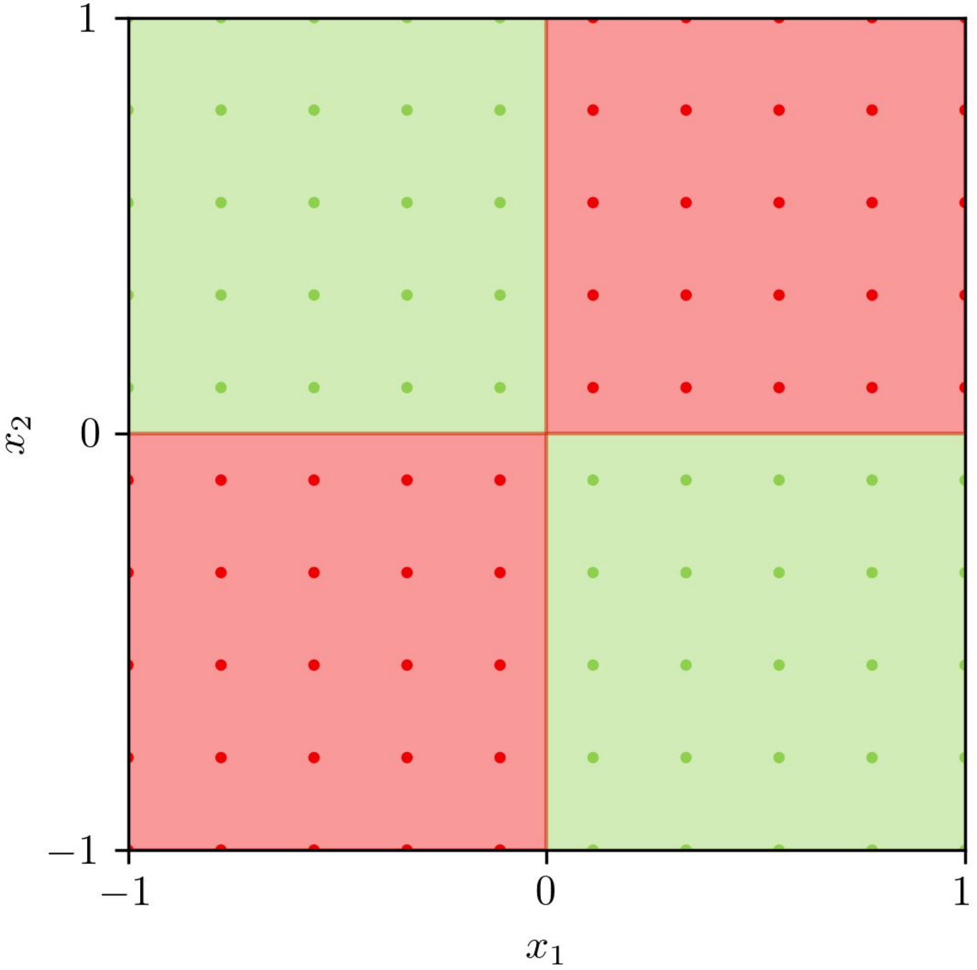
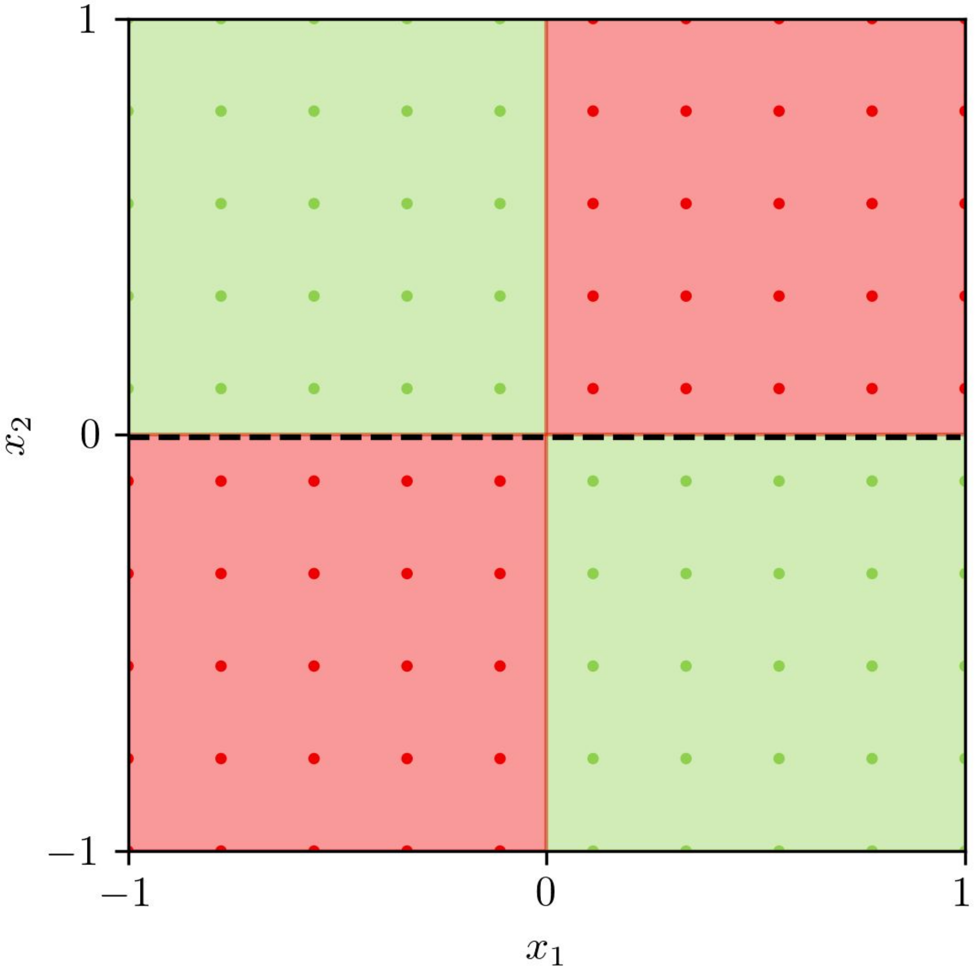
Unifying Predictive Multiplicity for Classification and Link Prediction

Lukas Harsch und Jonathan Schnitzler



Classification task:

