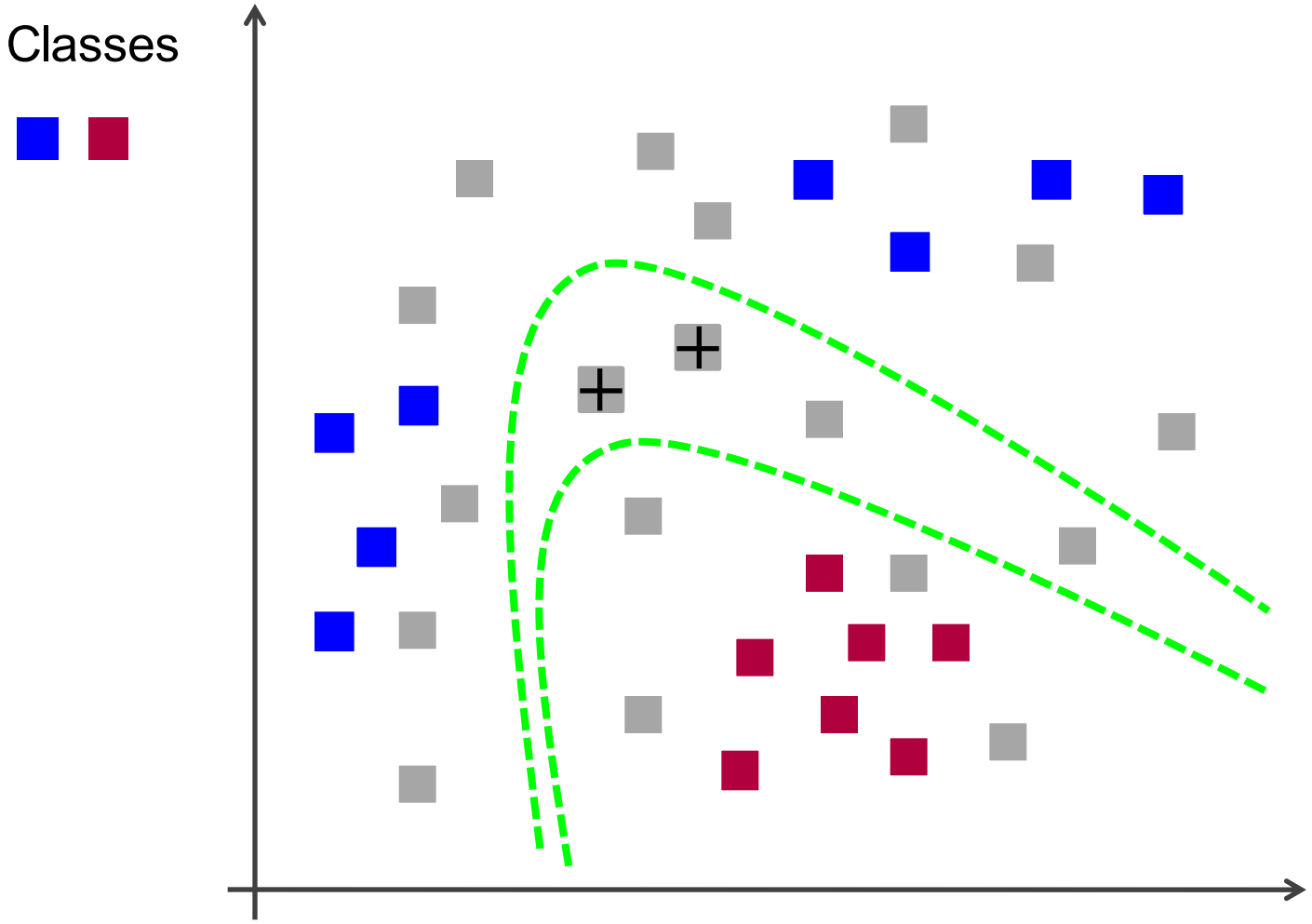
DATA DISAMBIGUATION

INTELLIGENT SYSTEMS



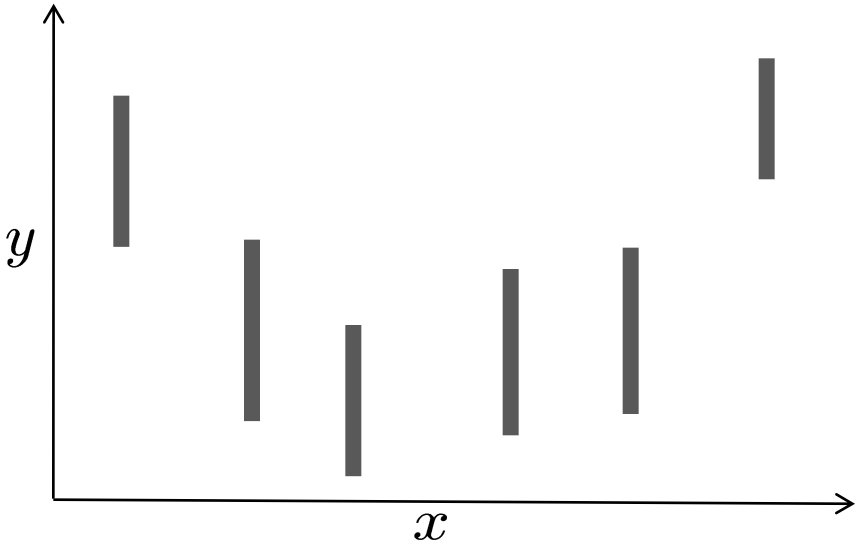
DATA DISAMBIGUATION

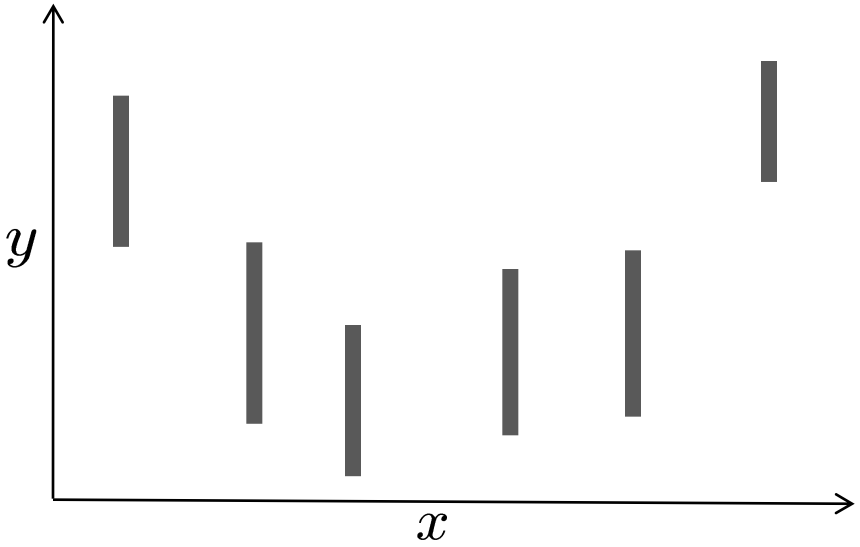
INTELLIGENT SYSTEMS



DATA DISAMBIGUATION

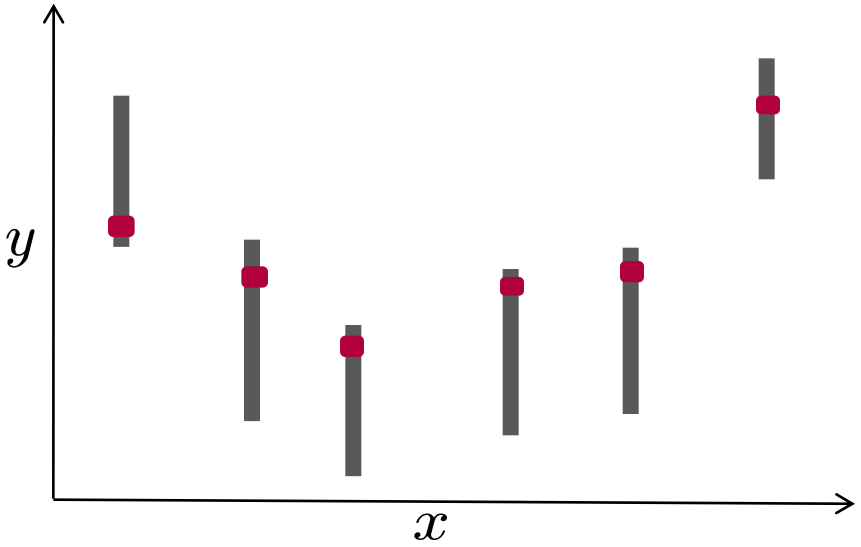
INTELLIGENT SYSTEMS

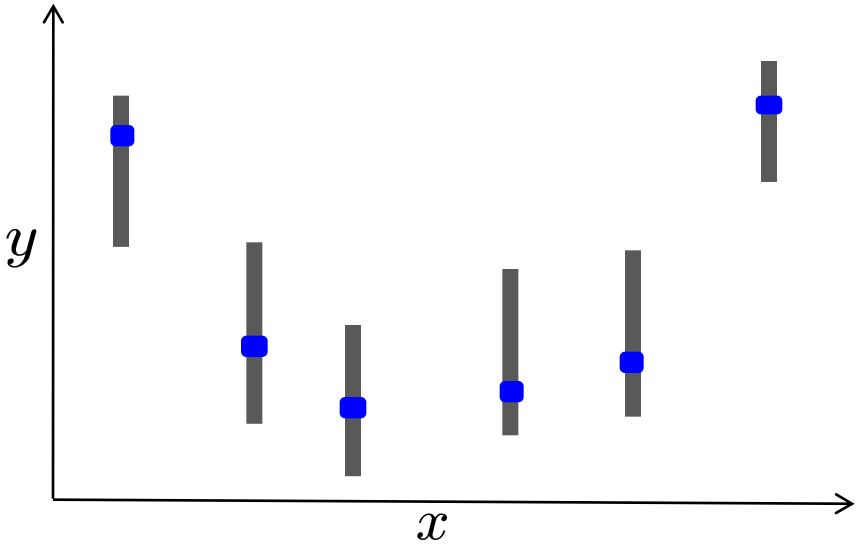




DATA DISAMBIGUATION

INTELLIGENT SYSTEMS

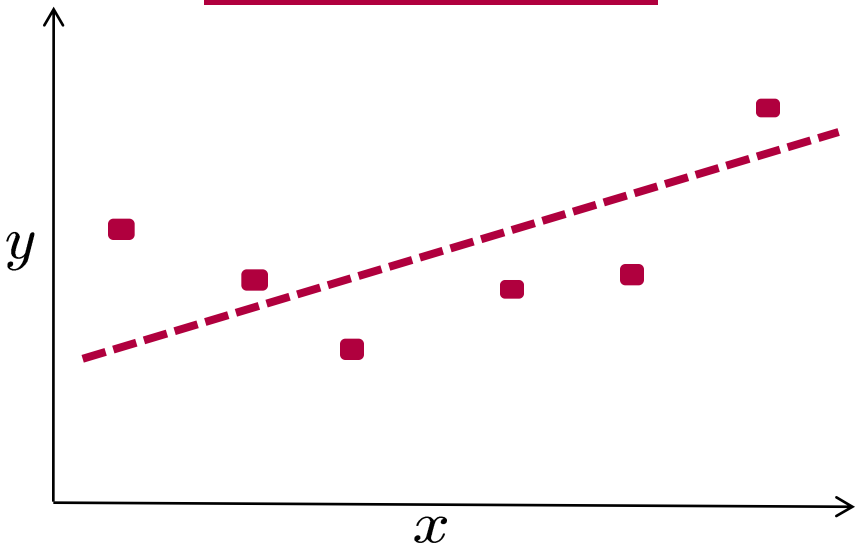




DATA DISAMBIGUATION

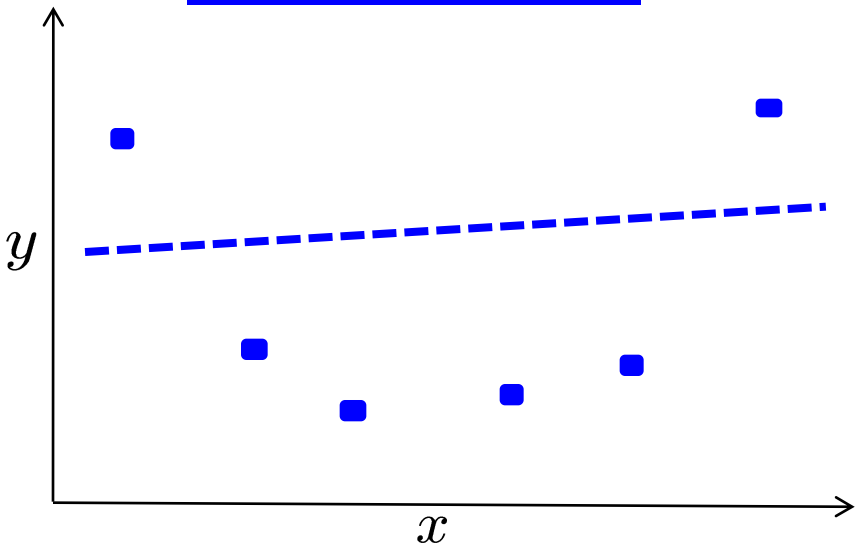
INTELLIGENT SYSTEMS

MORE PLAUSIBLE



A plausible instantiation that can be fitted reasonably well with a **LINEAR** model!

