Decision trees with missing values

Journées de Statistique 2019

Nicolas Prost, Julie Josse, Erwan Scornet, Gaël Varoquaux

CMAP, INRIA PARIETAL

June 4, 2019

Motivating data in health

More data $\Rightarrow$ more missing data

- Data filled manually
- Faulty sensors
- Data aggregation issues

Traumabase: 15 000 patients/ 250 variables/ 15 hospitals

Center	Age	Sex	Weight	Height	BMI	Lactates	Glasgow
Beaujon	54	m	85	NR	NR	NA	12
Lille	33	m	80	1.8	24.69	4.8	15
Pitie	26	m	NR	NR	NR	3.9	3
Beaujon	63	m	80	1.8	24.69	1.66	15
Pitie	30	w	NR	NR	NR	NM	15

- missing features: `Not Recorded`, `Not Made`, `Not Applicable`, etc.
- aim: predict the Glasgow score

Supervised learning

What it is

- A feature matrix $\mathbf{X}$ and a response vector Y
- A **training** data set and a **test** data set
- A loss function to minimize

Goal:

$$f^\star \in \operatorname*{argmin}_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}\left[(f(\mathbf{X}) - Y)^2\right].$$

Tool:

$$\hat{f}_n \in \operatorname*{argmin}_{f:\mathcal{X}\to\mathcal{Y}} \left(\frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{X}_i) - Y_i)^2\right).$$

- How to learn a prediction function when data is missing?
- How to predict on a test set when data is missing?

Classification And Regression Trees (CART)[1]

Decision trees

Method: Recursively, find which split minimises the (quadratic) loss, *i.e.* which

- feature $j^\star$
- threshold $z^\star$

minimise

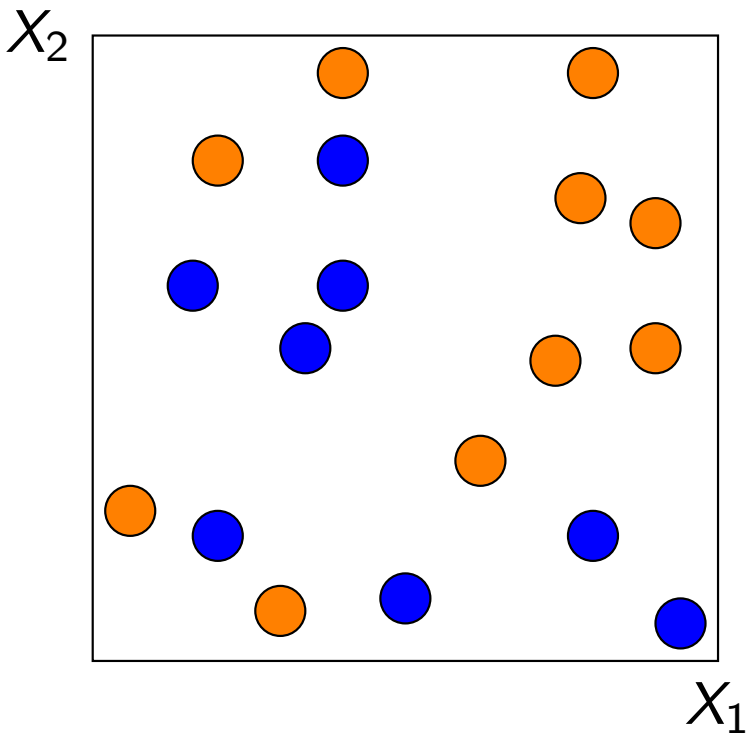
$$\mathbb{E}\Big[\big(Y - \mathbb{E}[Y|X_j \leq z]\big)^2 \cdot \mathbb{1}_{X_j \leq z} + \big(Y - \mathbb{E}[Y|X_j > z]\big)^2 \cdot \mathbb{1}_{X_j > z}\Big].$$

[1] Leo Breiman et al. *Classification and Regression Trees*. Wadsworth, 1984.

Classification And Regression Trees (CART)

Decision trees

Method: Recursively, find which split minimises the (quadratic) loss

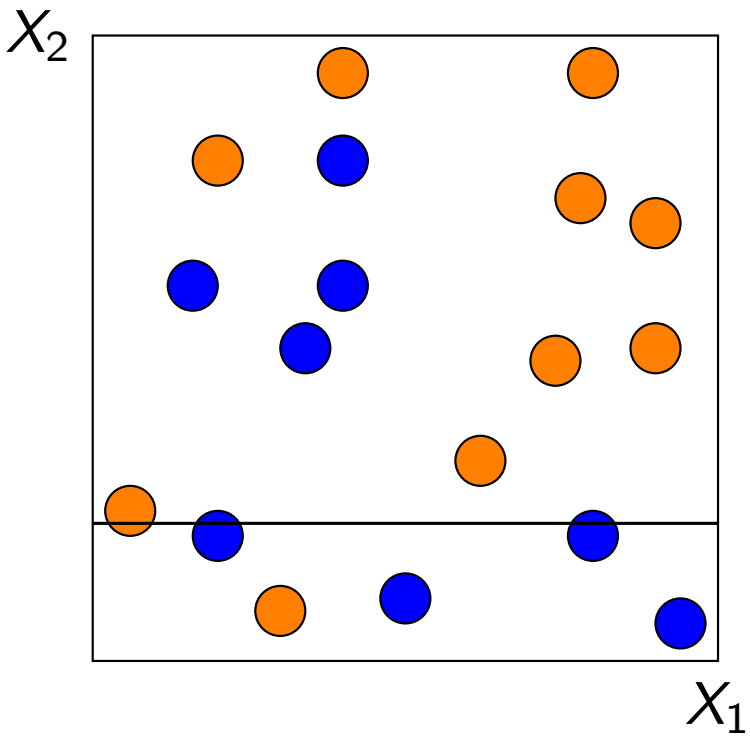


root

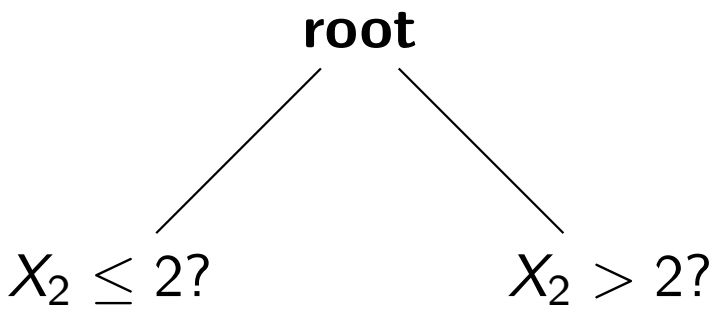
Classification And Regression Trees (CART)

Decision trees

Method: Recursively, find which split minimises the (quadratic) loss



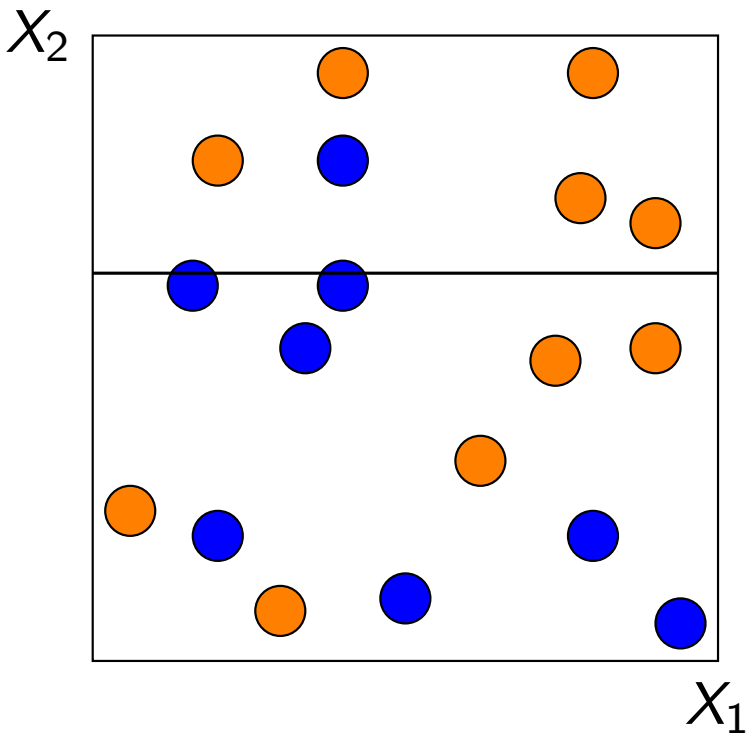
loss $= 3.57$



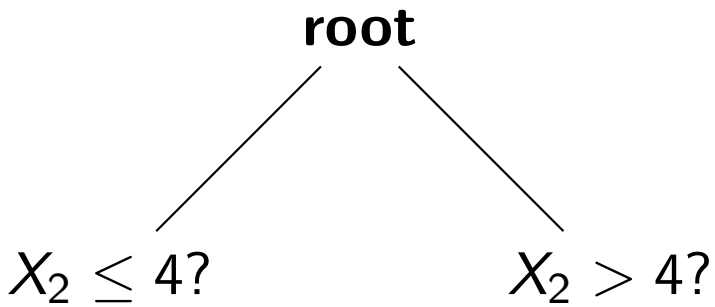
Classification And Regression Trees (CART)