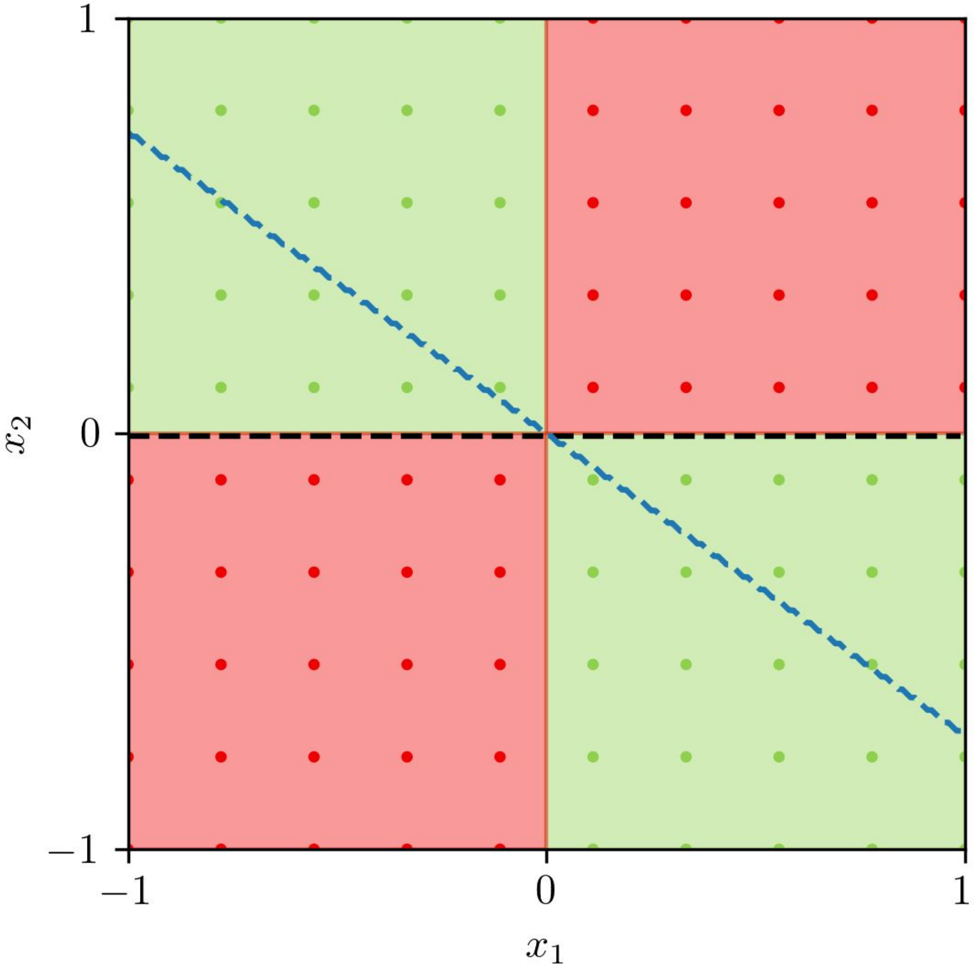
$$x = \{x_1, x_2\} \rightarrow y = \{+1, -1\}$$



Classification task:

$$x = \{x_1, x_2\} \to y = \{+1, -1\}$$

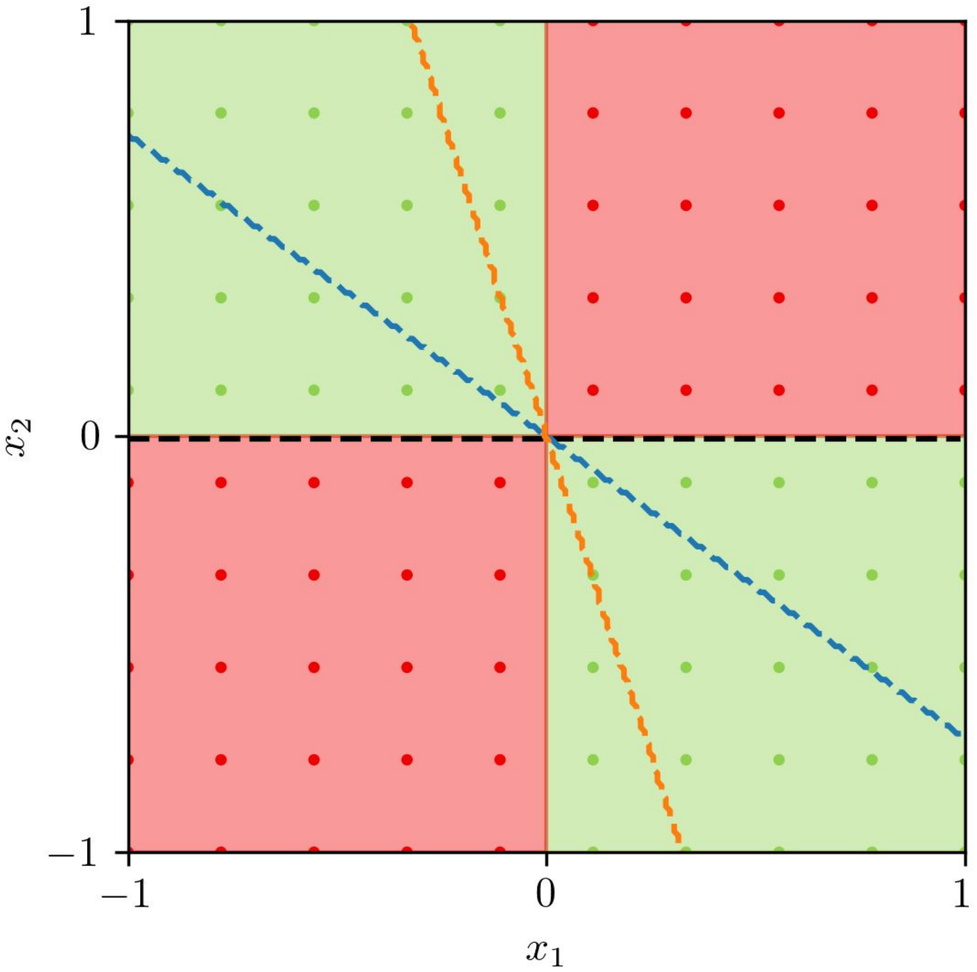
- SVM as classifier



Classification task:

$$x = \{x_1, x_2\} \rightarrow y = \{+1, -1\}$$

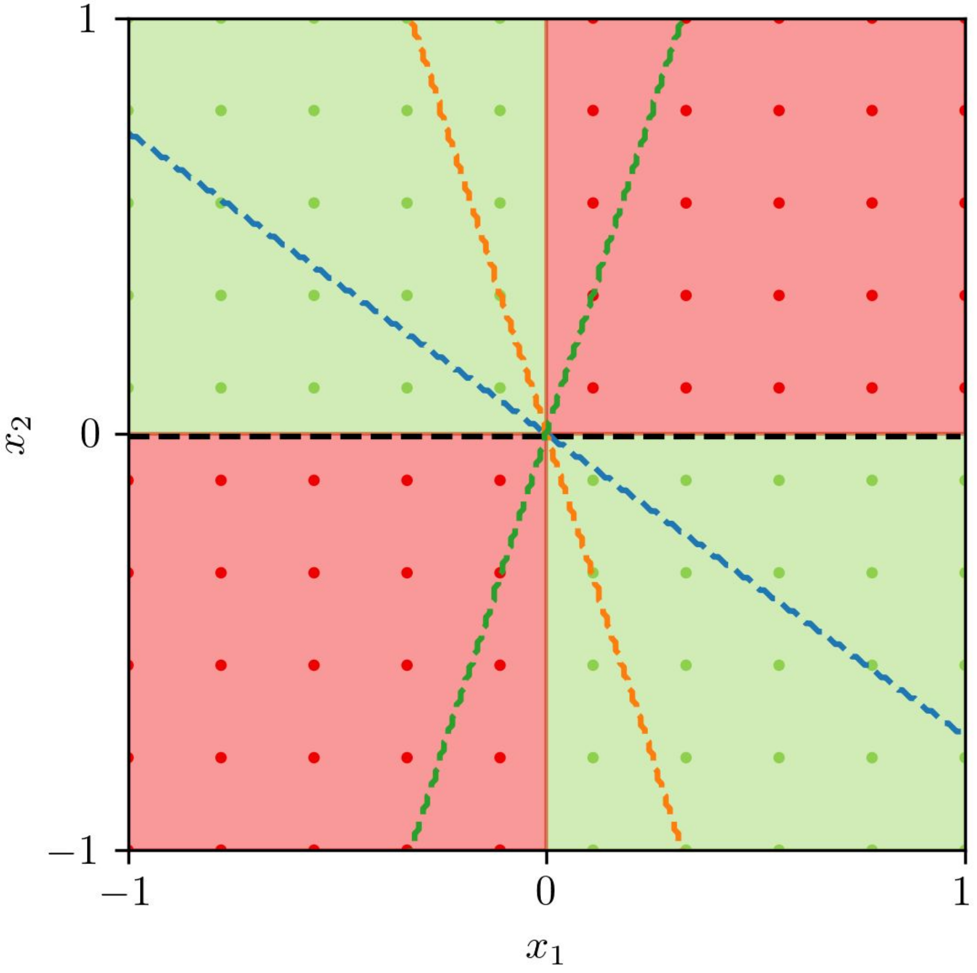
- SVM as classifier
- E.g. Rotation of decision boundary (DB) → SVMs with different parameters



Classification task:

$$x = \{x_1, x_2\} \rightarrow y = \{+1, -1\}$$

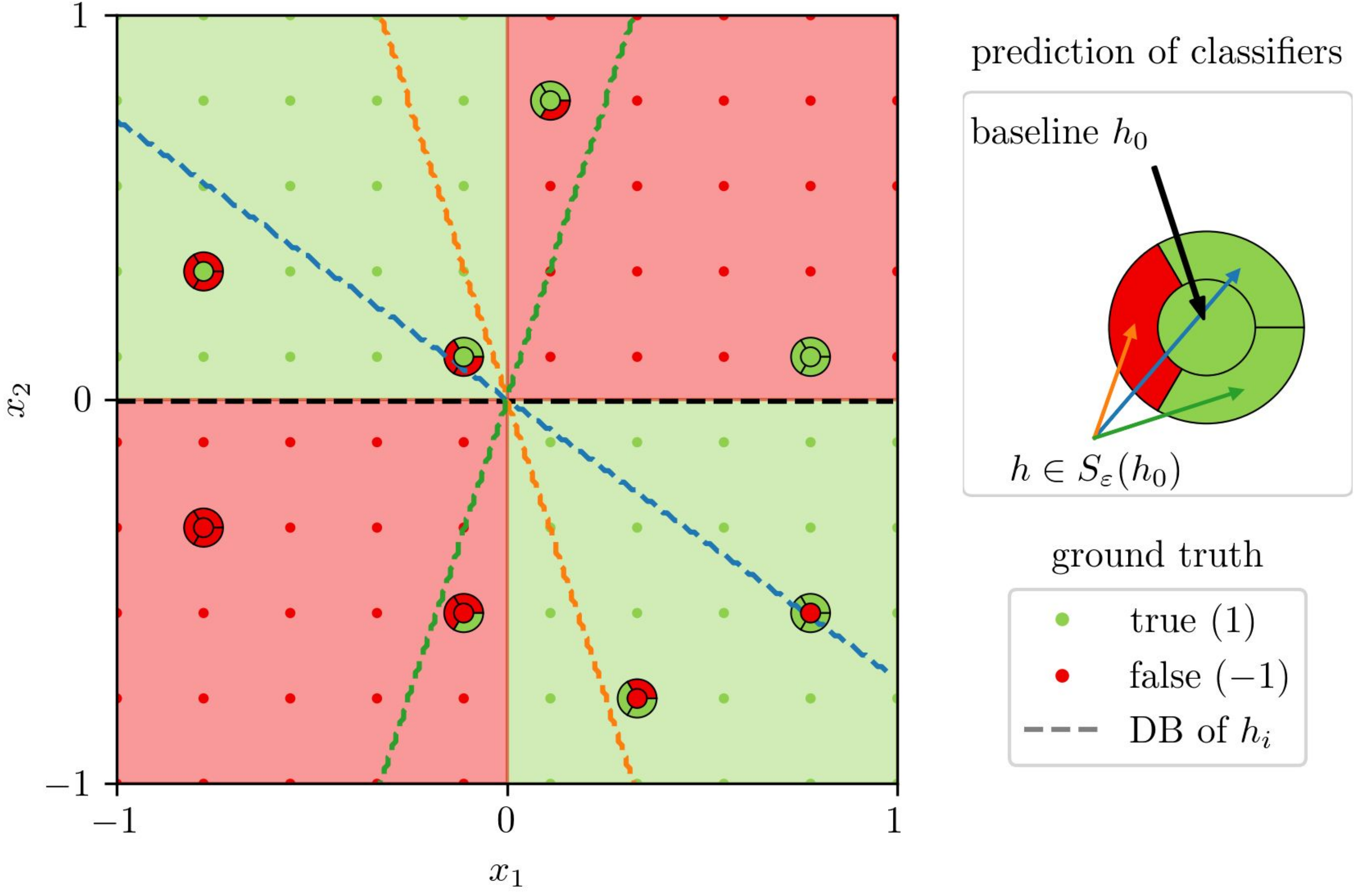
- SVM as classifier
- E.g. Rotation of decision boundary (DB) → SVMs with different parameters