LESS PLAUSIBLE

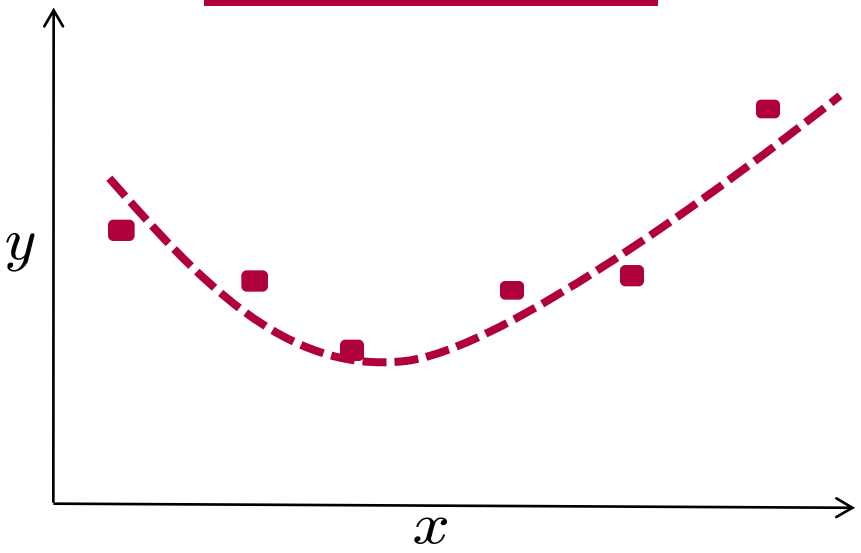


A less plausible instantiation, because there is no **LINEAR** model with a good fit!

DATA DISAMBIGUATION

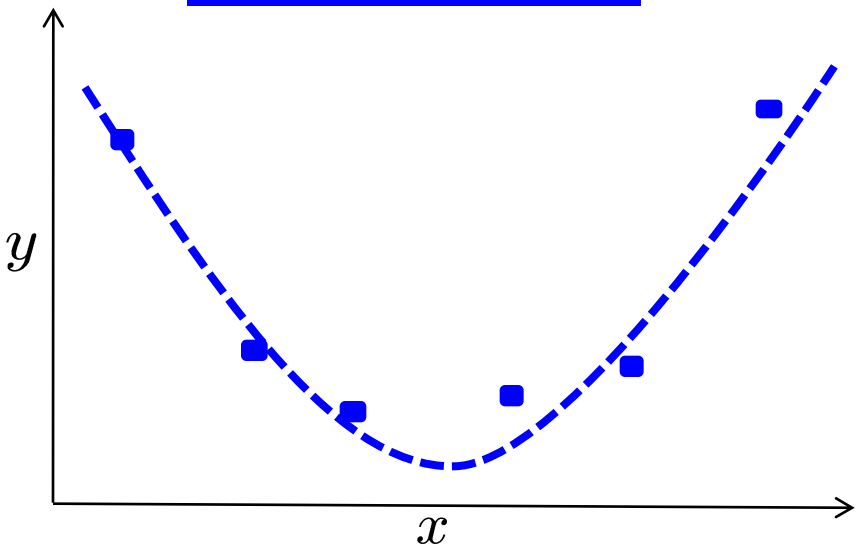
INTELLIGENT SYSTEMS

PLAUSIBLE



A plausible instantiation that can be fitted quite well with a **QUADRATIC** model!

PLAUSIBLE

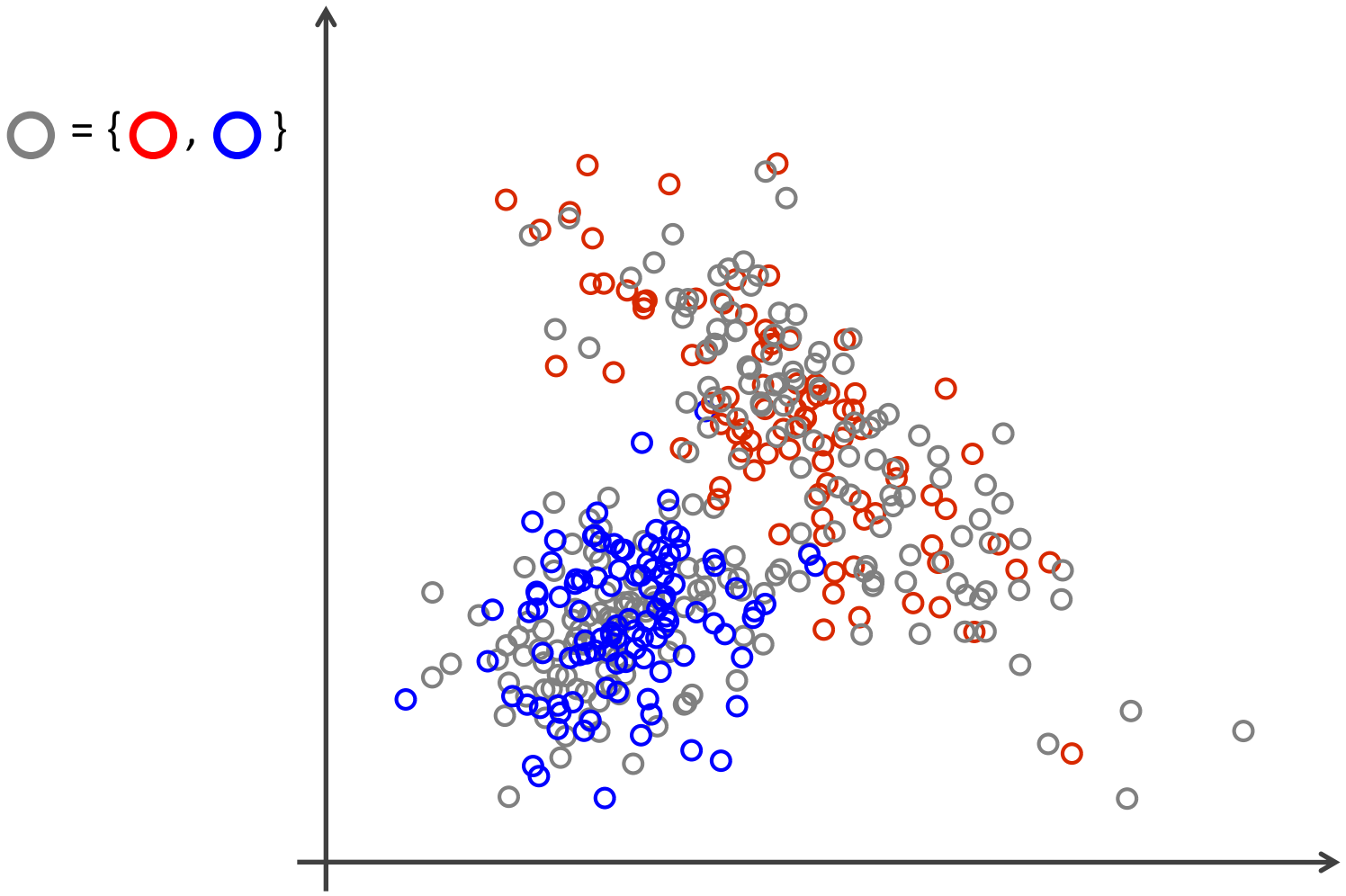


A plausible instantiation that can be fitted quite well with a **QUADRATIC** model!

It all depends on how you look at the data!