Decision trees

Method: Recursively, find which split minimises the (quadratic) loss



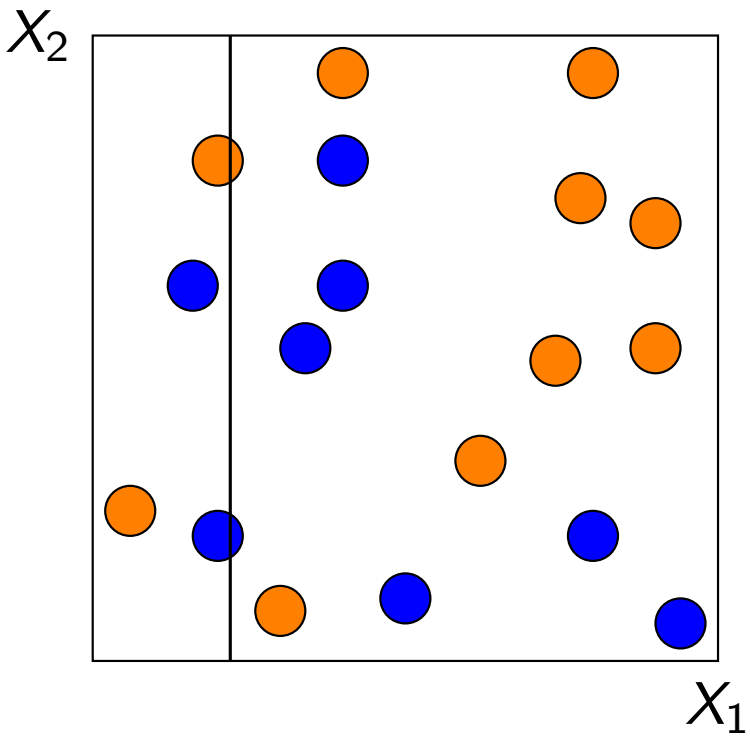
loss $= 3.75$



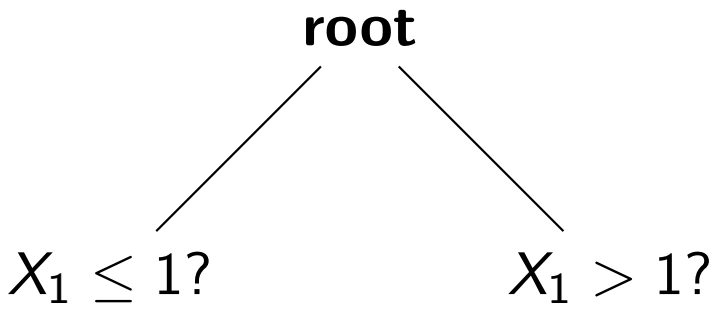
Classification And Regression Trees (CART)

Decision trees

Method: Recursively, find which split minimises the (quadratic) loss



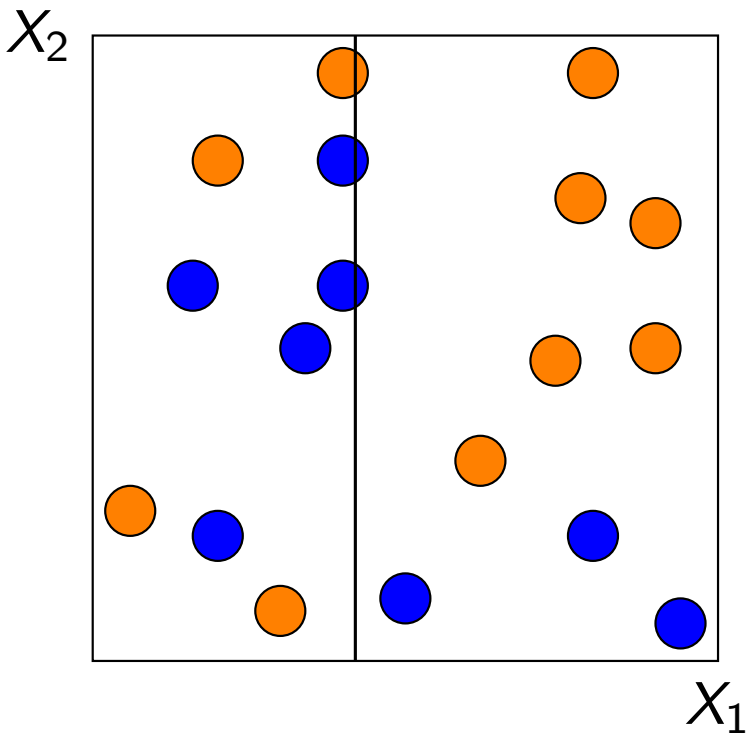
loss $= 4.43$



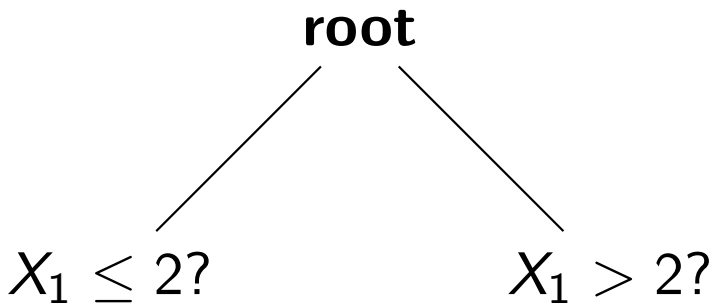
Classification And Regression Trees (CART)

Decision trees

Method: Recursively, find which split minimises the (quadratic) loss



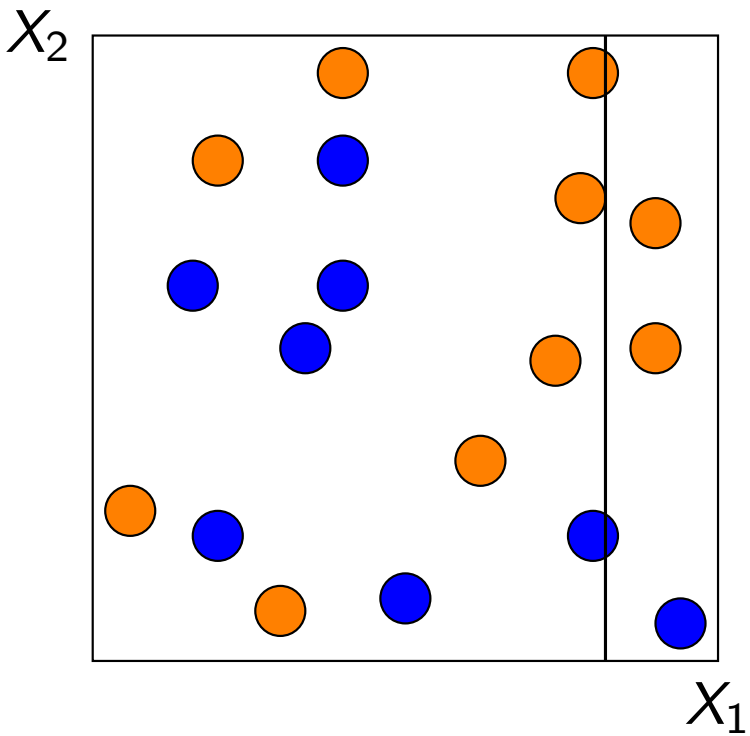
loss $= 4.22$



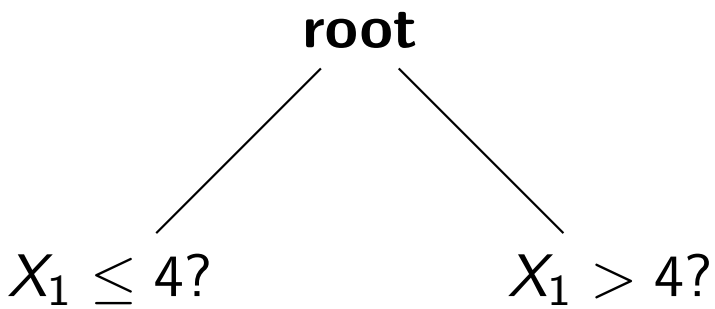
Classification And Regression Trees (CART)

Decision trees

Method: Recursively, find which split minimises the (quadratic) loss



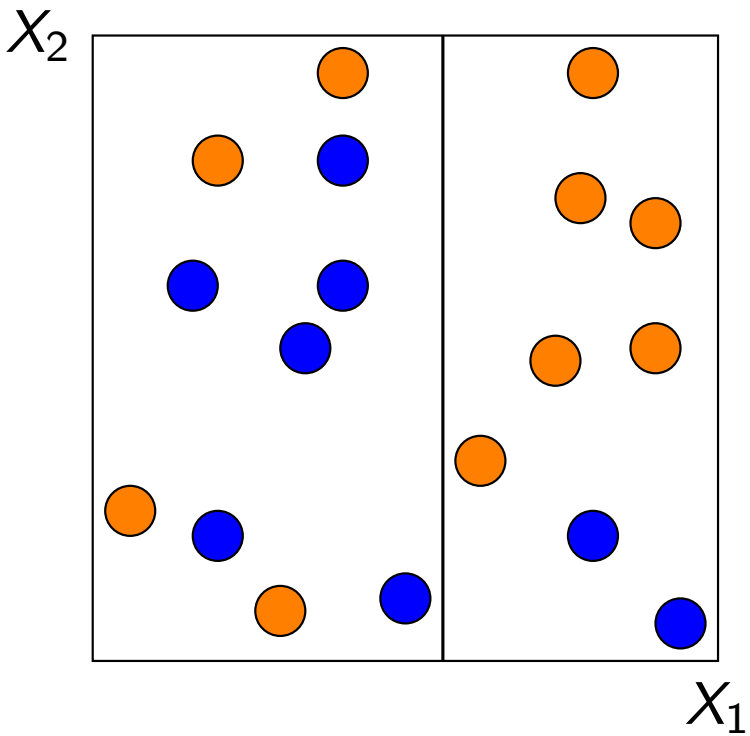
loss $= 4.40$



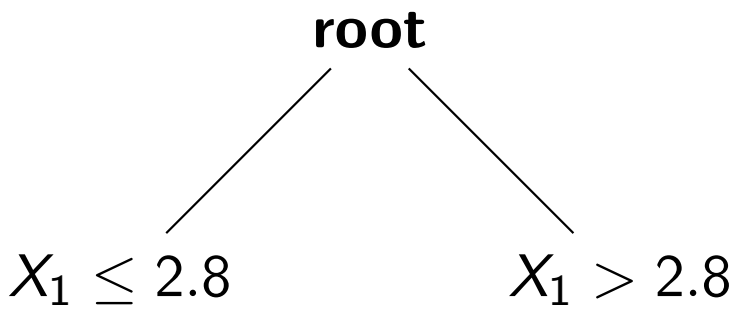
Classification And Regression Trees (CART)