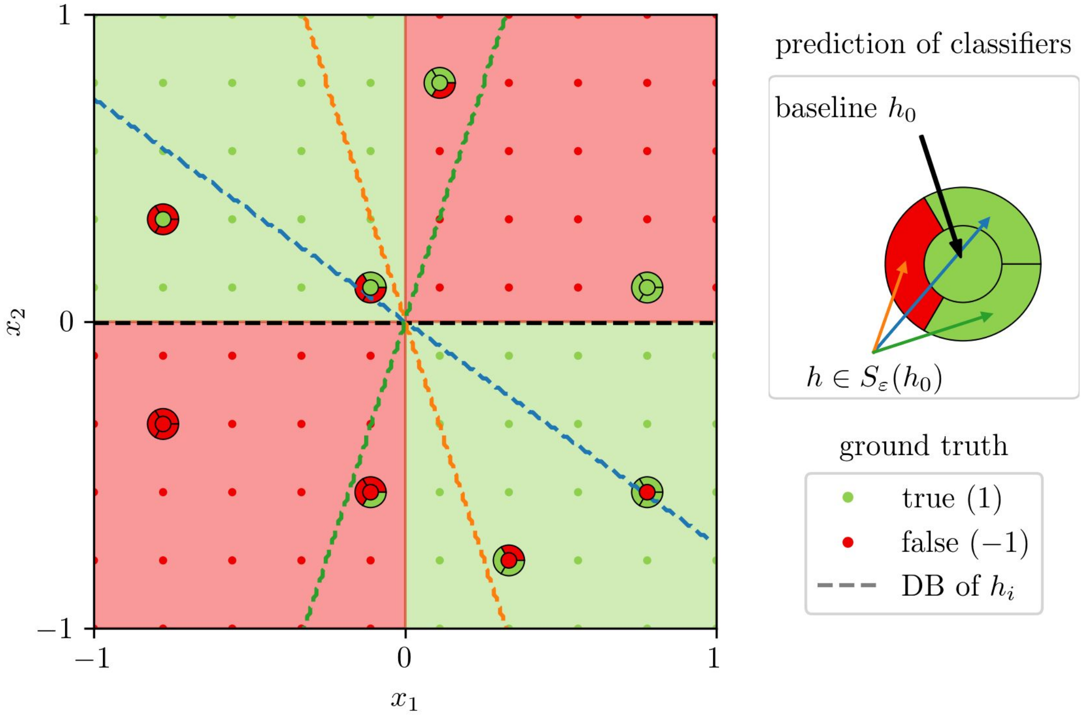


Classification task:

$$x = \{x_1, x_2\} \rightarrow y = \{+1, -1\}$$

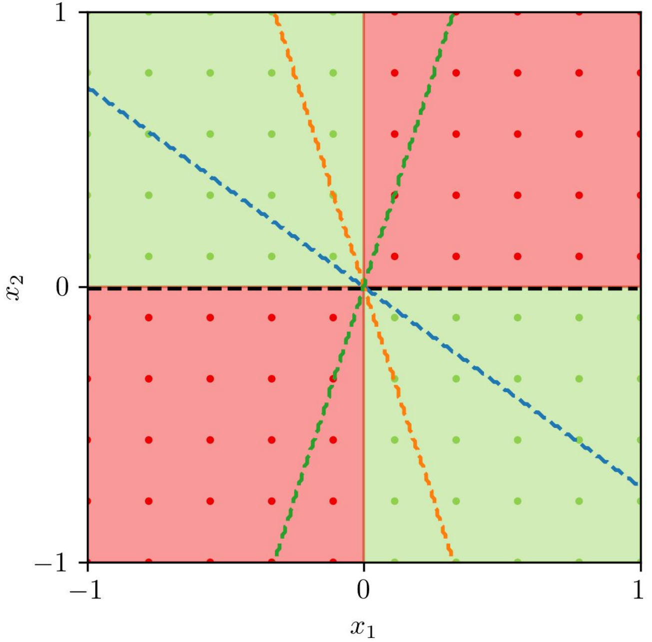
- SVM as classifier
- E.g. Rotation of decision boundary (DB)
  $\longrightarrow$ SVMs with different parameters
- All SVMs have same accuracy
  $\longrightarrow$ Multiplicity


prediction of classifiers
baseline h_0
$h \in S_\varepsilon(h_0)$
ground truth
true (1)
false (−1)
DB of h_i
x_1
x_2
−1
0
1



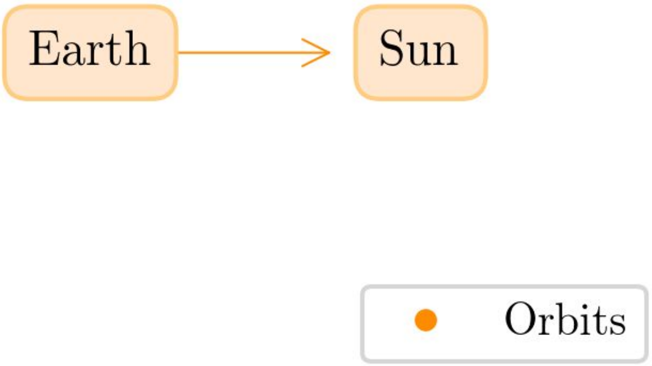
Classification task:

- For individual data points SVMs produce contradicting predictions
  → Predictive Multiplicity



$$\tau = (\mathbf{x}, y)$$

$$h : \mathbb{R}^d \to \{-1, 1\}$$



$$\tau = \langle h, r, t \rangle$$

$$h : E \times R \times E \to \mathbb{R}$$

Predictive Multiplicity

Classification [Marx et. al 2020]

- Baseline Model $h_0 \in \underset{h \in \mathcal{H}}{\arg\min} \hat{R}(h)$
- Error Rate:

$$\hat{R}(h) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[h(\mathbf{x}_i) \neq y_i]$$

- Hypothesis class $\mathcal{H}$
- Classifier $h \in \mathcal{H}$
- ε-level set:

$$S_\varepsilon(h_0) := \{h \in \mathcal{H} : \hat{R}(h) < \hat{R}(h_0) + \varepsilon\}$$

Knowledge Graph Embeddings [Zhu et al. 2024]

- Baseline Model $M_\theta^* \in \underset{M_\theta \in \mathcal{M}}{\arg\min} H_K(M_\theta)$
- Hit@K:

$$H_K(M_\theta) = \frac{1}{|\mathcal{T}|} \sum_{(q,e) \in \mathcal{T}} \mathbb{1}[R_{\succeq_{M_{\theta,q}}}(e) \leq K]$$

- Model class $\mathcal{M}$
- Model $M_\theta \in \mathcal{M}$
- ε-level set:

$$S_\varepsilon(M_\theta^*) := \{M_\theta \in \mathcal{M} | H_K(M_\theta^*) - H_K(M_\theta) \leq \varepsilon\}$$

Predictive Multiplicity

Classification [Marx et. al 2020]

- ε-level set:

$$S_\varepsilon(h_0) := \{h \in \mathcal{H} : \hat{R}(h) < \hat{R}(h_0) + \varepsilon\}$$

- Ambiguity

$$\alpha_\epsilon(h_0) := \frac{1}{n}\sum_{i=1}^{n} \max_{h \in S_\epsilon(h_0)} \mathbb{1}[h(\boldsymbol{x}_i) \neq h_0(\boldsymbol{x}_i)].$$

- Discrepancy

$$\delta_\epsilon(h_0) := \max_{h \in S_\epsilon(h_0)} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[h(\boldsymbol{x}_i) \neq h_0(\boldsymbol{x}_i)].$$

Knowledge Graph Embeddings [Zhu et al. 2024]