DATA DISAMBIGUATION

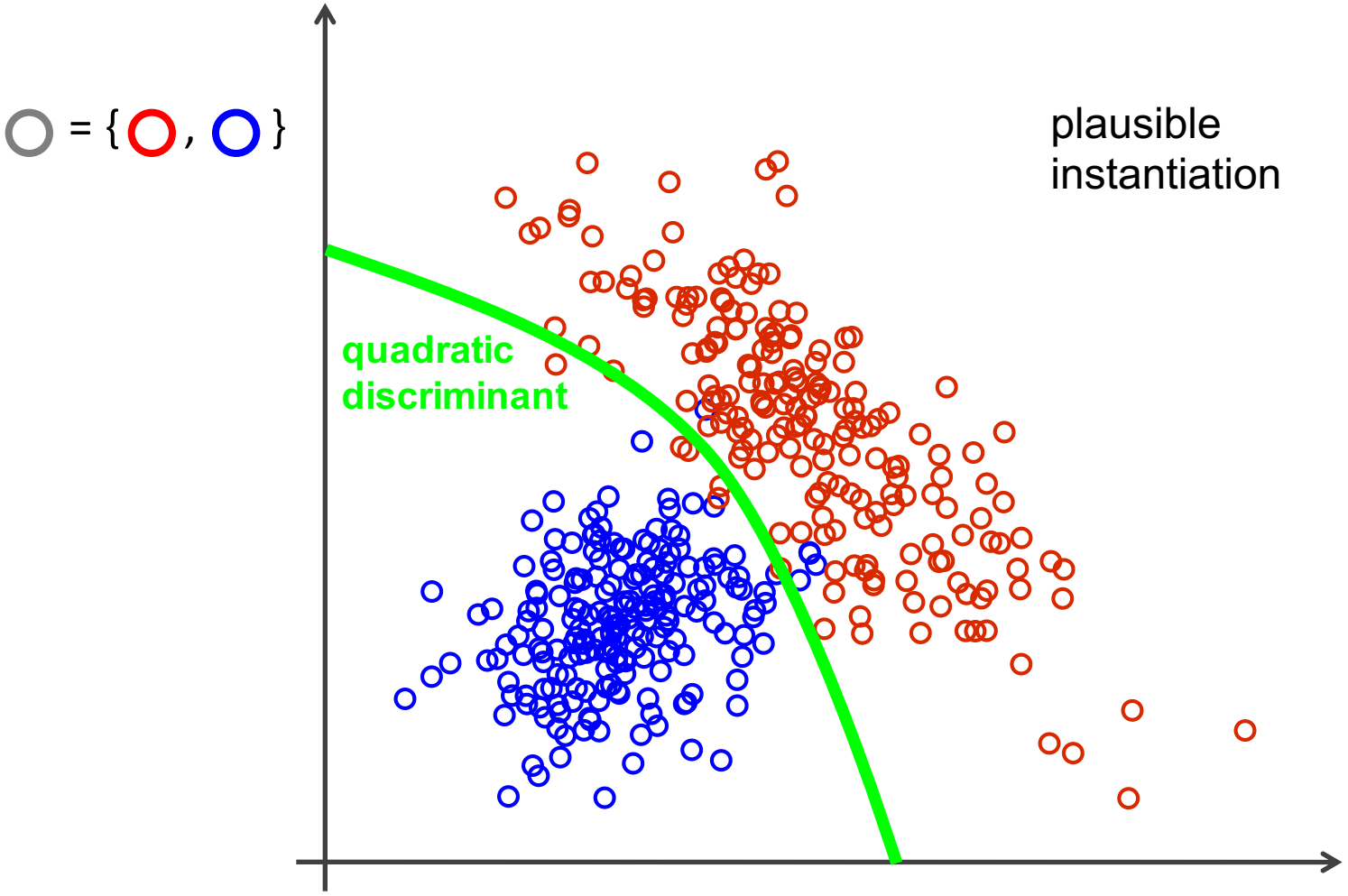
INTELLIGENT SYSTEMS



assume both class distributions to be Gaussian

DATA DISAMBIGUATION

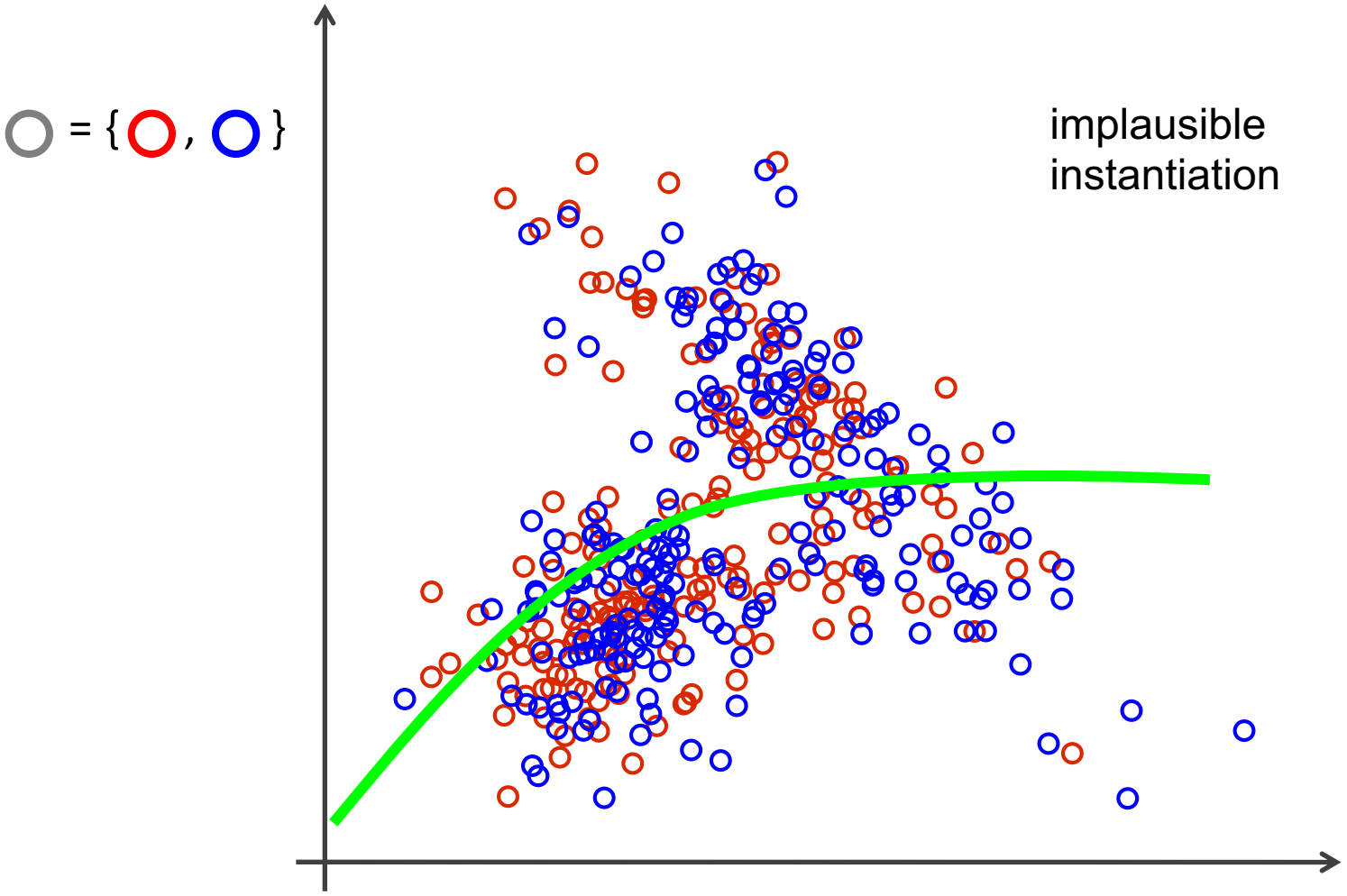
INTELLIGENT SYSTEMS



assume both class distributions to be Gaussian

DATA DISAMBIGUATION

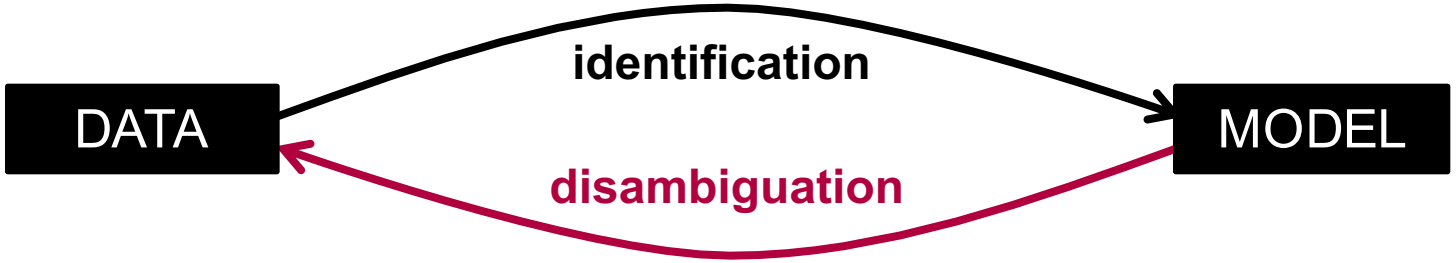
INTELLIGENT SYSTEMS



assume both class distributions to be Gaussian

INTELLIGENT SYSTEMS

Model identification and data disambiguation should be performed simultaneously:



... quite natural from a Bayesian perspective:

$$\begin{aligned}\mathbf{P}(h, \mathcal{D}) &= \mathbf{P}(h)\,\mathbf{P}(\mathcal{D} \mid h) \\ &= \mathbf{P}(\mathcal{D})\,\mathbf{P}(h \mid \mathcal{D})\end{aligned}$$

OUTLINE

INTELLIGENT SYSTEMS

PART 1	PART 2	PART 3
Superset learning	Optimistic loss minimization	Data imprecisiation

Likelihood of a model $h \in \mathcal{H}$:

$$\ell(h) = \mathbf{P}(\mathcal{O}, \mathcal{D} \,|\, h) = \mathbf{P}(\mathcal{D} \,|\, h)\, \mathbf{P}(\mathcal{O} \,|\, \mathcal{D}, h)$$
$$= \mathbf{P}(\mathcal{D} \,|\, h)\, \mathbf{P}(\mathcal{O} \,|\, \mathcal{D})$$

Imprecise observation only depends on true data, not on the model.

SUPERSET ASSUMPTION

INTELLIGENT SYSTEMS

Superset assumption:

$$\mathbf{P}(\mathcal{O} \mid \mathcal{D}) = \begin{cases} \text{const} & \text{if } \mathcal{O} \ni \mathcal{D} \\ 0 & \text{if } \mathcal{O} \not\ni \mathcal{D} \end{cases}$$

Imprecise data is a superset, but no other assumption.

GENERALIZED ERM

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

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

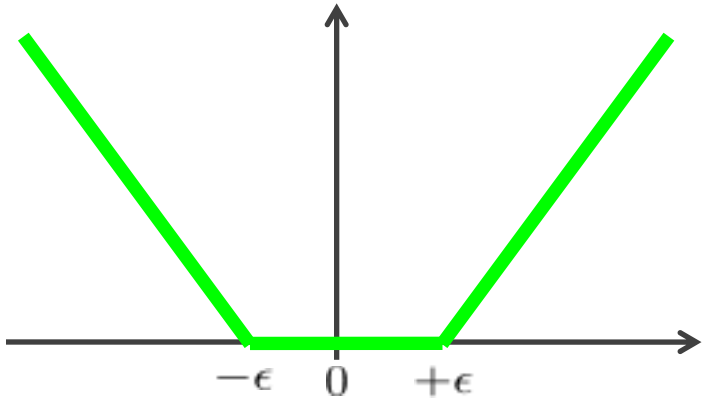
and the **optimistic superset loss** (OSL) function

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

how well the (precise) model fits the imprecise data

SPECIAL CASES

INTELLIGENT SYSTEMS

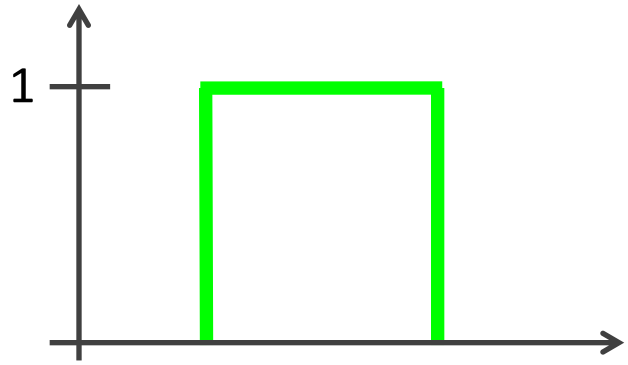


The ϵ-insensitive loss $L(y, \hat{y}) = \max(|y - \hat{y}| - \epsilon, 0)$ used in support vector regression corresponds to L^* with L the standard L_1 loss $L(y, \hat{y}) = |y - \hat{y}|$ and precise data y_n being replaced by interval-valued data $Y_n = [y_n - \epsilon, y_n + \epsilon]$.

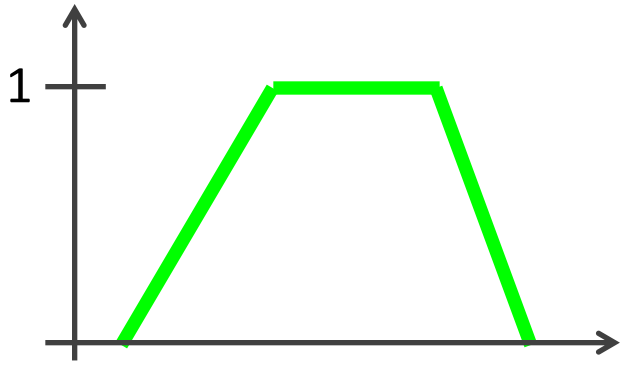
GENERALIZATION TO FUZZY DATA

INTELLIGENT SYSTEMS

- Y_n is a **subset** of $\mathcal{Y}$ (with characteristic function $\mathcal{Y} \longrightarrow \{0, 1\}$)
- Y_n is a **fuzzy subset** of $\mathcal{Y}$, characterized in terms of a membership function $\mathcal{Y} \longrightarrow [0, 1]$