Decision trees

Method: Recursively, find which split minimises the (quadratic) loss



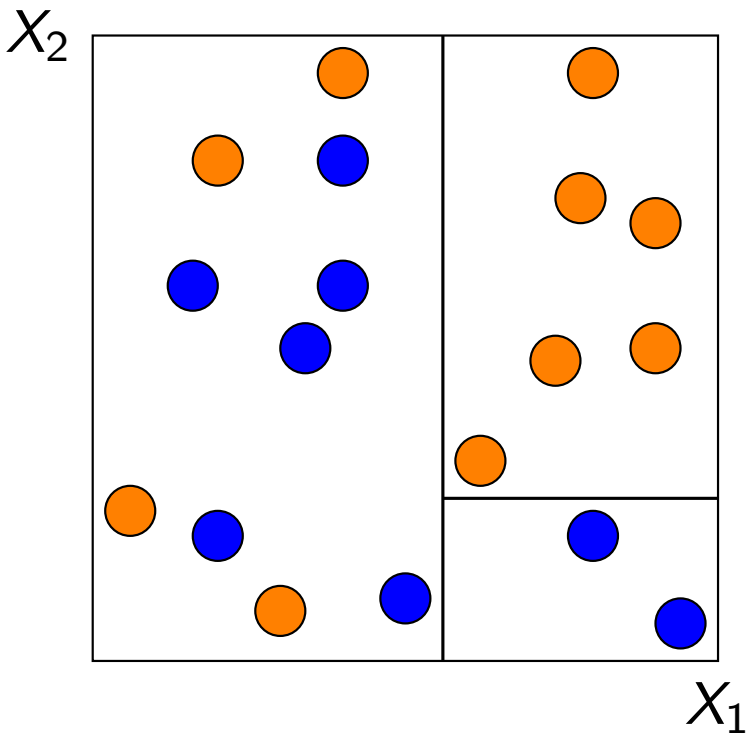
loss = 3.90

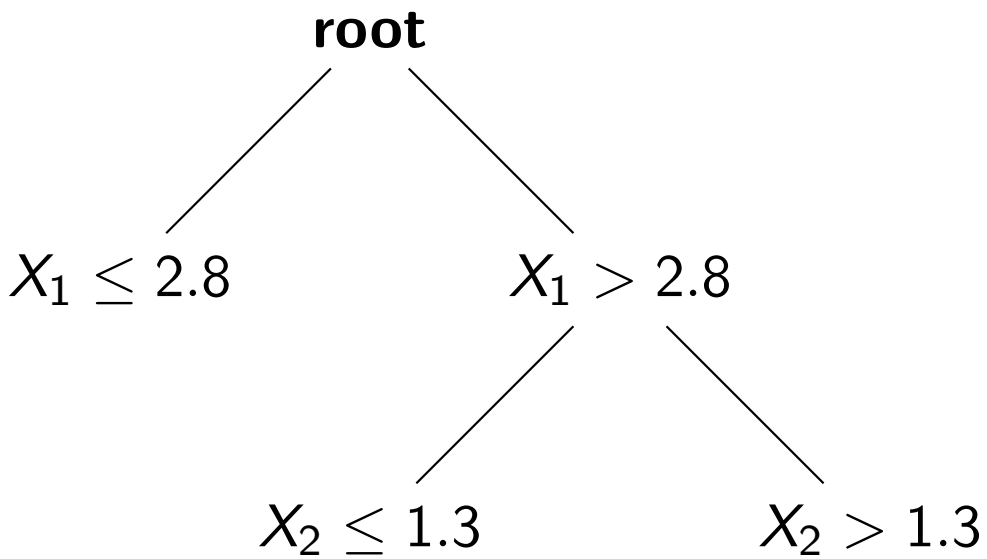


Classification And Regression Trees (CART)

Decision trees

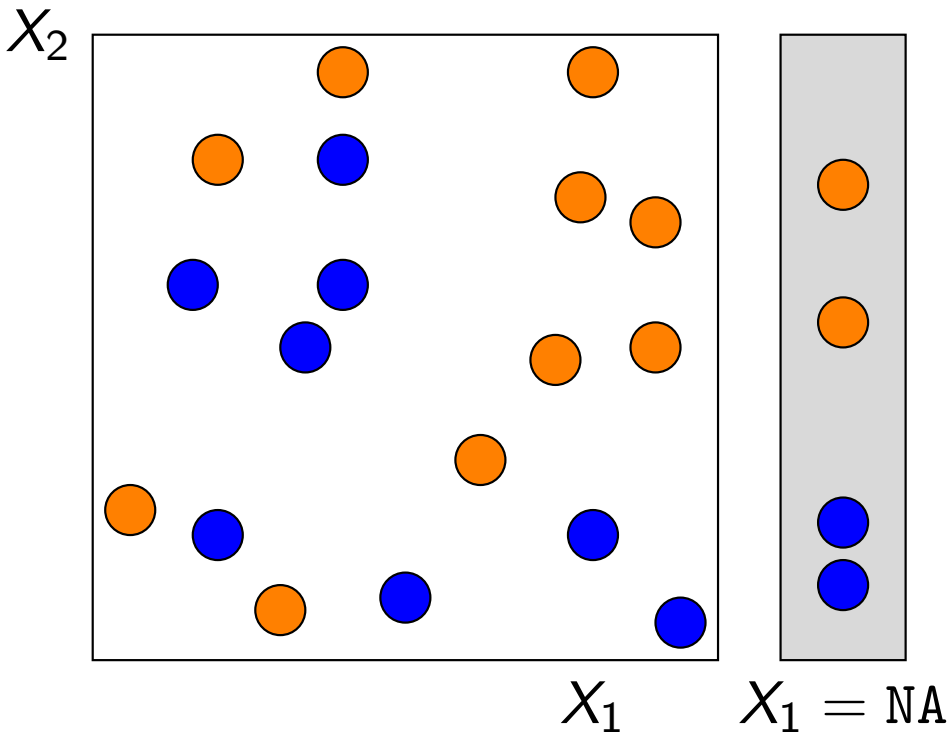
Method: Recursively, find which split minimises the (quadratic) loss





CART with missing values

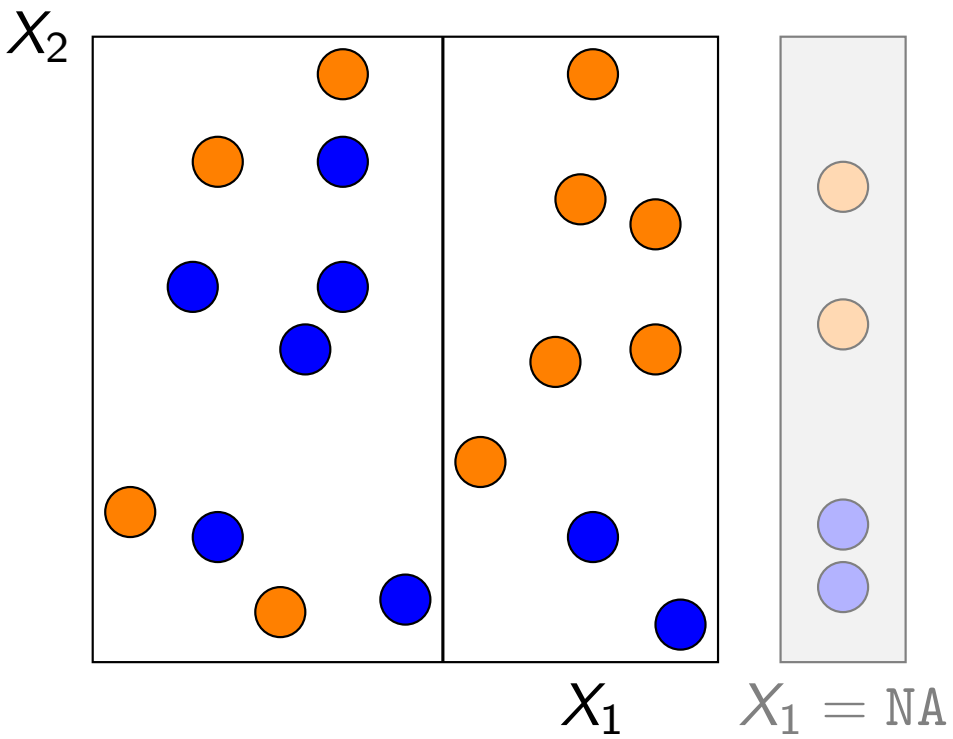
Splitting criterion discarding missing values