- ε-level set:

$$S_\varepsilon(M_\theta^*) := \{M_\theta \in \mathcal{M} | H_K(M_\theta^*) - H_K(M_\theta) \leq \varepsilon\}$$

- Ambiguity

$$\alpha_\epsilon(M_\theta^*) := \frac{1}{|\mathcal{T}|}\sum_{\tau \in \mathcal{T}} \max_{M_\theta \in S_\epsilon(M_\theta^*)} \Delta(M_\theta, \tau)$$

- Discrepancy

$$\delta_\epsilon(M_\theta^*) := \max_{M_\theta \in S_\epsilon(M_\theta^*)} \frac{1}{|\mathcal{T}|}\sum_{\tau \in \mathcal{T}} \Delta(M_\theta, \tau)$$

Predictive Multiplicity - Unified Definitions

- Unified Accuracy of a model $h \in \mathcal{H}$ over $\tau \in \mathcal{T}$ samples

$$\mathcal{L}(h) = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \mathbb{1}[S(h, \tau)]$$

Predictive Multiplicity - Unified Definitions

- Unified Accuracy of a model $h \in \mathcal{H}$ over $\tau \in \mathcal{T}$ samples

$$\mathbb{1}[R_{\preceq_{h,q}}(e) \leq K]$$ KGE - Top K Prediction

$$\mathcal{L}(h) = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \mathbb{1}[S(h, \tau)]$$

Predictive Multiplicity - Unified Definitions

- Unified Accuracy of a model $h \in \mathcal{H}$ over $\tau \in \mathcal{T}$ samples

$$\mathcal{L}(h) = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \mathbb{1}[S(h, \tau)]$$

$$\mathbb{1}[R_{\preceq_{h,q}}(e) \leq K]$$ KGE - Top K Prediction

$$\mathbb{1}[h(\mathbf{x}) = y]$$ Classification

Predictive Multiplicity - Unified Definitions

- Unified Accuracy of a model $h \in \mathcal{H}$ over $\tau \in \mathcal{T}$ samples

$$\mathcal{L}(h) = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \mathbb{1}[S(h, \tau)]$$

$$\mathbb{1}[R_{\preceq_{h,q}}(e) \leq K]$$ KGE - Top K Prediction

$$\mathbb{1}[h(\mathbf{x}) = y]$$ Classification

- With the Ranking of entity e given query q

$$R_{\preceq_{h,q}}(e) = |\{d \in E | e \preceq_{h,q} d\}|$$

Predictive Multiplicity - Unified Definitions

- Baseline Classifier h_0
- Epsilon set of h_0

$$D(h, h_0) := \mathcal{L}(h_0) - \mathcal{L}(h)$$

$$S_\varepsilon(h_0) := \{h \in \mathcal{H} : D(h_0, h) \leq \varepsilon\}$$

Measuring Predictive Multiplicity

$$\mathit{\Delta}(h, \tau) := \mathbb{1}[S(h, \tau) \neq S(h_0, \tau)]$$

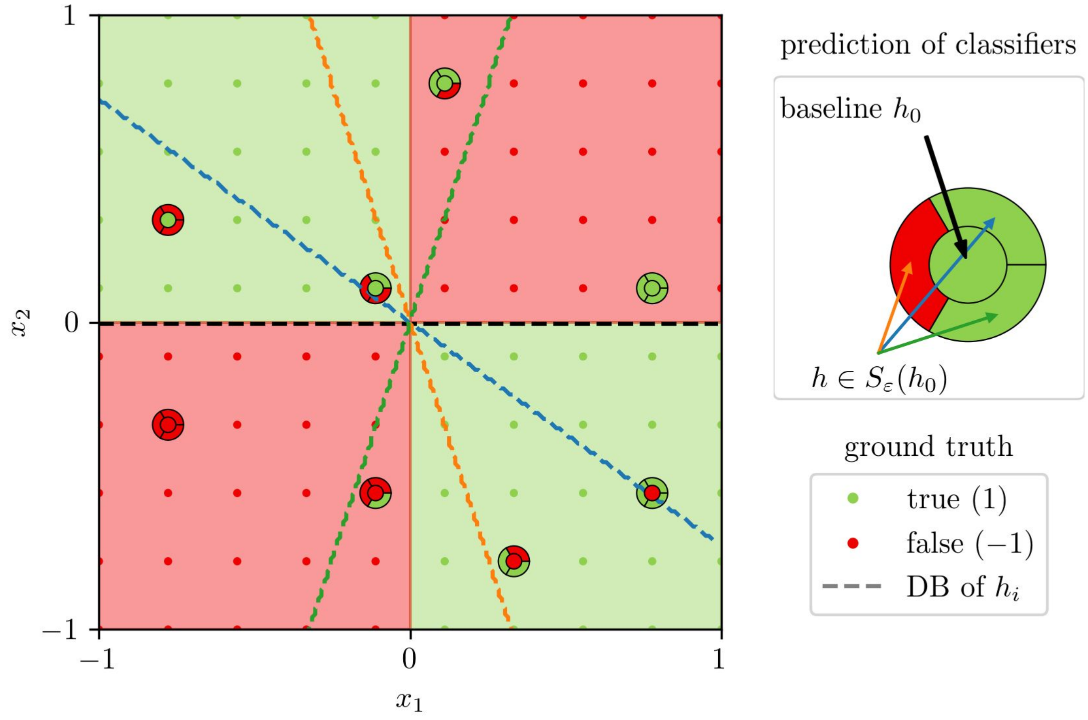
- Ambiguity

$$\alpha_\varepsilon(h_0) := \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \max_{h \in S_\varepsilon(h_0)} \mathit{\Delta}(h, \tau)$$

- Discrepancy

$$\delta_\varepsilon(h_0) := \max_{h \in S_\varepsilon(h_0)} \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \mathit{\Delta}(h, \tau)$$