interval



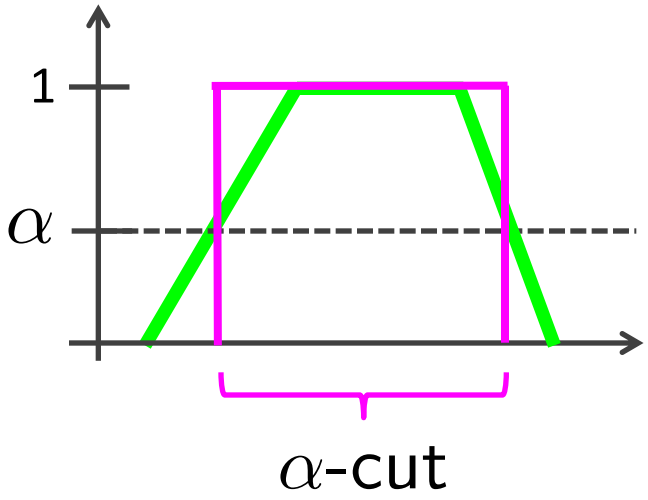
fuzzy interval

GENERALIZATION TO FUZZY DATA

INTELLIGENT SYSTEMS

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

LOSS



GENERALIZATION TO FUZZY DATA

INTELLIGENT SYSTEMS

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

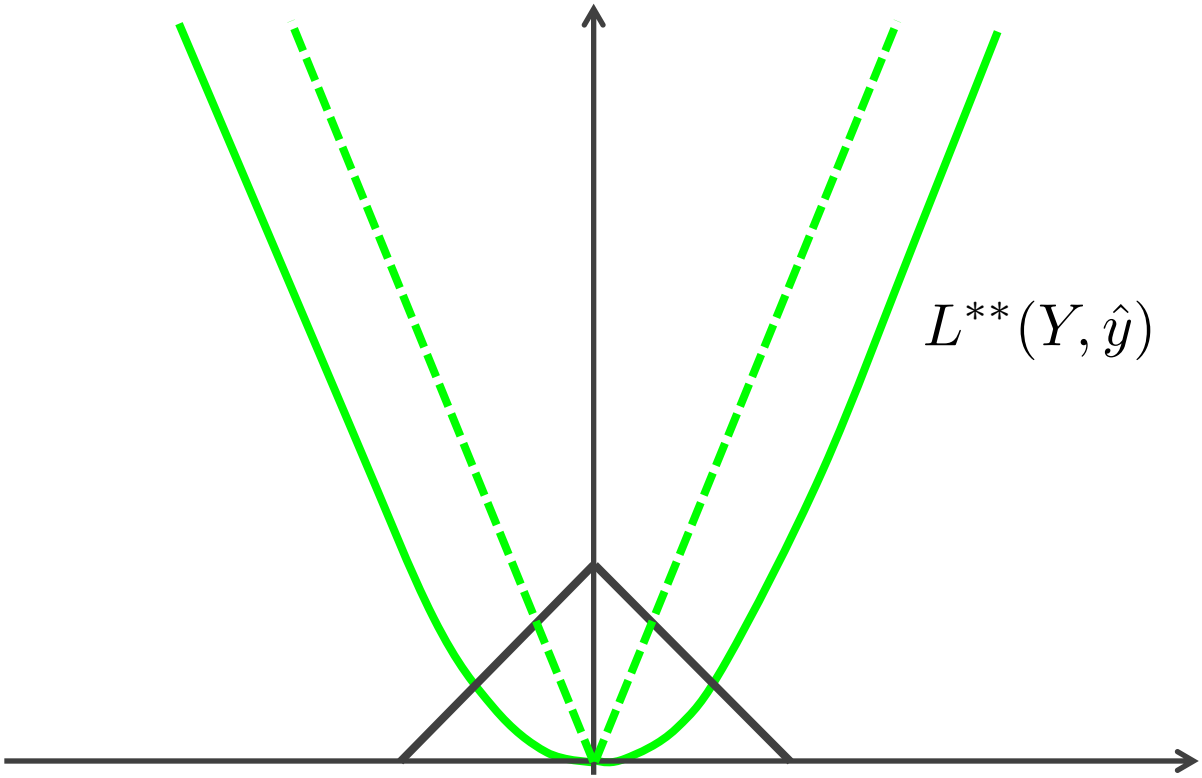
LOSS

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

RISK

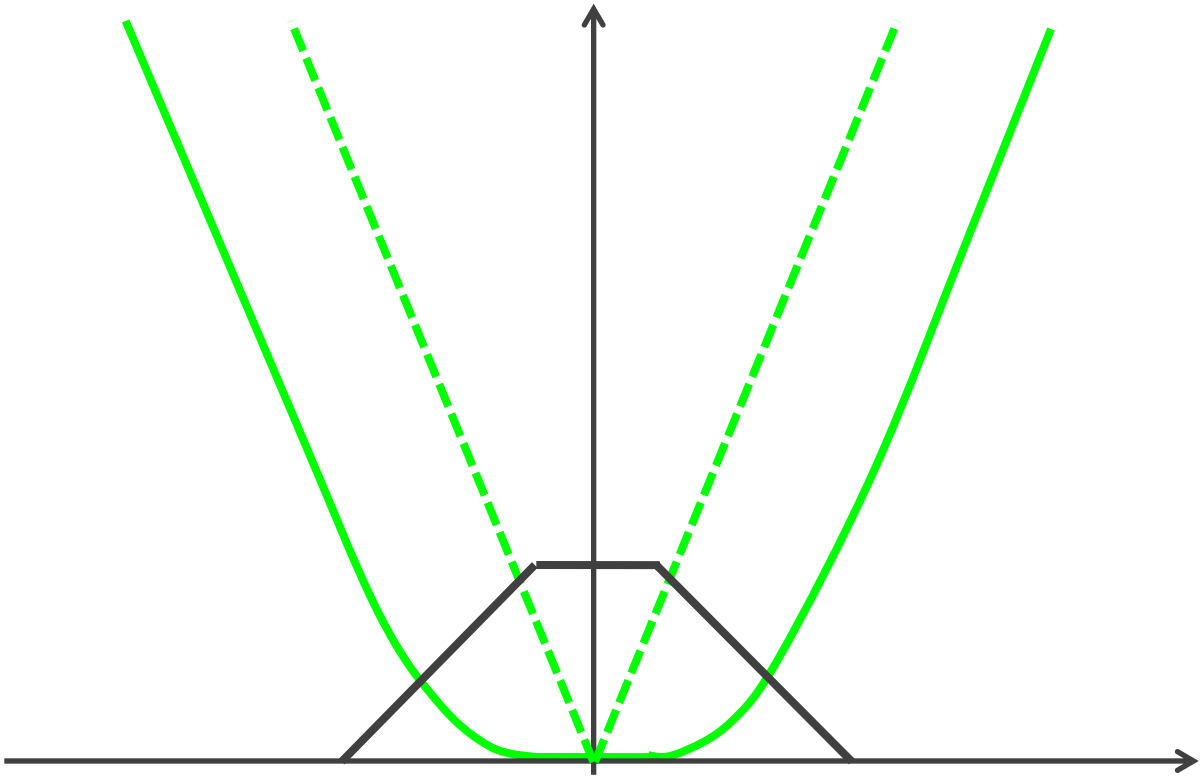
GENERALIZATION TO FUZZY DATA

INTELLIGENT SYSTEMS



→ *Huber loss* !

GENERALIZATION TO FUZZY DATA



→ *(generalized) Huber loss !*

STRUCTURED OUTPUT PREDICTION

INTELLIGENT SYSTEMS

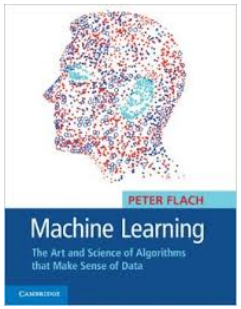
Superset learning naturally applies to learning problems with ***structured outputs****, which are often only* ***partially specified*** *and can then be associated with the* ***set of all consistent completions****.*

LABEL RANKING

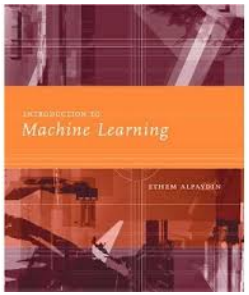
INTELLIGENT SYSTEMS

... is the problem to learn a model that maps instances to **TOTAL ORDERS** over a fixed set of alternatives/labels:

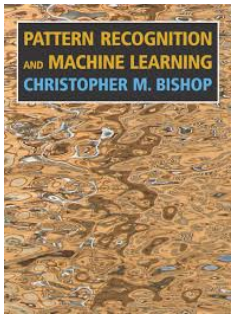
A



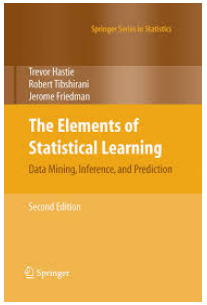
B



C



D



$$\mathcal{Y} = \{\text{ABCD}, \text{ABDC}, \ldots, \text{DCBA}\}$$

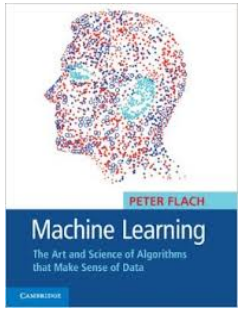
LABEL RANKING

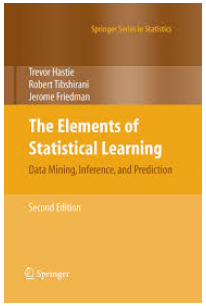
INTELLIGENT SYSTEMS

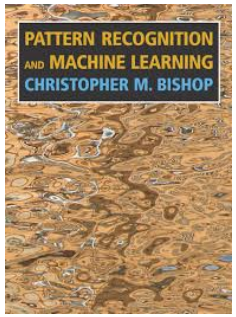
... is the problem to learn a model that maps instances to **TOTAL ORDERS** over a fixed set of alternatives/labels:

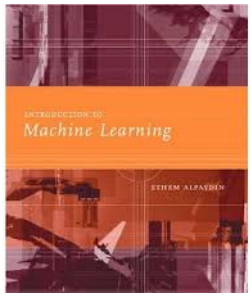
(0,37,46,325,1,0)

A ≻ D ≻ C ≻ B









... likes more
... reads more
... recommends more
...

LABEL RANKING

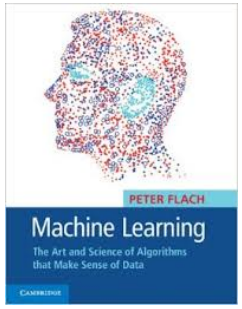
INTELLIGENT SYSTEMS

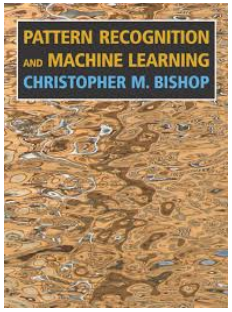
... is the problem to learn a model that maps instances to **TOTAL ORDERS** over a fixed set of alternatives/labels:

(0,37,46,325,1,0)

$\mapsto$

A $\succ$ C





Training data is typically incomplete!

LABEL RANKING

INTELLIGENT SYSTEMS

... is the problem to learn a model that maps instances to **TOTAL ORDERS** over a fixed set of alternatives/labels:

(0,37,46,325,1,0)

:

$$\left\{ \begin{array}{l} A \succ C \succ B \succ D \\ A \succ C \succ D \succ B \\ A \succ B \succ C \succ D \\ \vdots \succ \vdots \succ \vdots \succ \vdots \\ D \succ B \succ A \succ C \end{array} \right\}$$

set of linear extensions

Training data is typically incomplete!

LABEL RANKING LOSSES

INTELLIGENT SYSTEMS

KENDALL

$$L(\pi, \pi^*) = \sum_{1 \le i < j \le M} [\![(\pi(i) - \pi(j))(\pi^*(i) - \pi^*(j)) < 0]\!]$$

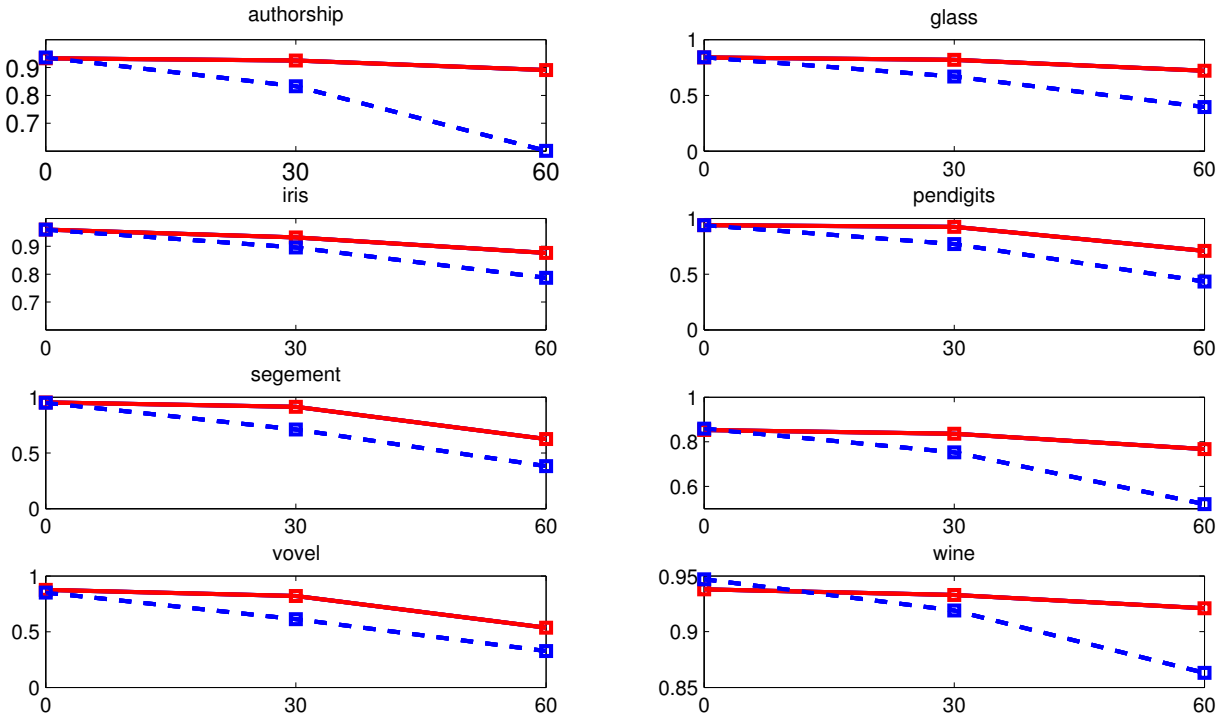
SPEARMAN

$$L(\pi, \pi^*) = \sum_{1 \le i \le M} |\pi^*(i) - \pi(i)|$$

EXPERIMENTAL STUDIES

INTELLIGENT SYSTEMS

- Cheng and H. (2015) compare an approach to label ranking based on superset learning with a state-of-the-art label ranker based on the Plackett-Luce model (PL).
- Two missing label scenarios: missing at random, top-rank
- General conclusion: more robust toward incompleteness



OUTLINE

INTELLIGENT SYSTEMS

PART 1	PART 2	PART 3
Superset learning	Optimistic loss minimization	Data imprecisiation

So far:

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

Now:

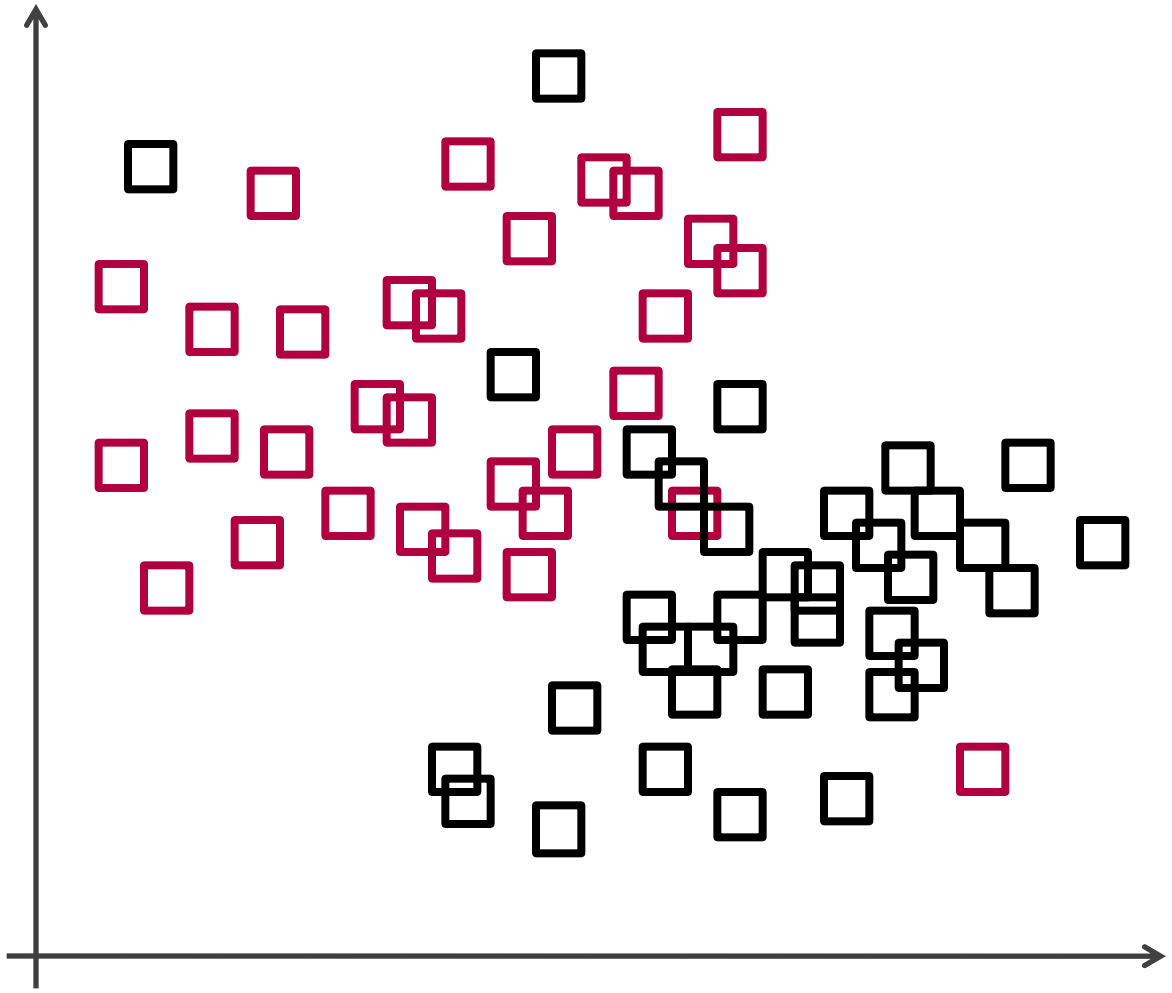
Deliberately turn precise into imprecise data, so as to modulate the influence of an observation on the learning process!

Motivated by the following observation:

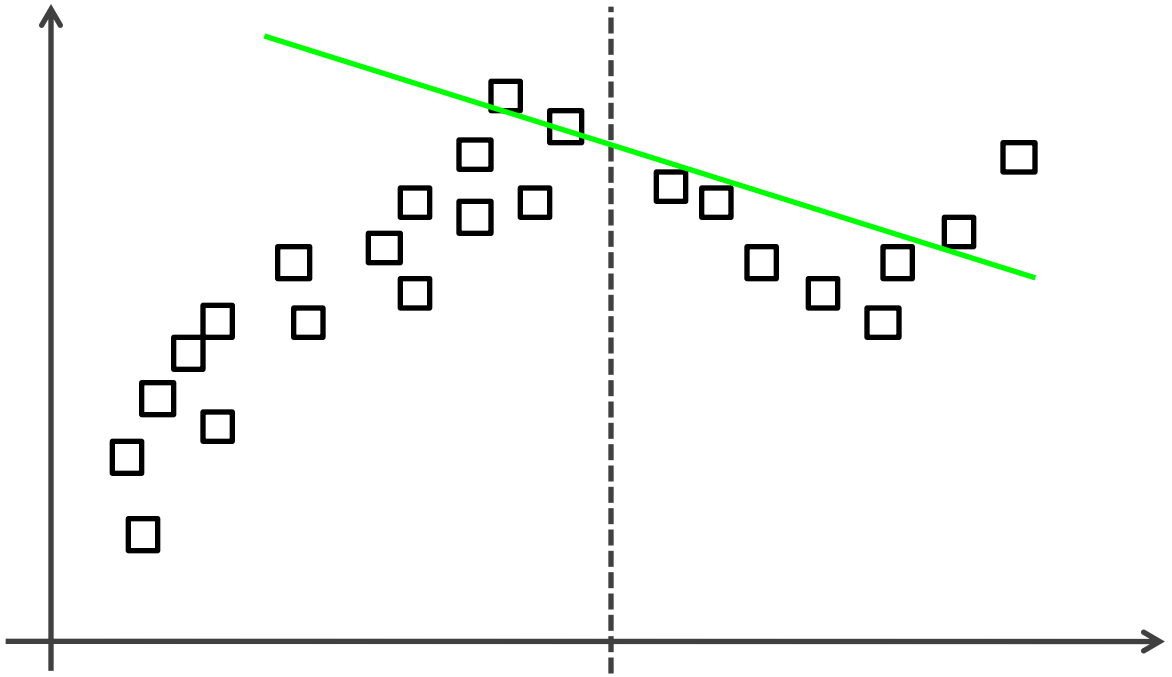
$$(Y \subset Y') \quad \Rightarrow \quad \big(L^*(Y,\cdot) \geq L^*(Y',\cdot)\big)$$

EXAMPLE WEIGHING

INTELLIGENT SYSTEMS



EXAMPLE WEIGHING

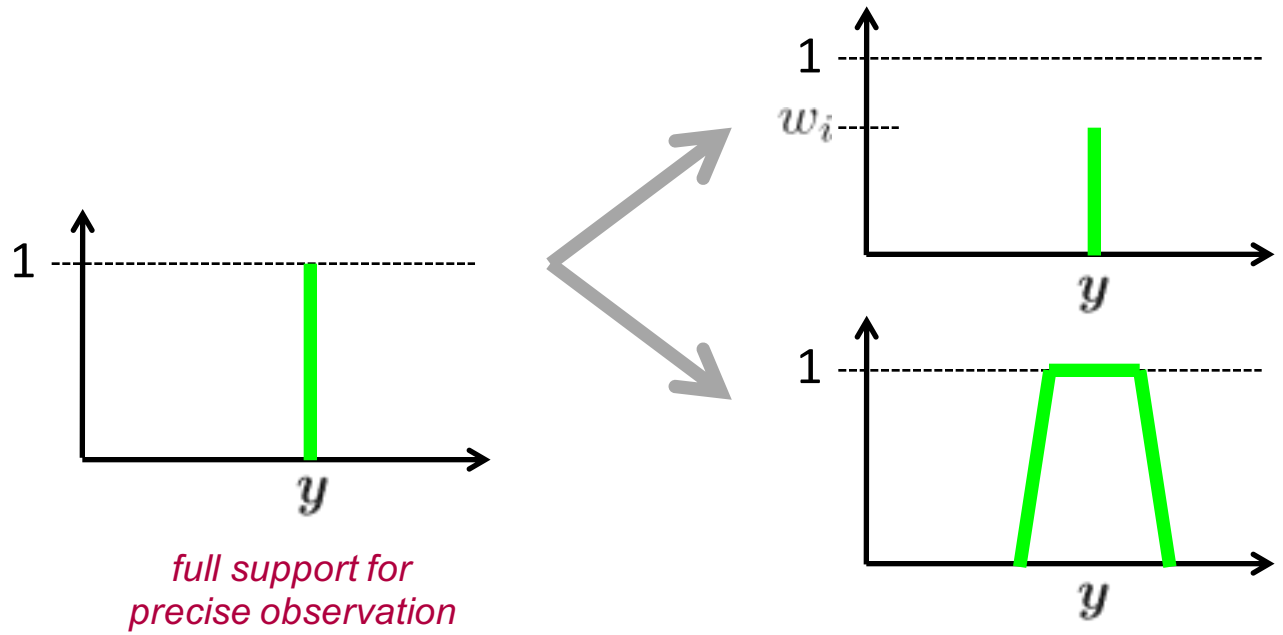


$$\text{minimize} \sum_{i=1}^{n} w_i \cdot (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2$$