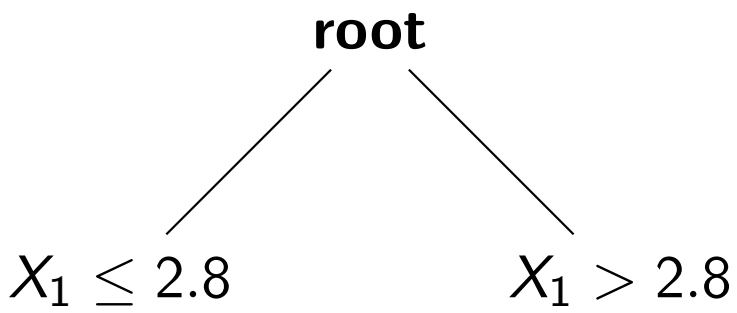


root

CART with missing values

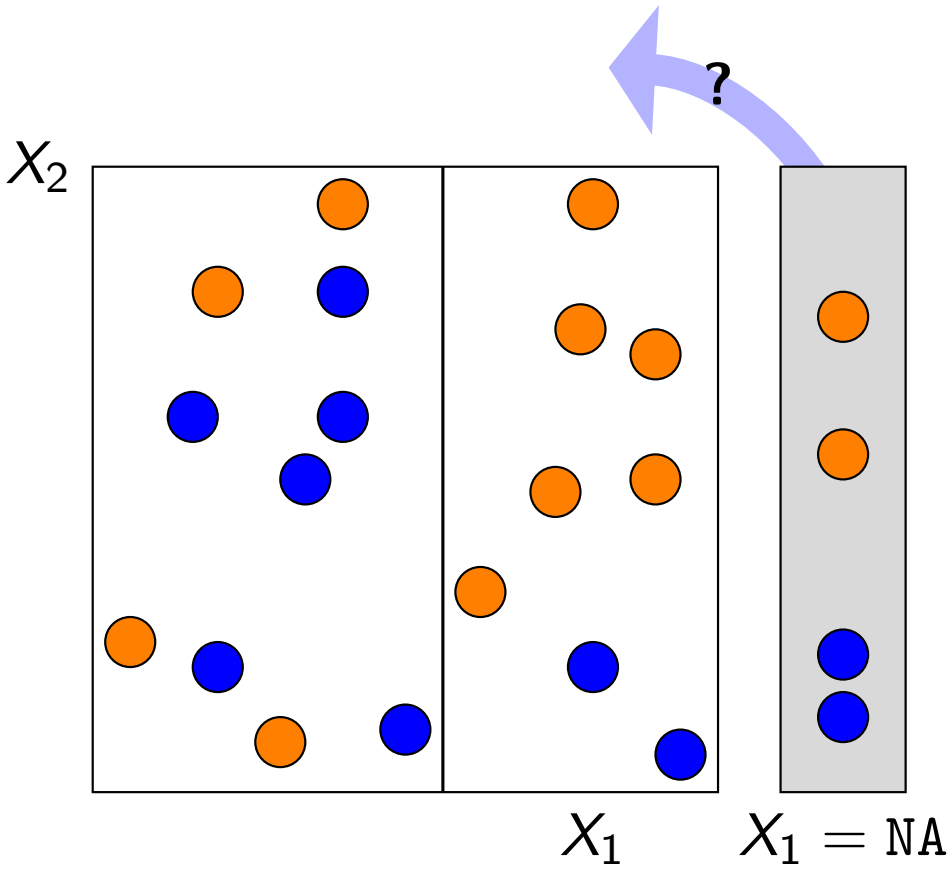
Splitting criterion discarding missing values

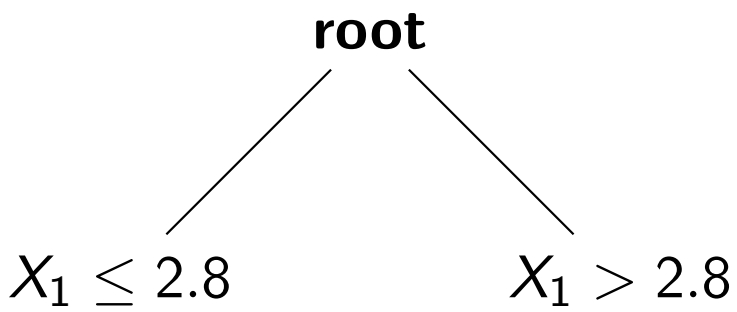




CART with missing values

Splitting criterion discarding missing values



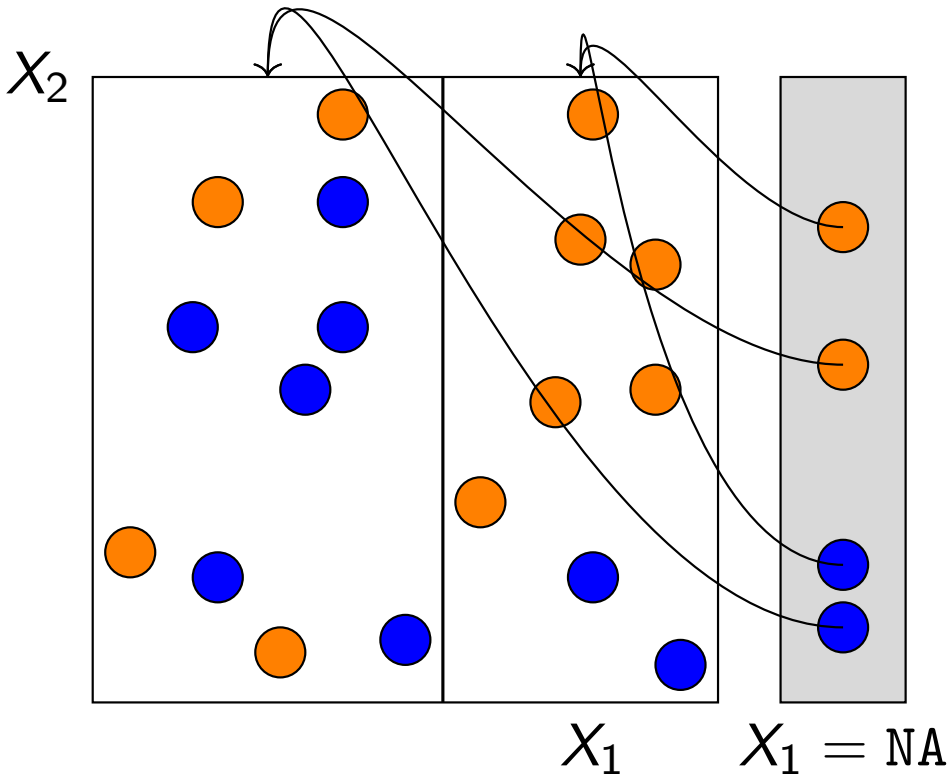


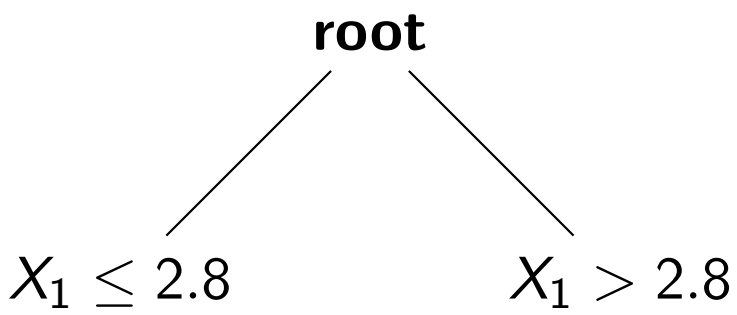
Where to send the incomplete observations?

Probabilistic split[2]

Completion strategies

Send incomplete observations to either side, flipping a coin.



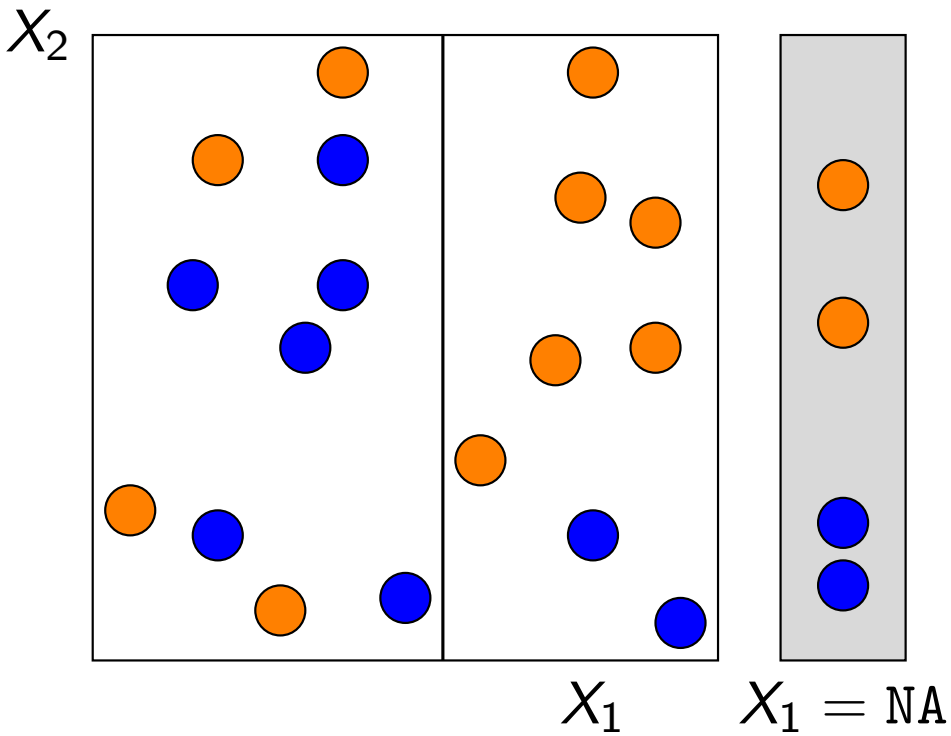


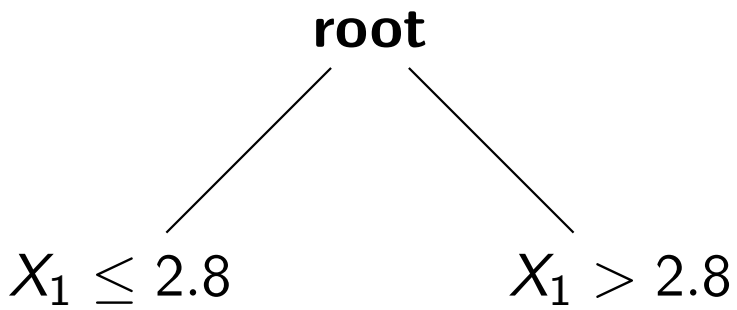
[2] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

Block propagation[3]

Completion strategies

Send all incomplete observations to the side which minimises the loss.