Example Classification

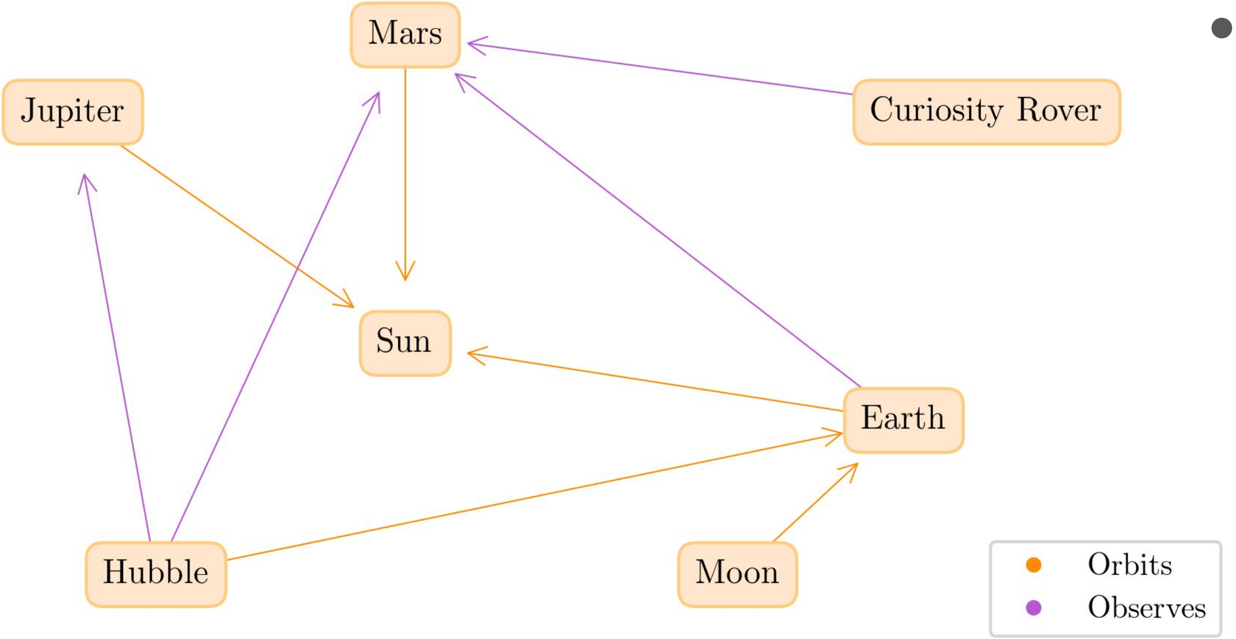


$$\alpha_0(h_0) = \frac{6}{8}$$

$$\delta_0(h_0) = \max\left[\frac{2}{8}, \frac{4}{8}, \frac{6}{8}\right]$$

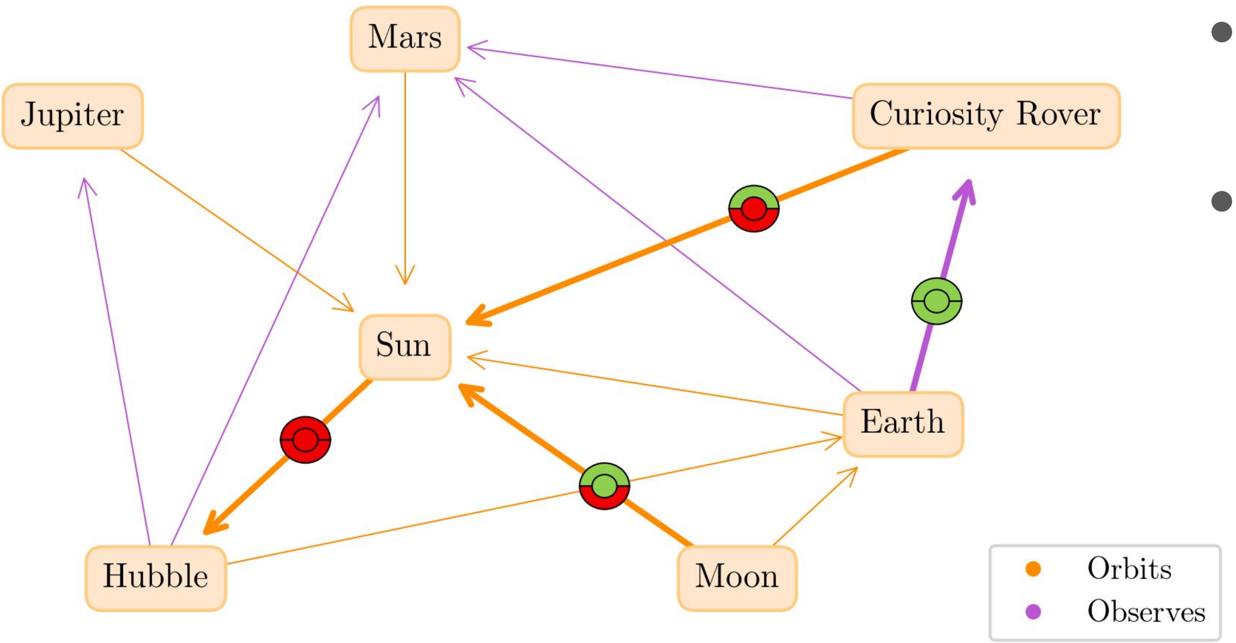
$$= \frac{6}{8}$$

Example Link Prediction for KGE



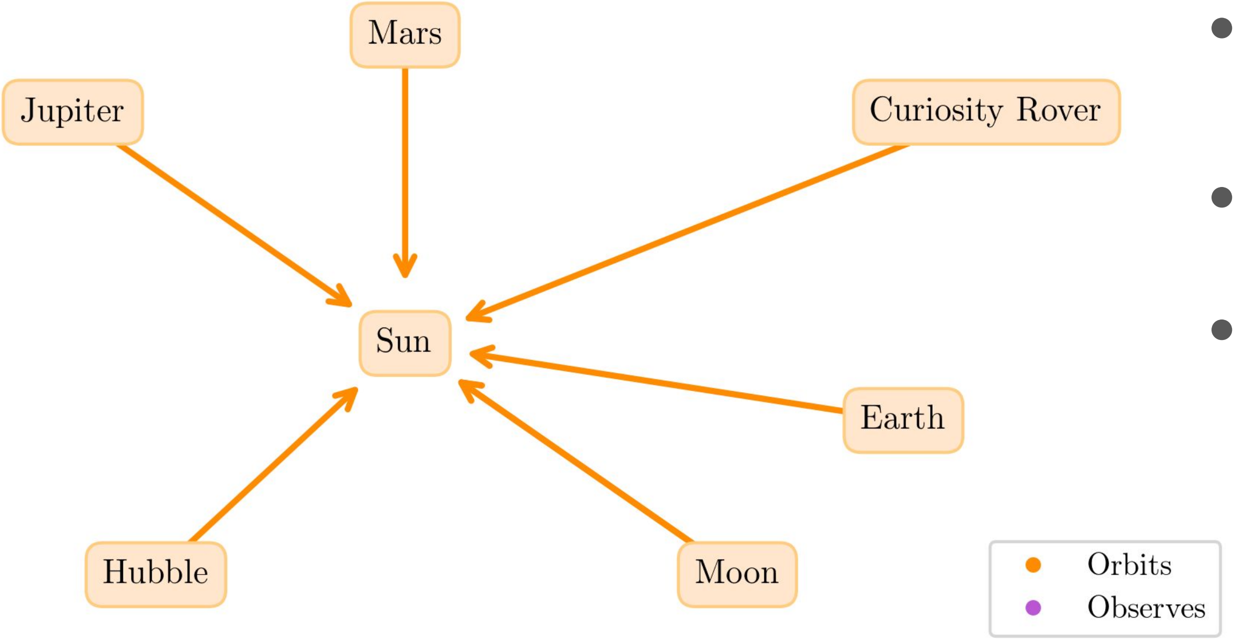
- KGE Embeddings
  - TransE
  - RotE

Example Link Prediction for KGE



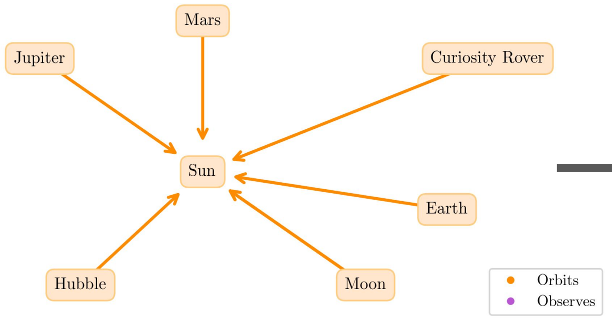
- KGE Embeddings
  - TransE
  - RotE
- Binary Glyph shows Top-4

Example Link Prediction for KGE: What orbits the sun?