EXAMPLE WEIGHING

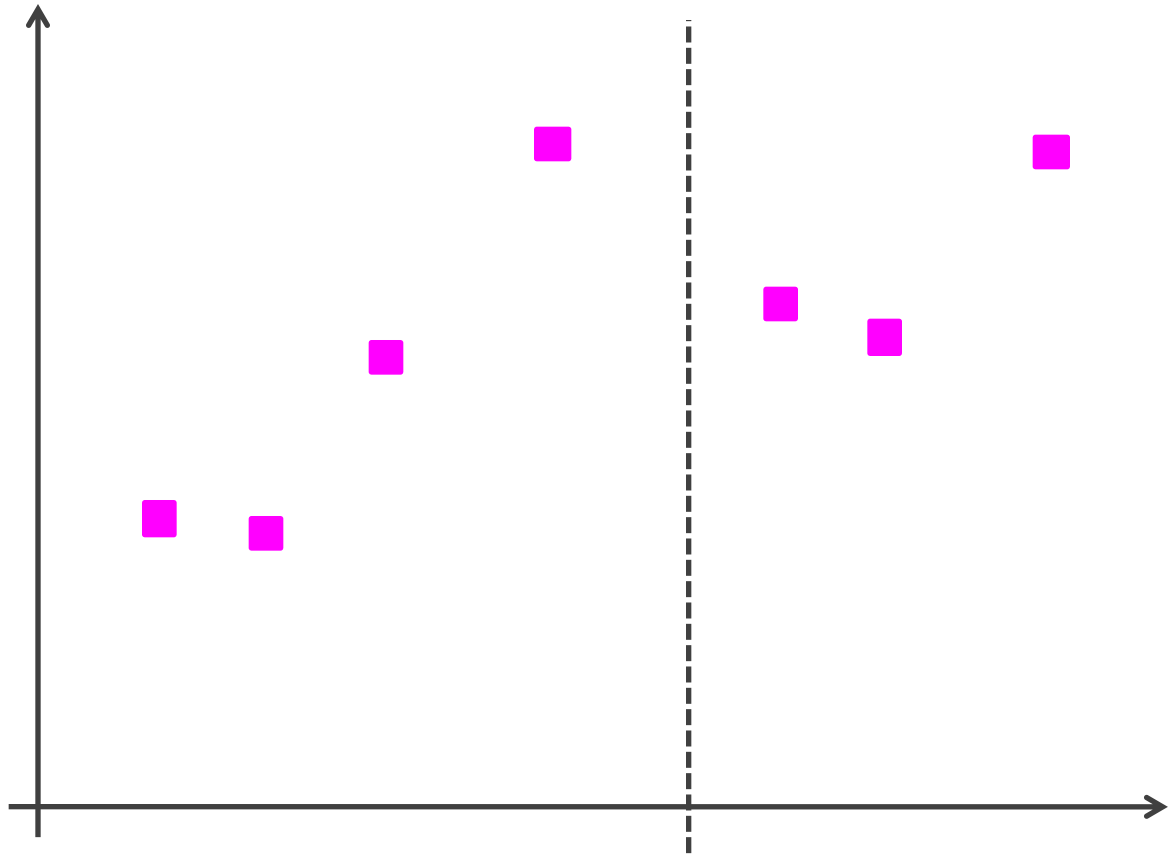
INTELLIGENT SYSTEMS

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

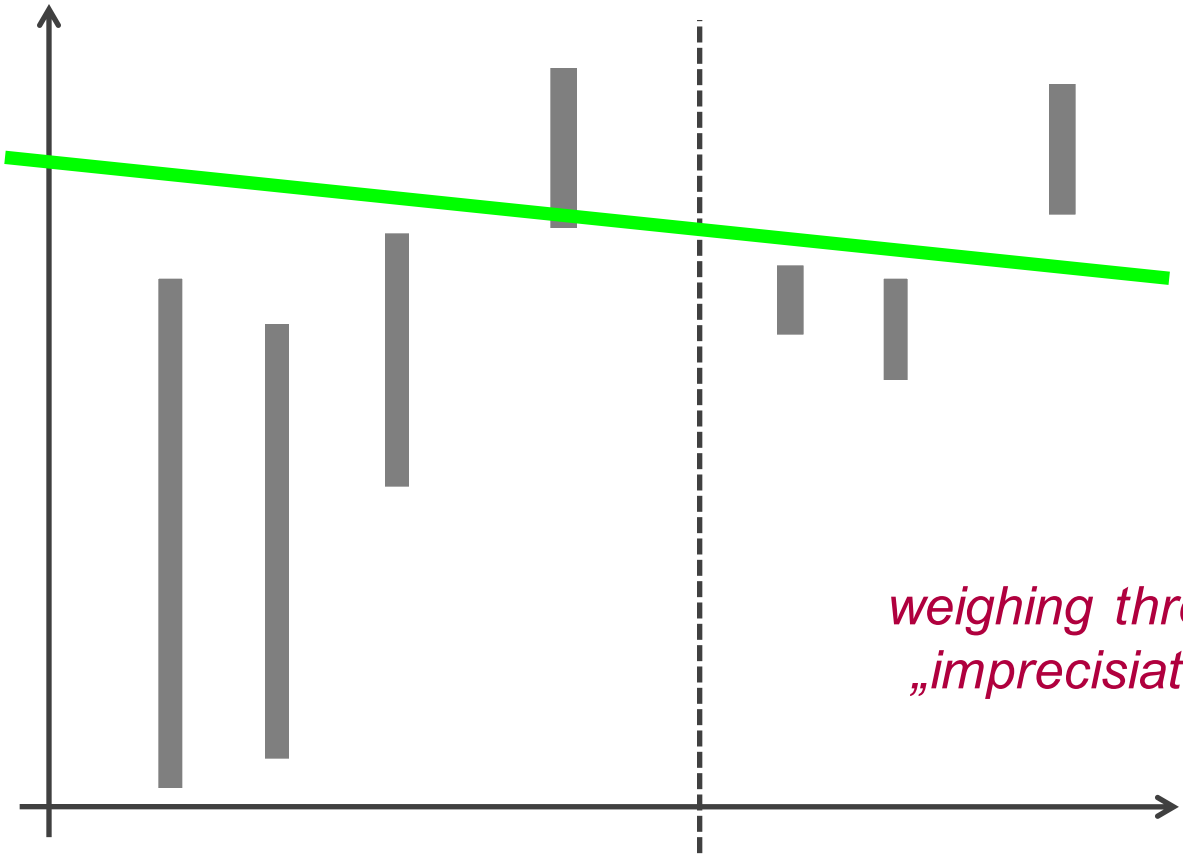


EXAMPLE WEIGHING

INTELLIGENT SYSTEMS



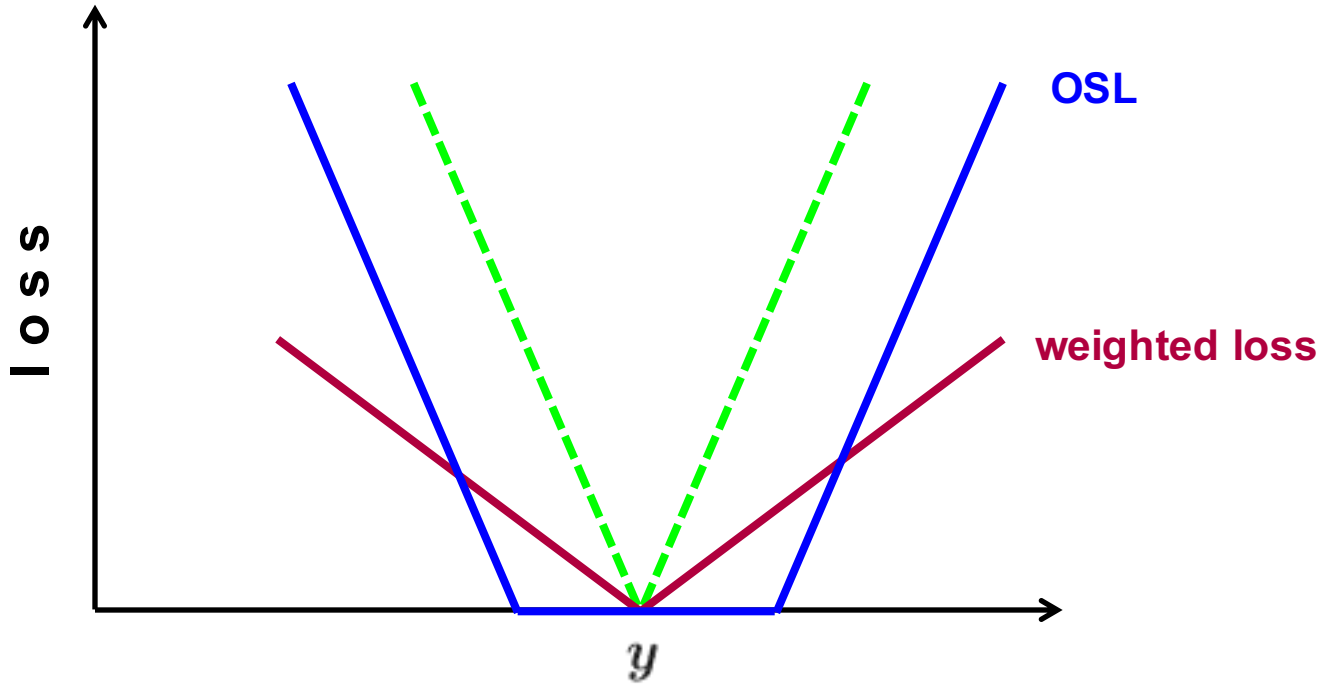
EXAMPLE WEIGHING



weighing through „imprecisiation“

$$\text{minimize} \sum_{n=1}^{N} L_2^* \left(Y_n, \boldsymbol{w}^\top \boldsymbol{x}_n \right)$$

EXAMPLE WEIGHING

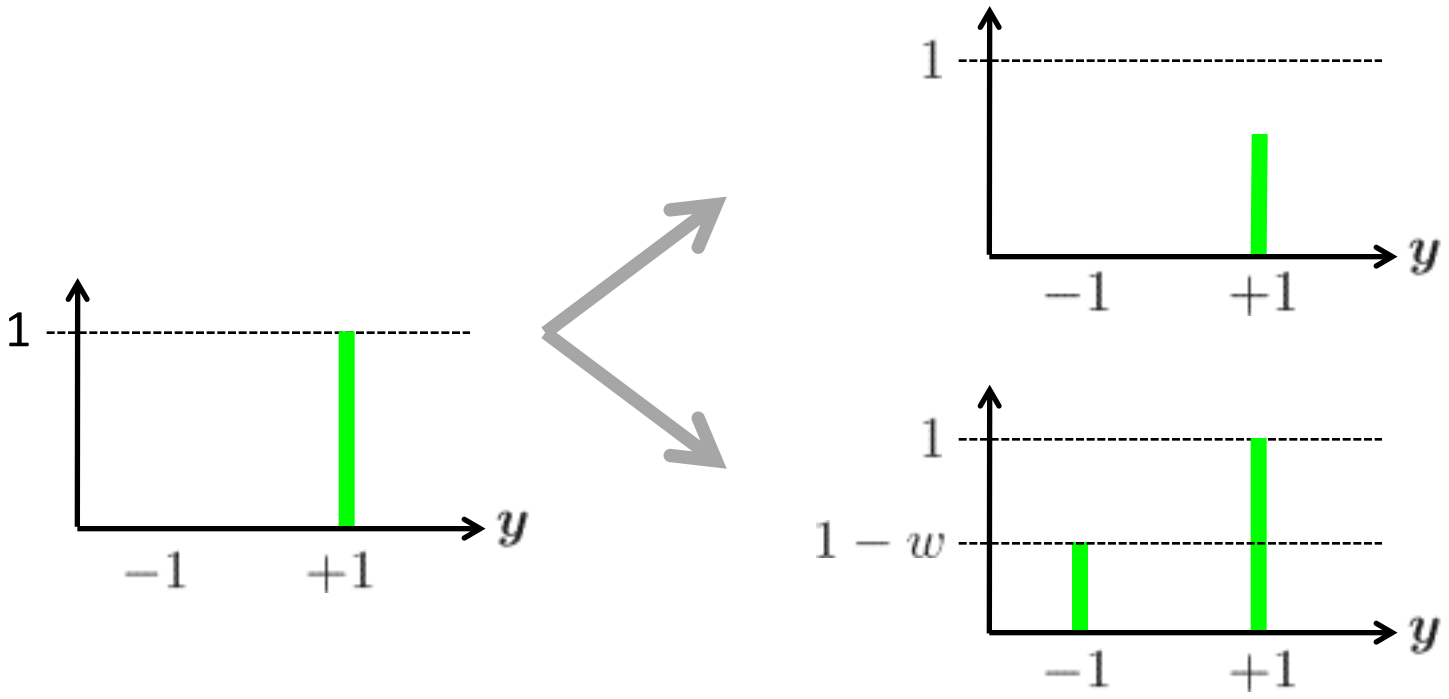


Different ways of (individually) discounting the loss function.

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

INTELLIGENT SYSTEMS

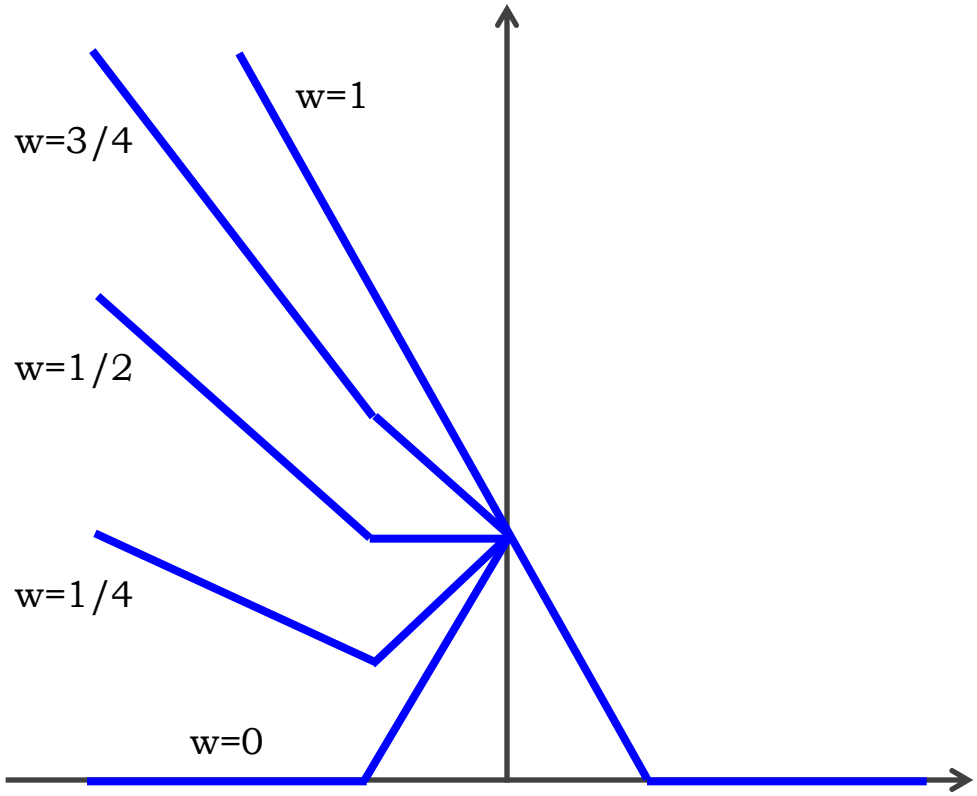
We suggest an alternative way of weighing examples, namely, via **„data impreciſiation“** ...