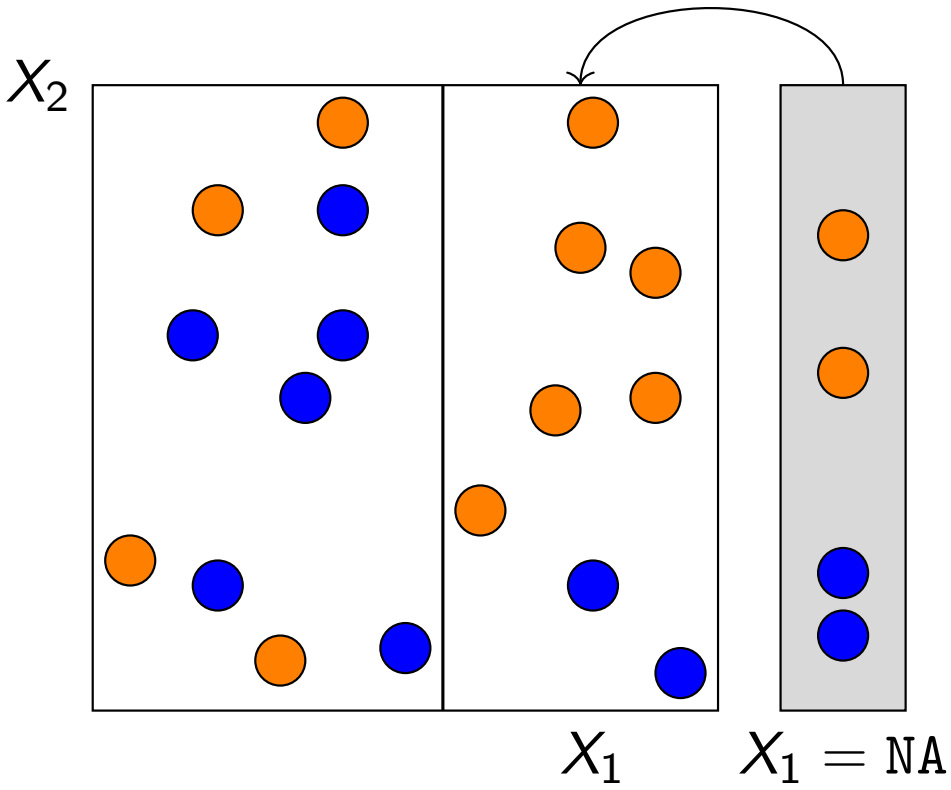


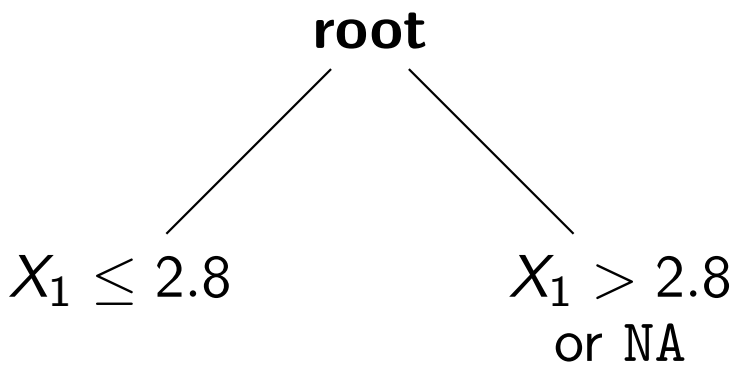
[3] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *sigkdd international conference on knowledge discovery and data mining*. ACM. 2016,

Block propagation[3]

Completion strategies

Send all incomplete observations to the side which minimises the loss.



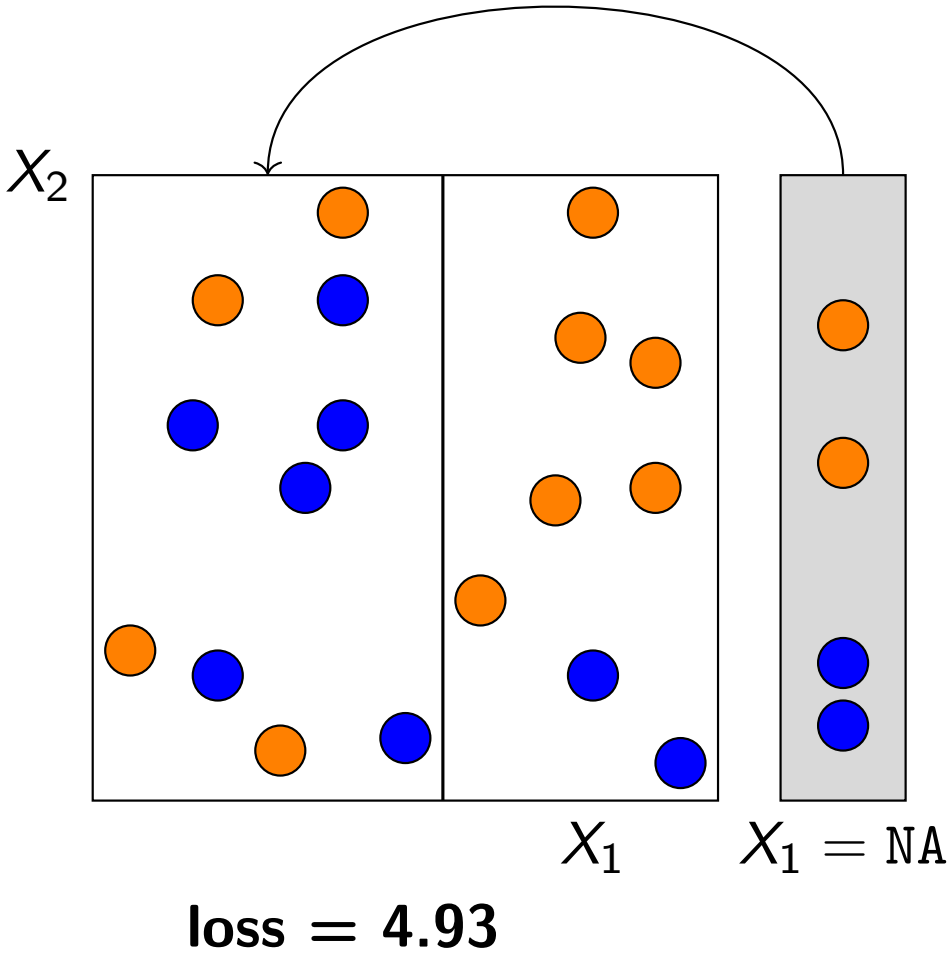


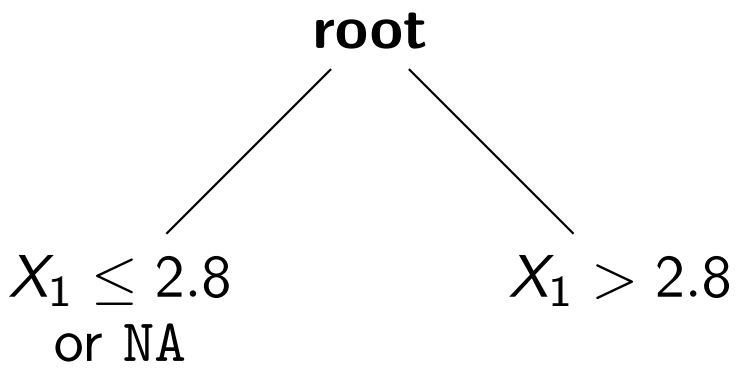
[3] Chen and Guestrin, "Xgboost: A scalable tree boosting system".

Block propagation[3]

Completion strategies

Send all incomplete observations to the side which minimises the loss.



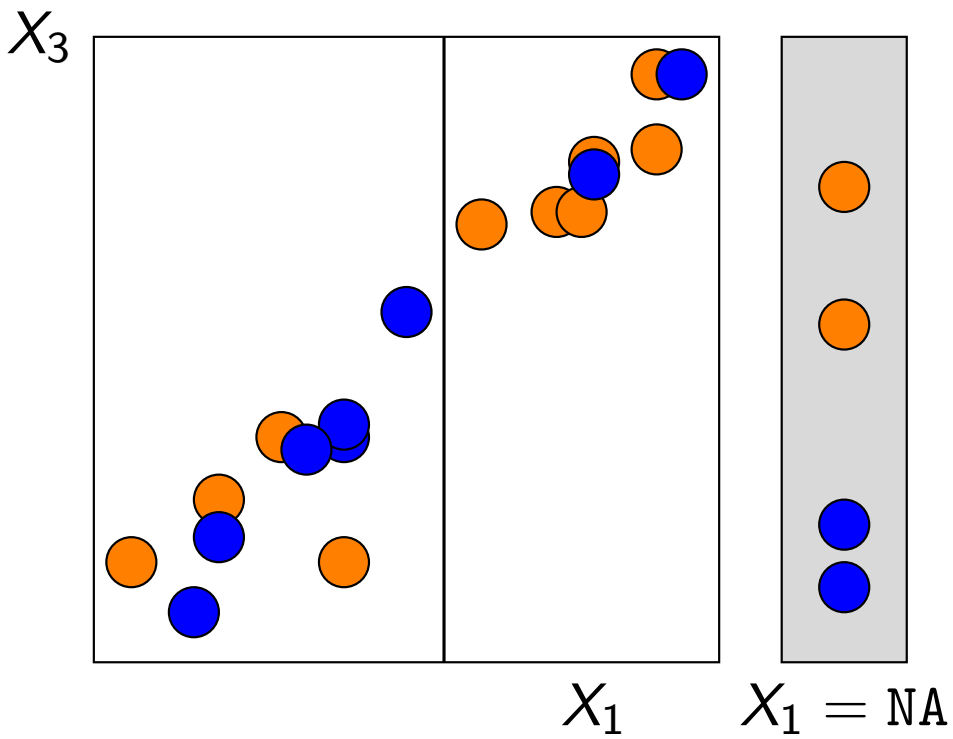


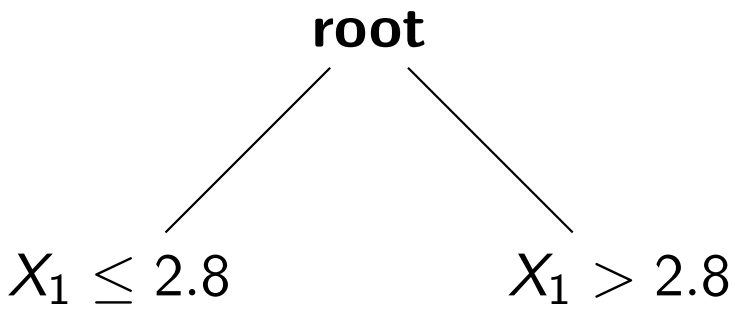
[3] Chen and Guestrin, "Xgboost: A scalable tree boosting system".