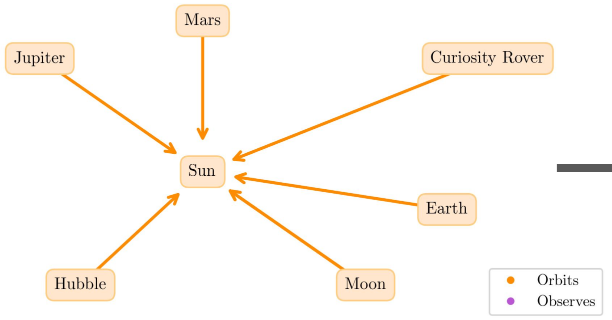
- KGE Embeddings
  - TransE
  - RotE
- Binary Glyph shows Top-4
- $\langle h, r, t \rangle \to \mathbb{R}$

Example Link Prediction for KGE: What orbits the sun?



Rank	$h_0(q, e)$
1	Earth (1.0)
2	Jup. (1.0)
3	Mars (1.0)
4	Moon (0.4)
5	Rover (0.3)
6	Hubble (0.2)
7	Sun (0.0)

Example Link Prediction for KGE: What orbits the sun?



Rank	$h_0(q, e)$	$h_1(q, e)$	$h_2(q, e)$
1	Earth (1.0)	Mars (109.5)	Mars (3.4)
2	Jup. (1.0)	Jup. (77.7)	Jup. (3.1)
3	Mars (1.0)	Earth (74.8)	Earth (2.8)
4	Moon (0.4)	Hubble (50.4)	Moon (0.0)
5	Rover (0.3)	Moon (29.1)	Hubble (-1.9)
6	Hubble (0.2)	Rover (25.3)	Rover (-1.9)
7	Sun (0.0)	Sun (20.6)	Sun (-2.9)

Voting methods in Link Prediction

- Ensemble learning
- Majority Voting:

Rank	$h_0(q,e)$	$h_1(q,e)$	$h_2(q,e)$
1	Earth (1.0)	Mars (109.5)	Mars (3.4)
2	Jup. (1.0)	Jup. (77.7)	Jup. (3.1)
3	Mars (1.0)	Earth (74.8)	Earth (2.8)
4	Moon (0.4)	Hubble (50.4)	Moon (0.0)
5	Rover (0.3)	Moon (29.1)	Hubble (-1.9)
6	Hubble (0.2)	Rover (25.3)	Rover (-1.9)
7	Sun (0.0)	Sun (20.6)	Sun (-2.9)

→

Majority
Mars (3)
Earth (1)
Jup (1)
Moon (0)
Hubble (0)
Sun (0)
Rover (0)

Conclusion

Predictive Multiplicity …

- … occurs in many ML tasks such as link prediction and classification.

Conclusion

Predictive Multiplicity …

- … occurs in many ML tasks such as link prediction and classification.
- … potentially undermines reliability and fairness in critical applications (e.g. recidivism prediction, granting loans).

Conclusion

Predictive Multiplicity …

- … occurs in many ML tasks such as link prediction and classification.
- … potentially undermines reliability and fairness in critical applications (e.g. recidivism prediction, granting loans).
- … quantifiable using ambiguity and discrepancy.

Conclusion

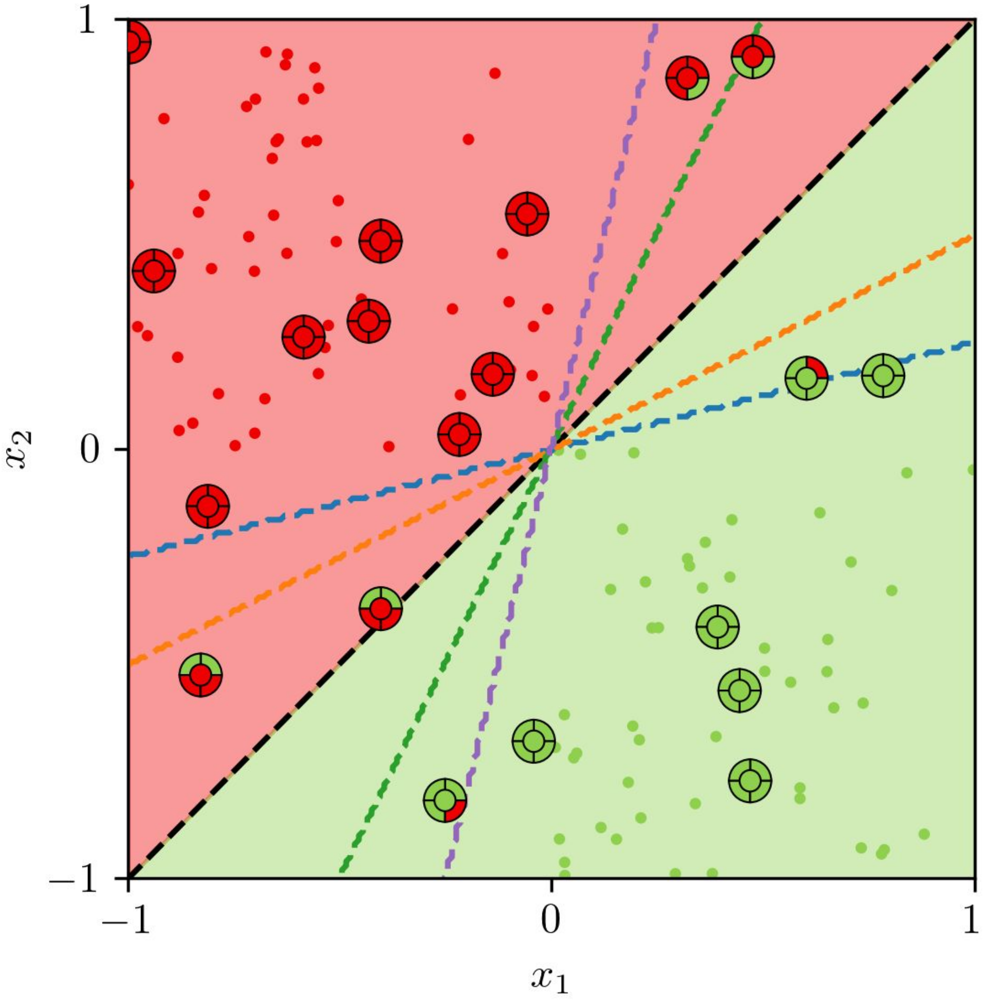
Predictive Multiplicity …

- … occurs in many ML tasks such as link prediction and classification.
- … potentially undermines reliability and fairness in critical applications (e.g. recidivism prediction, granting loans).
- … quantifiable using ambiguity and discrepancy.
- … mitigatable using voting methods (ensemble learning).