certainly positive

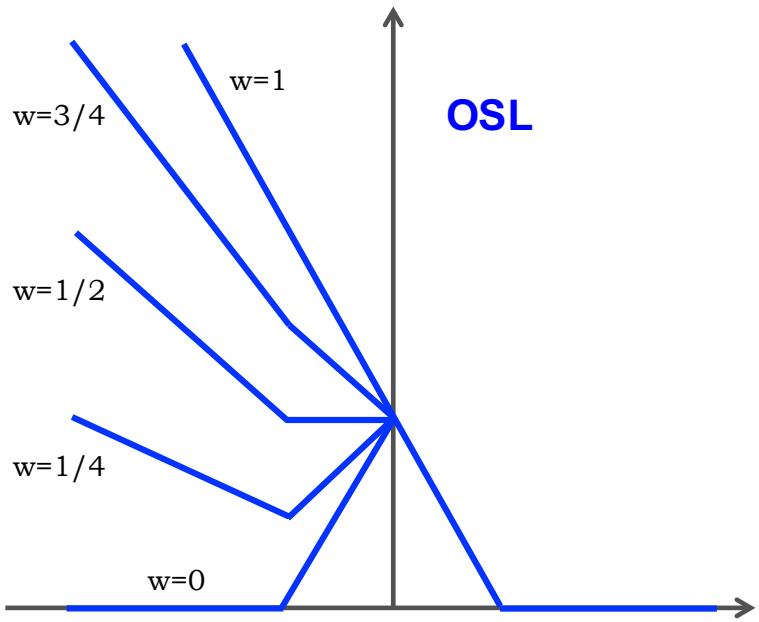
less certainly positive

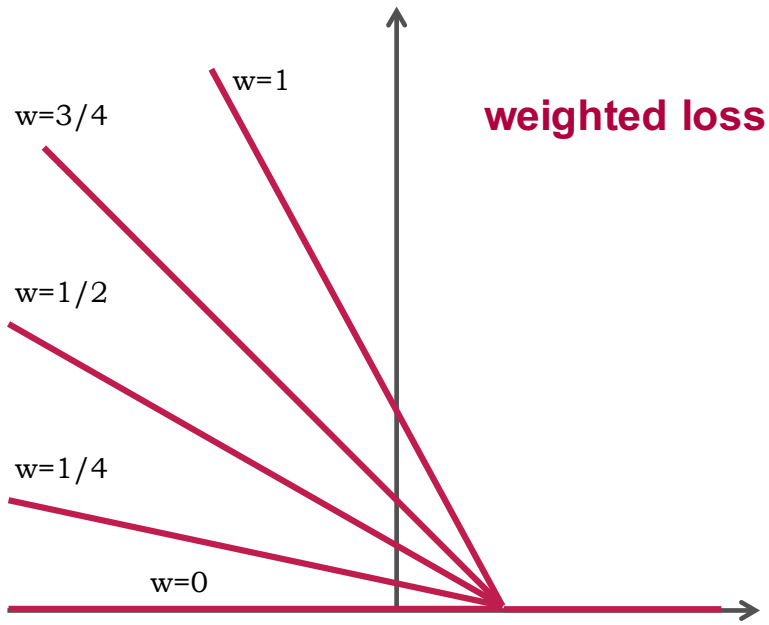
FUZZY MARGIN LOSSES



GENERALIZED HINGE LOSS

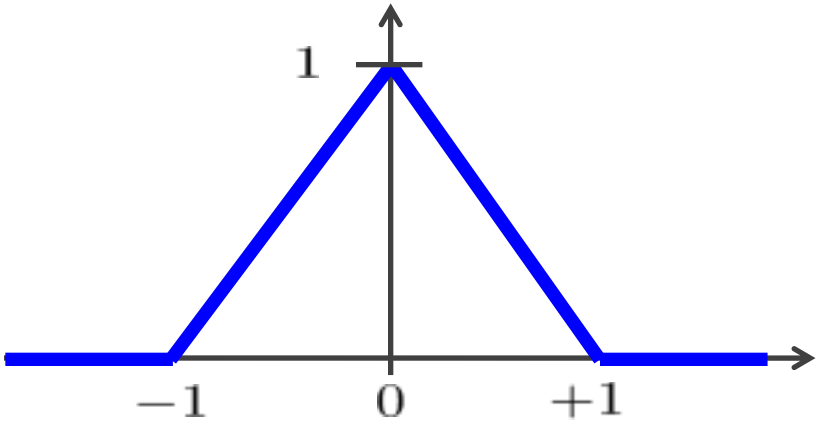
FUZZY MARGIN LOSSES





Different ways of (individually) discounting the loss function.

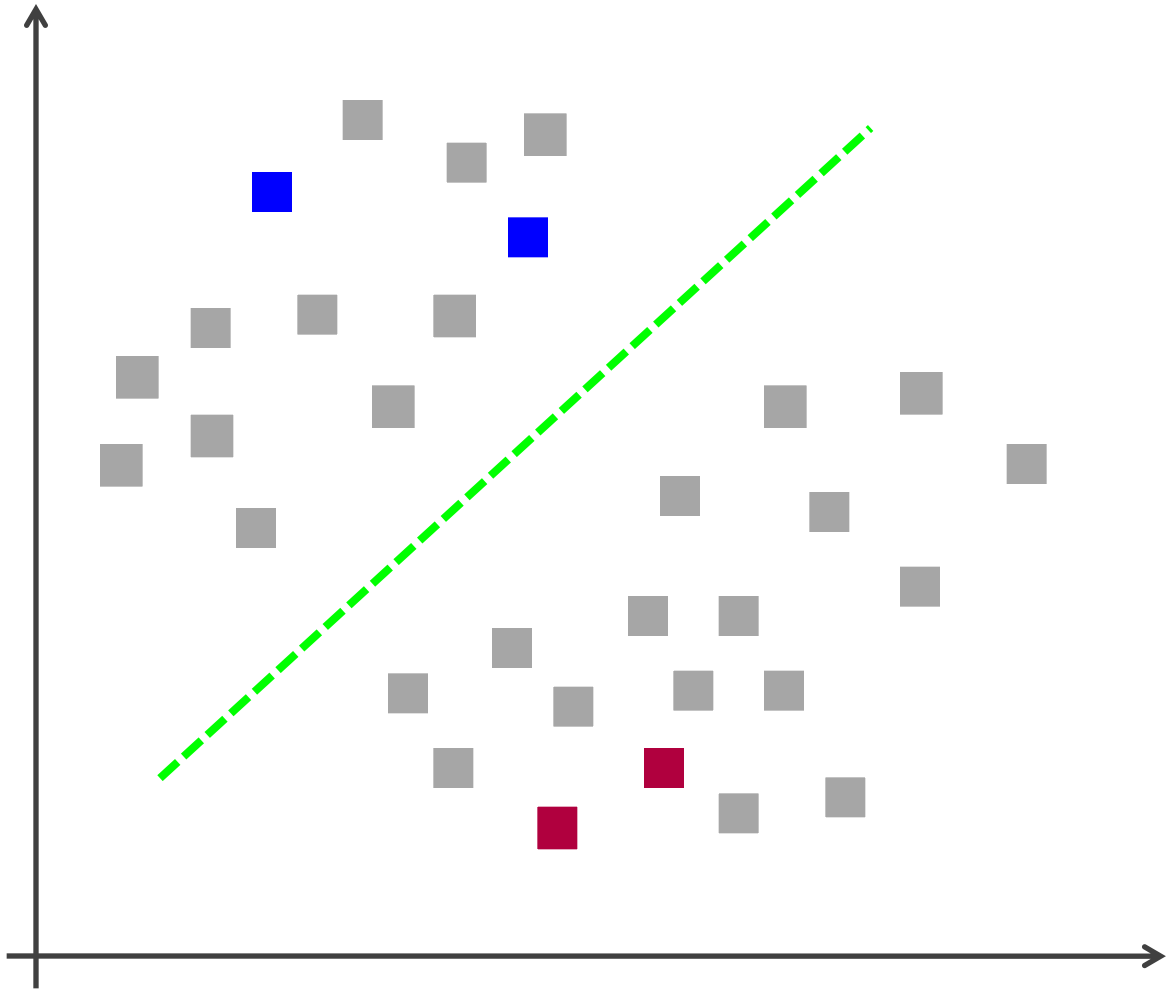
THE HAT LOSS



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".

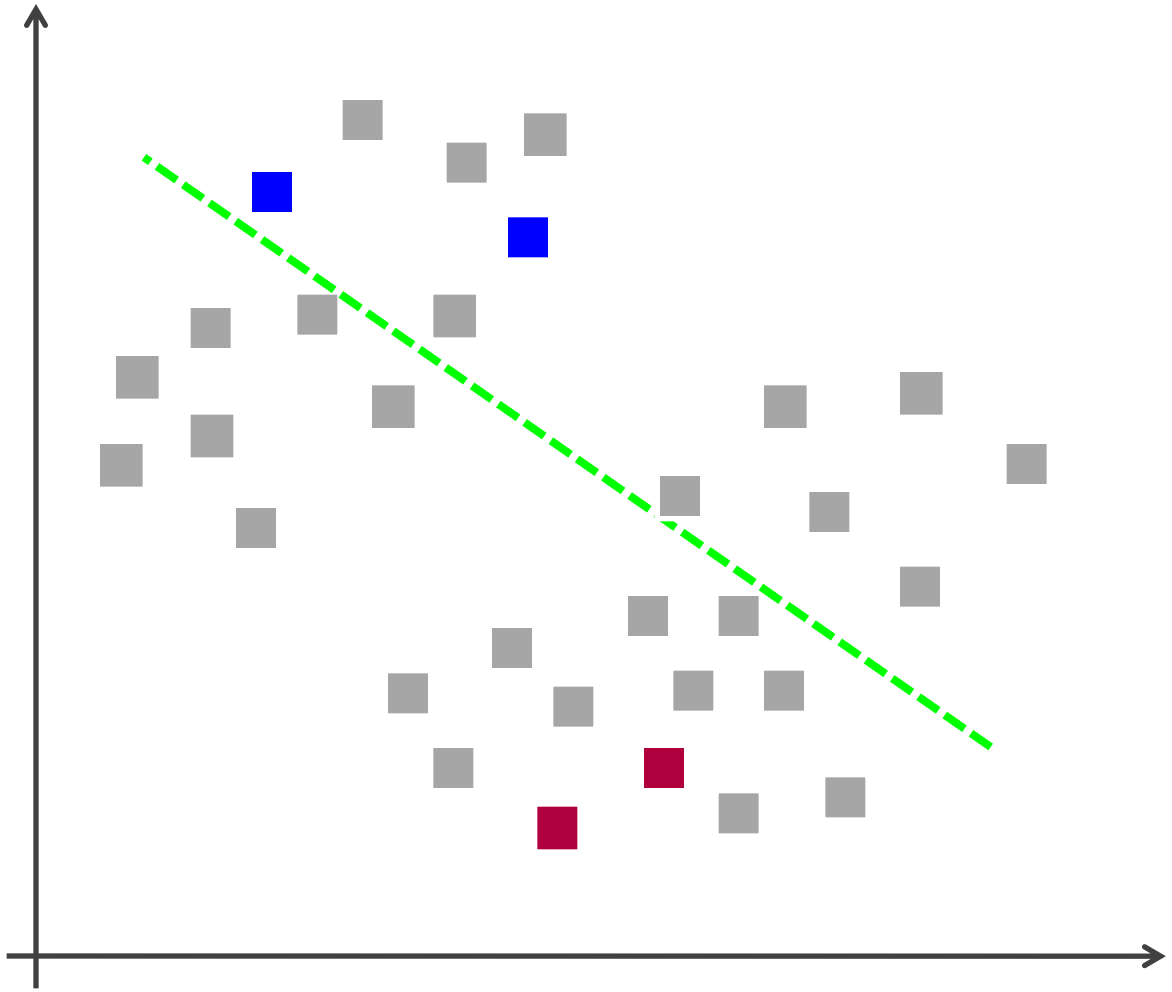
DATA DISAMBIGUATION

INTELLIGENT SYSTEMS



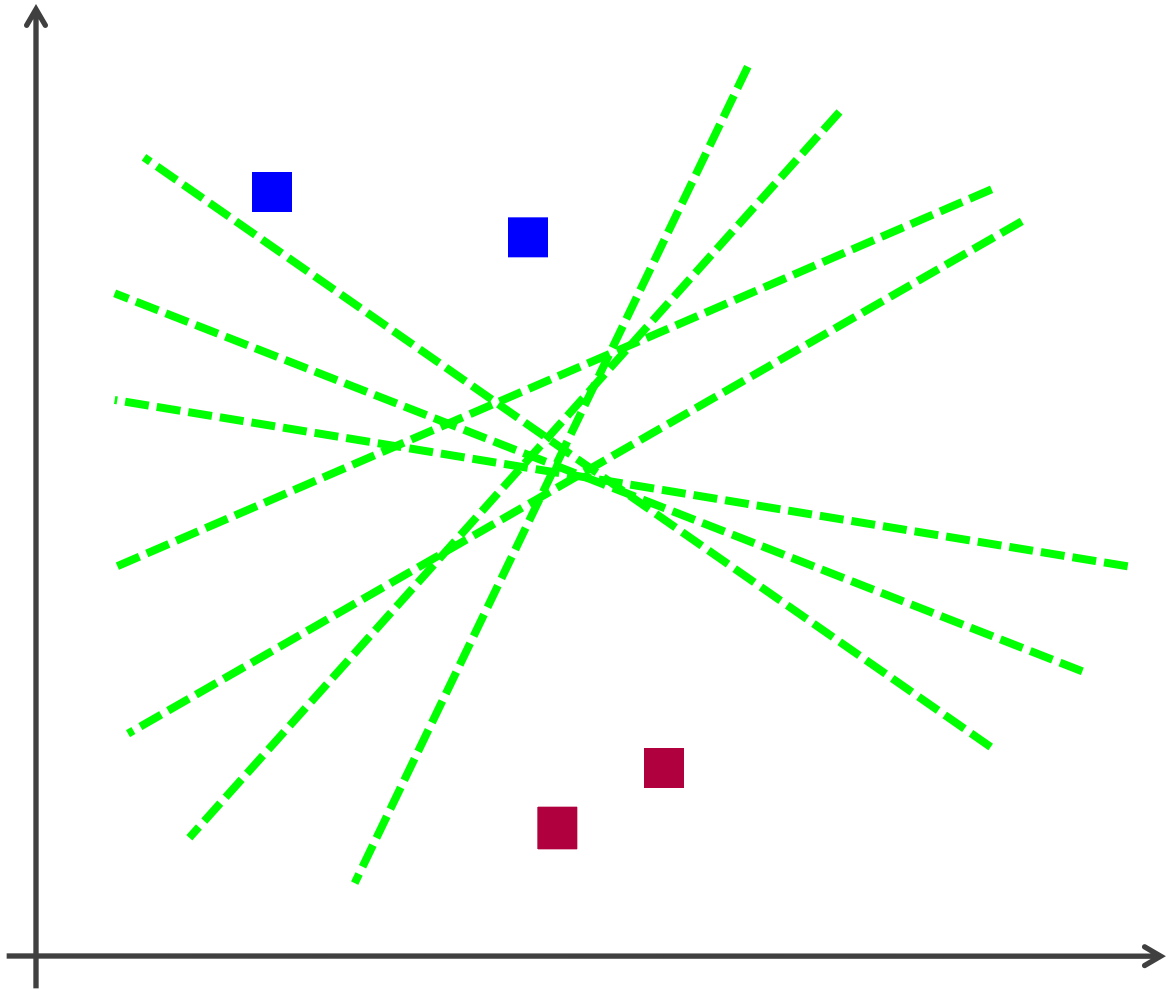
DATA DISAMBIGUATION

INTELLIGENT SYSTEMS



DATA DISAMBIGUATION

INTELLIGENT SYSTEMS



Robust loss minimization techniques:

- **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 optimistic fuzzy loss.