Surrogate splits[4]

Completion strategies

Find a related variable to split.



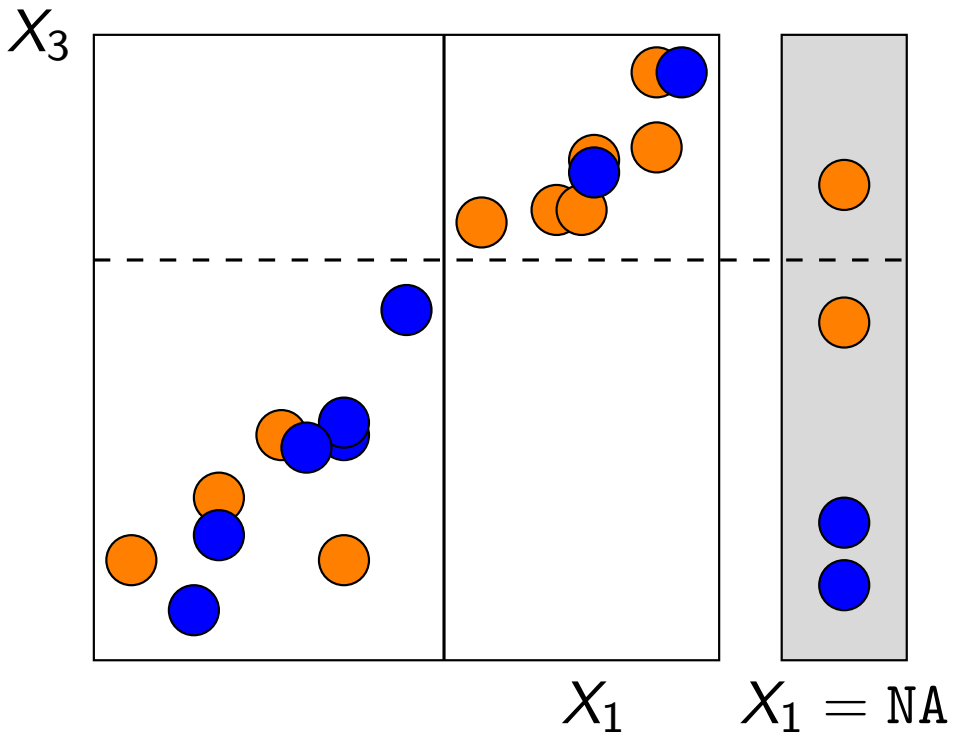


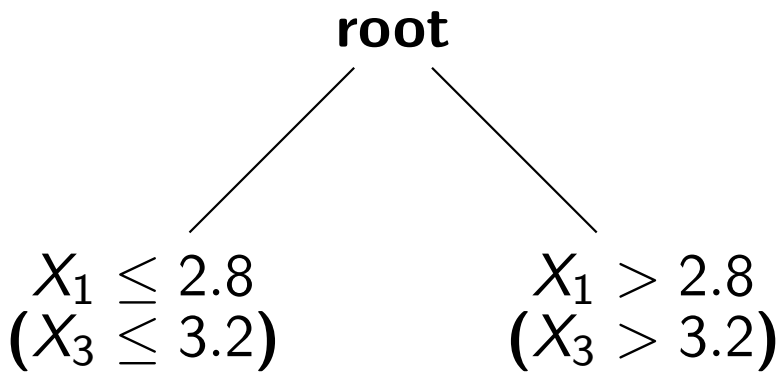
[4] Breiman et al., *Classification and Regression Trees*.

Surrogate splits[4]

Completion strategies

Find a related variable to split.





[4] Breiman et al., *Classification and Regression Trees*.

Variable selection bias

Issue with discarding missing values: variables with more values are selected more frequently.

Example: X_1 and X_2 are i.i.d., independent from Y

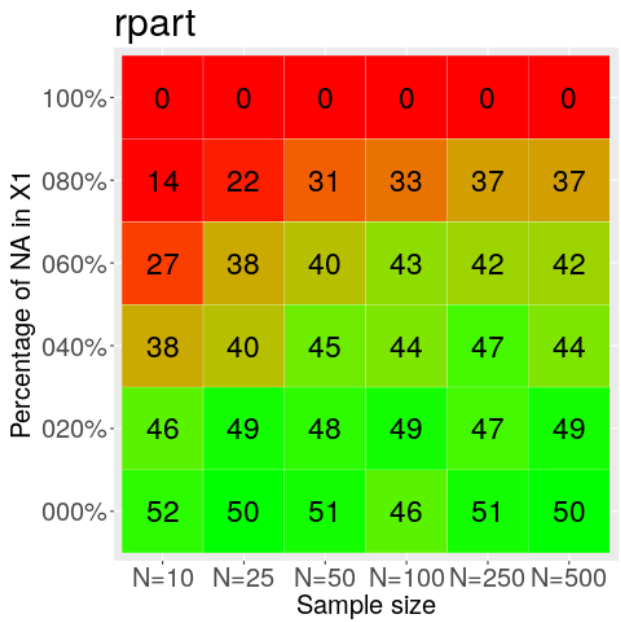


Figure: X_1 selection frequency for `rpart` (CART)

Conditional inference trees[6]

Alterative splitting strategy

Separate

- choice of splitting variable: test "$Y \perp\!\!\!\perp X^j$" based on

$$T(X_j) = \sum_{X_i^j \, observed} X_i^j Y_i$$

- threshold choice: usual impurity.

In order to test by permutation, consider the distribution of all the $\sum_{X_i^j \, observed} X_i^j Y_{\sigma(i)}$ where σ runs uniformly over the permutations.

This distribution is asymptotically normal[5],which enables the definition of a p-value.

---

[5] Helmut Strasser and Christian Weber. "On the asymptotic theory of permutation statistics". In: (1999).

[6] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. "Unbiased Recursive Partitioning: A Conditional Inference Framework". In: *Journal of Computational and Graphical Statistics* 15.3 (2006).

Empirical comparison

Example: X_1 and X_2 are i.i.d., independent from Y

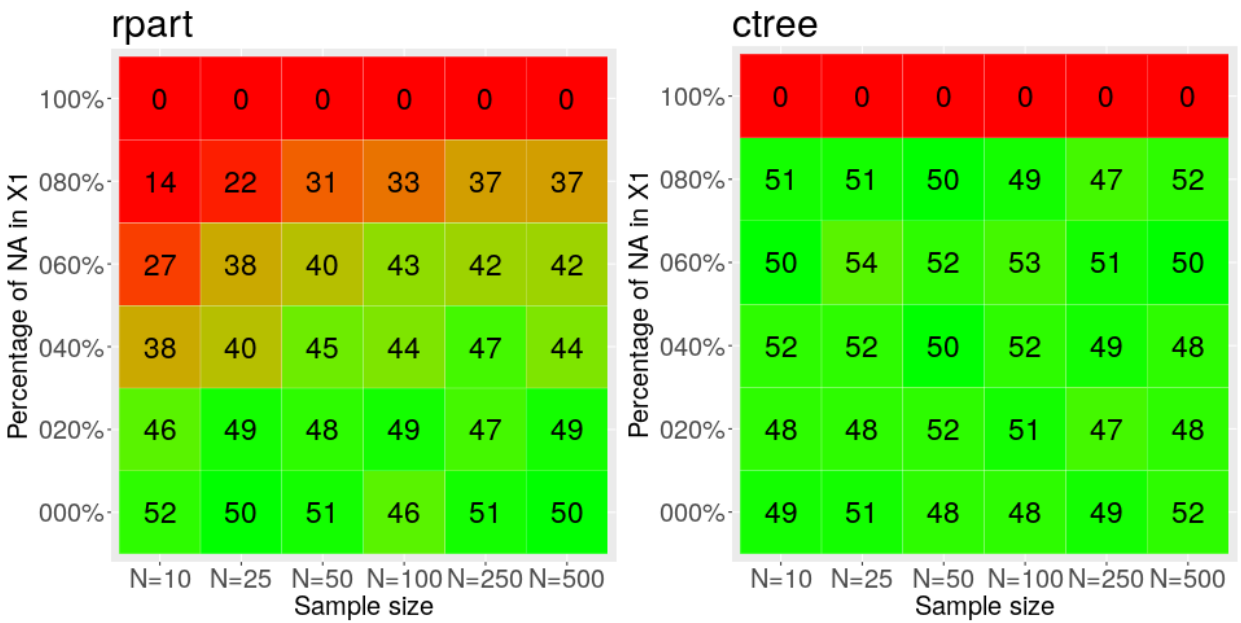


Figure: X_1 selection frequency for `rpart` (CART) vs `ctree` (conditional inference trees)

Missing incorporated in attribute[7]

CART with missing values

Method: Recursively, find which partition $\mathcal{P}$ minimises

$$\mathbb{E}\Big[\big(Y - \mathcal{P}(\widetilde{\mathbf{X}})\big)^2\Big],$$

with, for each feature j and each threshold z, there are three possible partitions,