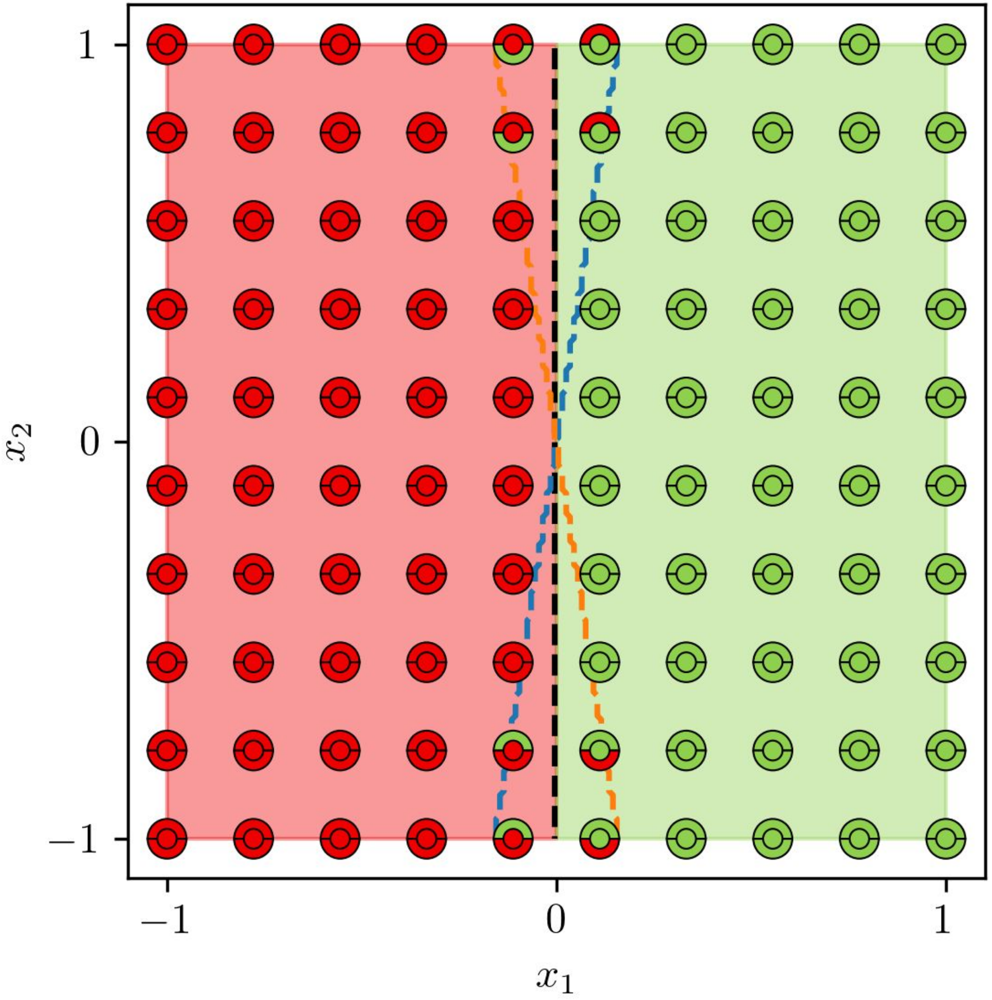
Conclusion

Predictive Multiplicity …

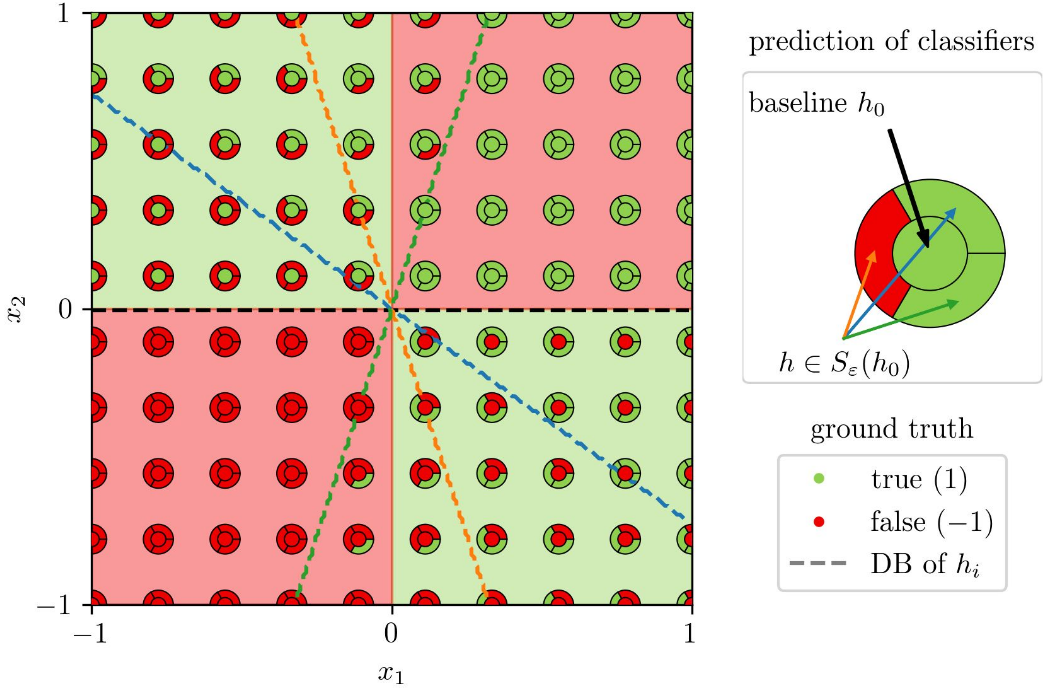
- … occurs in many ML tasks such as link prediction and classification.
- … potentially undermines reliability and fairness in critical applications (e.g. recidivism prediction, granting loans).
- … quantifiable using ambiguity and discrepancy.
- … mitigatable using voting methods (ensemble learning).
- … should be reported as inherent aspect of model performance alongside test error to ensure transparency for stakeholders.



1
0
−1
−1
0
1
x_2
x_1



1
0
−1
−1
0
1
x_2
x_1



Classification task:

- For individual data points SVMs produce contradicting predictions → Predictive Multiplicity