  Non-convex optimization problem solved by concave-convex procedure (Yuille and Rangaraja, 2002).

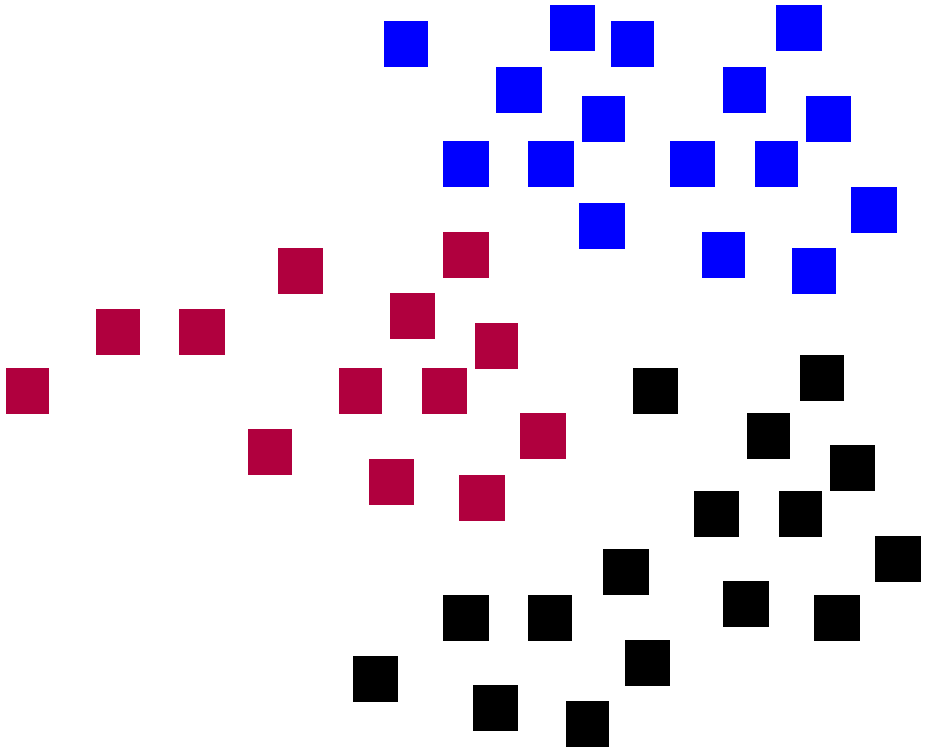
EXPERIMENTAL RESULTS

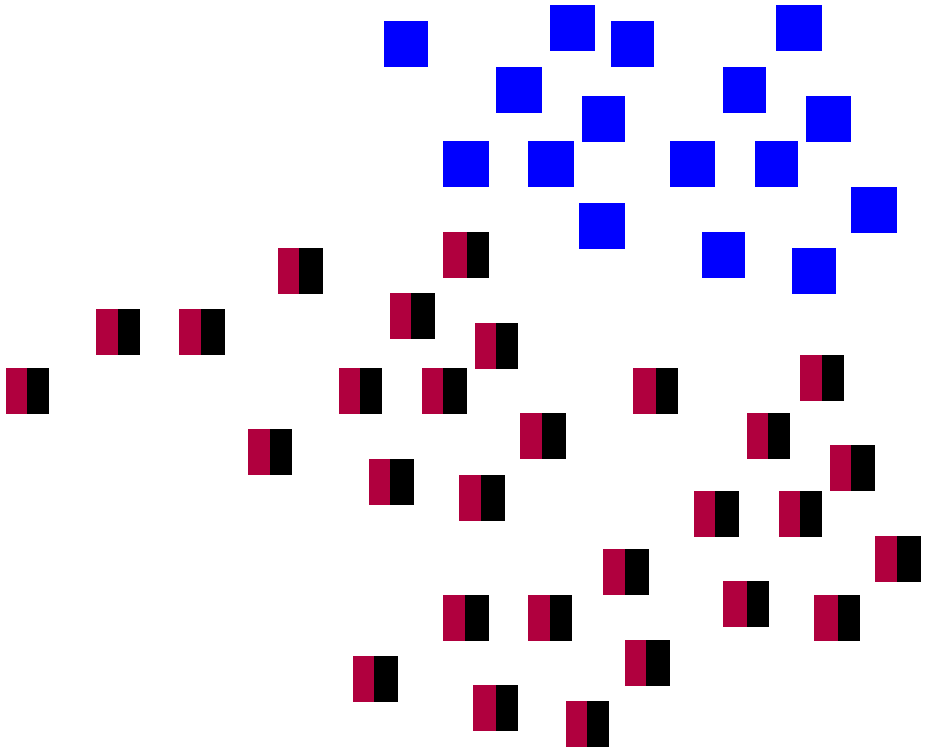
INTELLIGENT SYSTEMS

Table 1: Experimental results: Average misclassification rate on test data (with standard deviaton) for different methods, data sets, and noise levels.

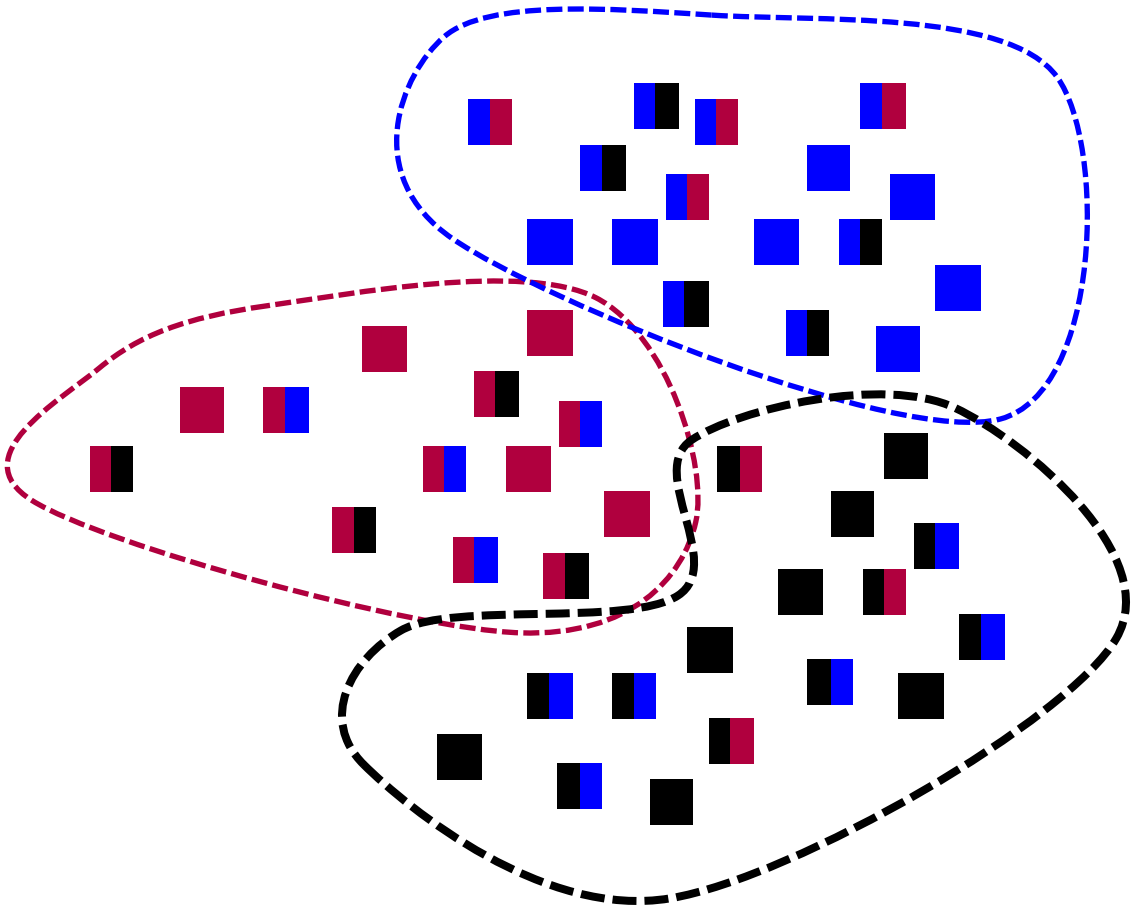
perc	data sets	SVM	OWSVM	RSVM	FLSVM
0%	Wdbc	0.0281 (0.0114)	0.0263 (0.0087)	**0.0228** (0.0100)	0.0374 (0.0159)
	Bupa	0.3188 (0.0928)	0.3043 (0.0774)	**0.3072** (0.0776)	0.3188 (0.0934)
	Banknote	0.0153 (0.0110)	**0.0095** (0.0050)	0.0153 (0.0083)	0.0124 (0.0101)
	Parkinsons	0.1333 (0.0334)	0.1128 (0.0292)	**0.1077** (0.0215)	**0.1077** (0.0215)
10%	Wdbc	0.0387 (0.0133)	**0.0281** (0.0144)	0.0316 (0.0146)	0.0334 (0.0130)
	Bupa	0.3391 (0.0442)	**0.3304** (0.0527)	0.3159 (0.0635)	0.3159 (0.0720)
	Banknote	0.0233 (0.0063)	0.0168 (0.0110)	0.0146 (0.0068)	**0.0131** (0.0095)
	Parkinsons	0.1385 (0.0229)	0.1231 (0.0215)	0.1231 (0.0215)	**0.1179** (0.0215)
20%	Wdbc	0.0615 (0.0124)	0.0474 (0.0100)	**0.0386** (0.0171)	0.0422 (0.0209)
	Bupa	0.3855 (0.0364)	0.3478 (0.0369)	**0.3275** (0.0873)	0.3362 (0.0601)
	Banknote	0.0241 (0.0050)	0.0248 (0.0056)	0.0211 (0.0125)	**0.0175** (0.0075)
	Parkinsons	0.1385 (0.0466)	0.1333 (0.0493)	**0.1279** (0.0429)	0.1436 (0.0493)
30%	Wdbc	0.0791 (0.0270)	0.0633 (0.0245)	0.0633 (0.0314)	**0.0580** (0.0302)
	Bupa	0.3884 (0.0854)	**0.3710** (0.0861)	0.3826 (0.1033)	**0.3710** (0.1018)
	Banknote	0.0313 (0.0110)	0.0277 (0.0077)	0.0270 (0.0092)	**0.0255** (0.0215)
	Parkinsons	0.1846 (0.0459)	0.1897 (0.0693)	0.1846 (0.0712)	**0.1692** (0.0716)

Under what conditions is (successful) learning in the superset setting actually possible?





systematic imprecisiation



non-systematic imprecisiation

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

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

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.

The **balanced benefit condition**:

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

where $\mathcal{R}^S(h)$ 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 h^* is the Bayes predictor and $\hat{h}$ the empirical (superset) risk minimizer; in general, Δ cannot be made arbitrarily small.

SUMMARY AND OUTLOOK

INTELLIGENT SYSTEMS

- 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.

REFERENCES

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

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.

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

D.B. Rubin. **Inference and missing data**. Biometrika, 63(3):581–592, 1976.

D. F. Heitjan and D. B. Rubin. **Ignorability and coarse data**. The Annals of Statistics, 19(4):2244–2253, 1991.