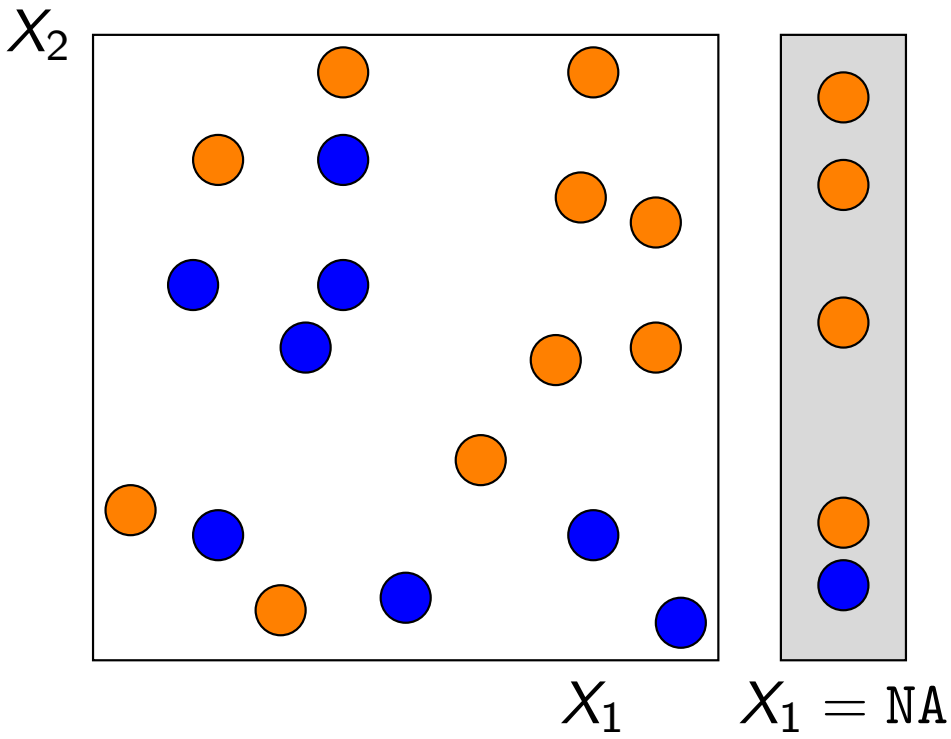
$$\begin{aligned} \{\widetilde{X}_j \leq z \text{ or } \widetilde{X}_j = \texttt{NA}\} \quad & \textbf{VS} \quad \{\widetilde{X}_j > z\} \\ \{\widetilde{X}_j \leq z\} \quad & \textbf{VS} \quad \{\widetilde{X}_j > z \text{ or } \widetilde{X}_j = \texttt{NA}\} \\ \{\widetilde{X}_j \neq \texttt{NA}\} \quad & \textbf{VS} \quad \{\widetilde{X}_j = \texttt{NA}\} \end{aligned}$$

$\rightarrow$ targets $\mathbb{E}[Y|\widetilde{\mathbf{X}}]$

[7] B. E. T. H. Twala, M. C. Jones, and D. J. Hand. "Good Methods for Coping with Missing Data in Decision Trees". In: *Pattern Recogn. Lett.* 29.7 (May 2008).

Missing incorporated in attribute

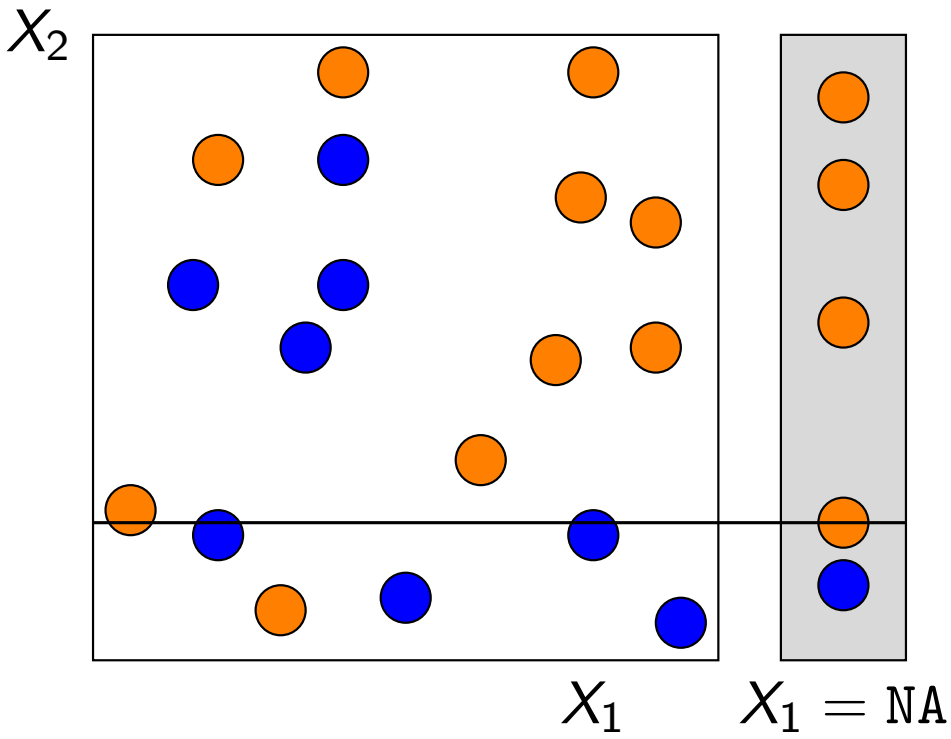
CART with missing values



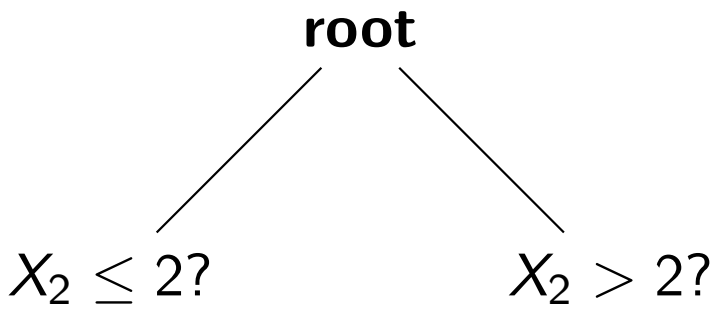
root

Missing incorporated in attribute

CART with missing values

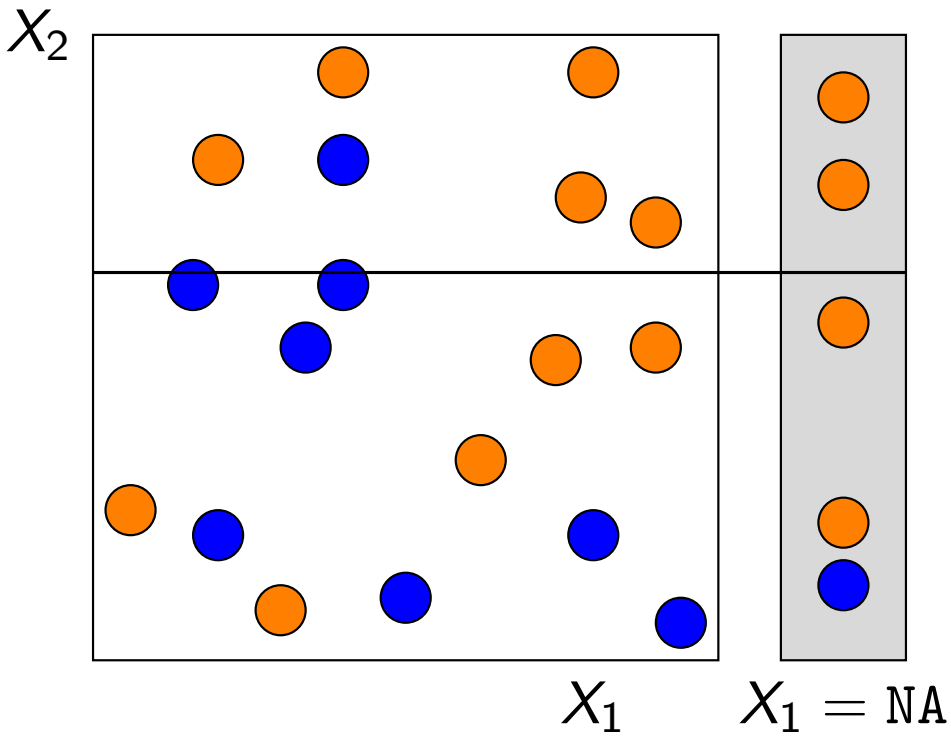


loss $= 4.60$

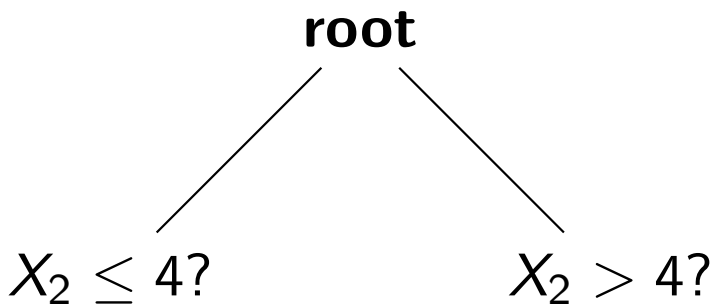


Missing incorporated in attribute

CART with missing values

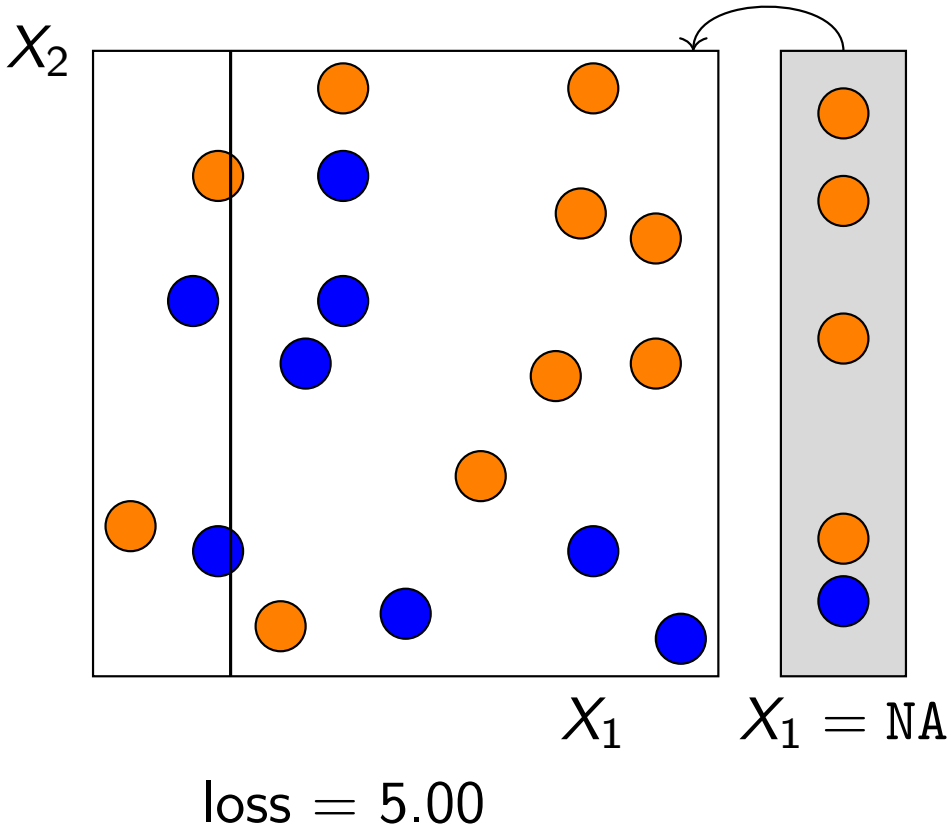


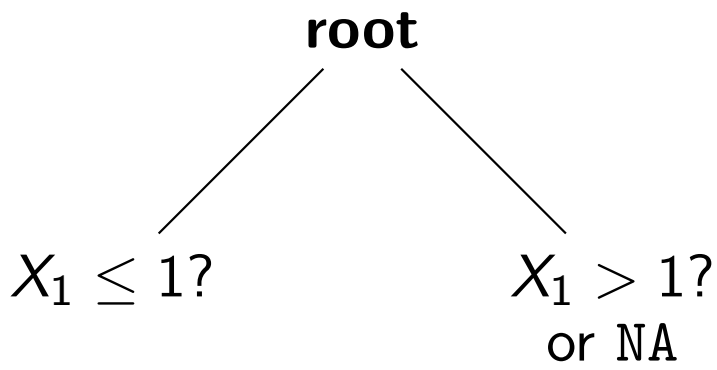
loss $= 4.79$



Missing incorporated in attribute

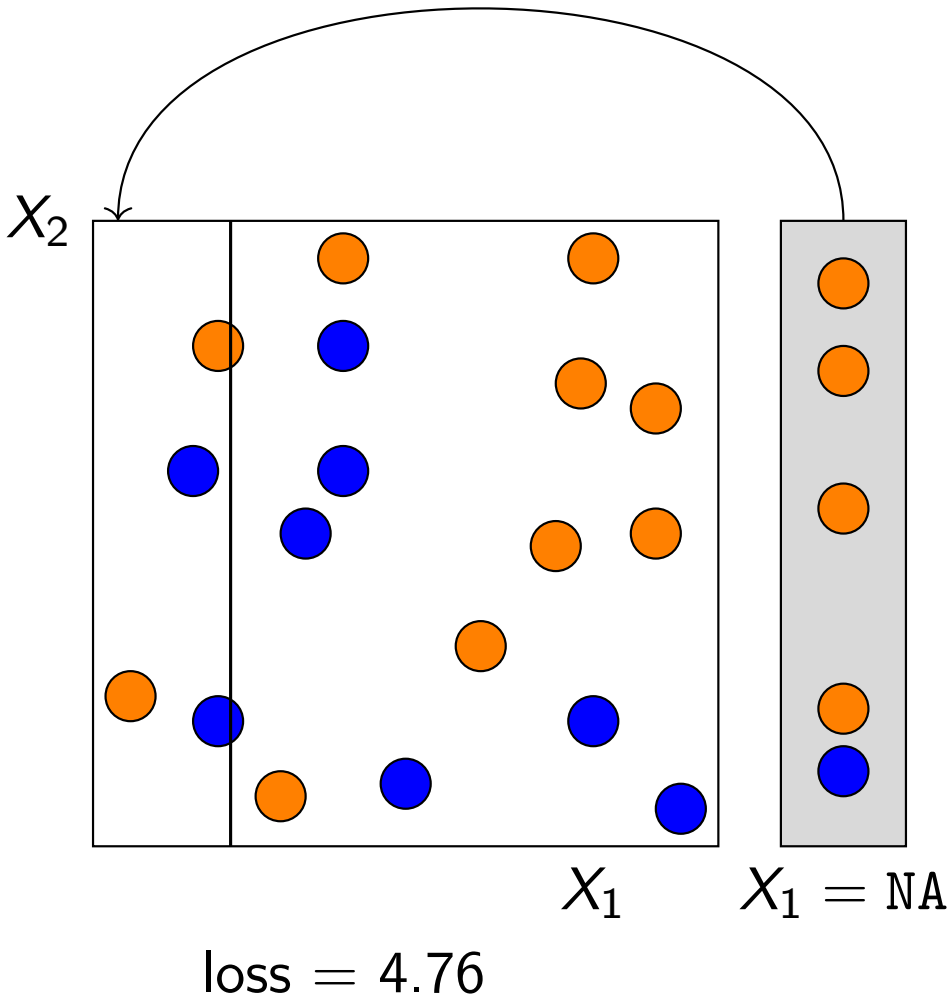
CART with missing values

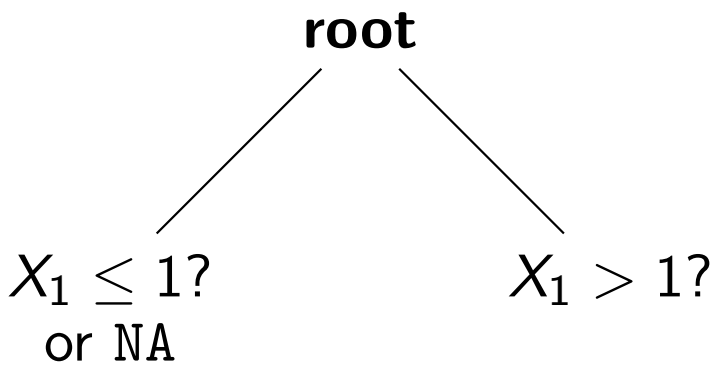




Missing incorporated in attribute

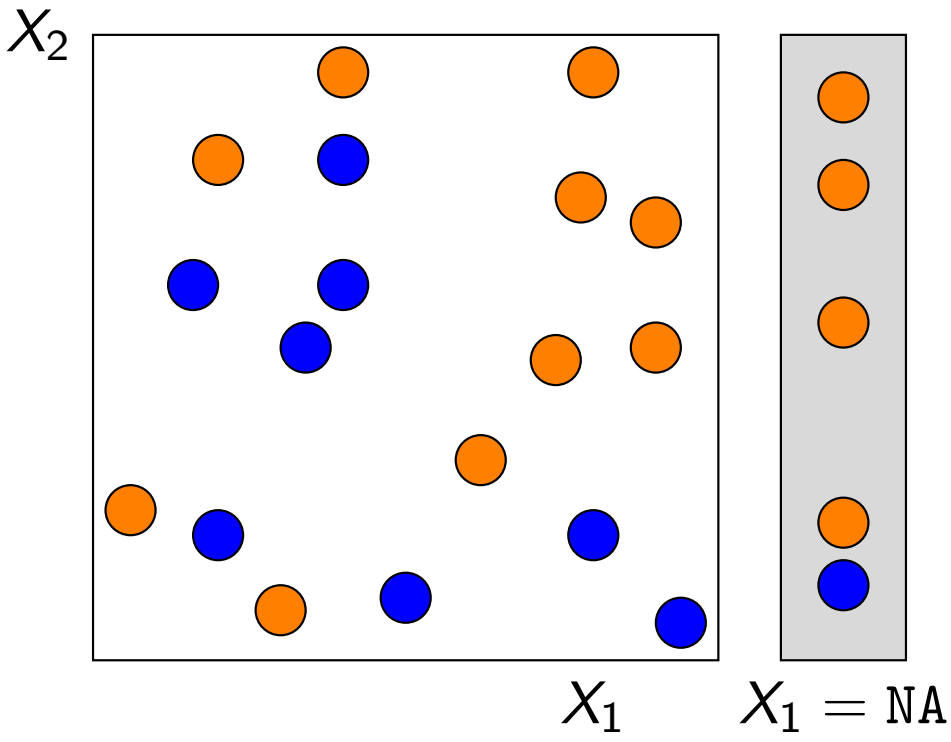
CART with missing values



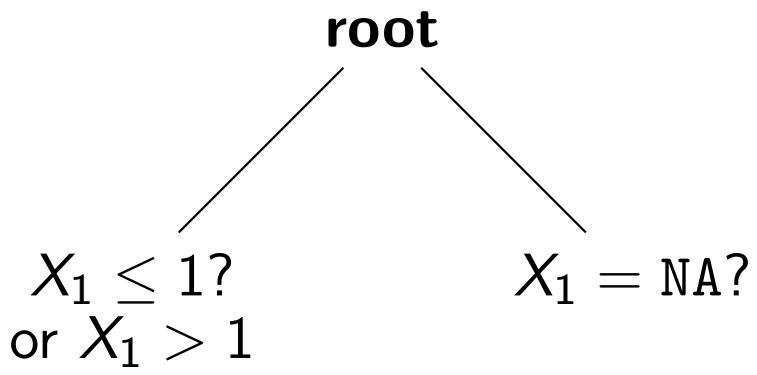


Missing incorporated in attribute

CART with missing values

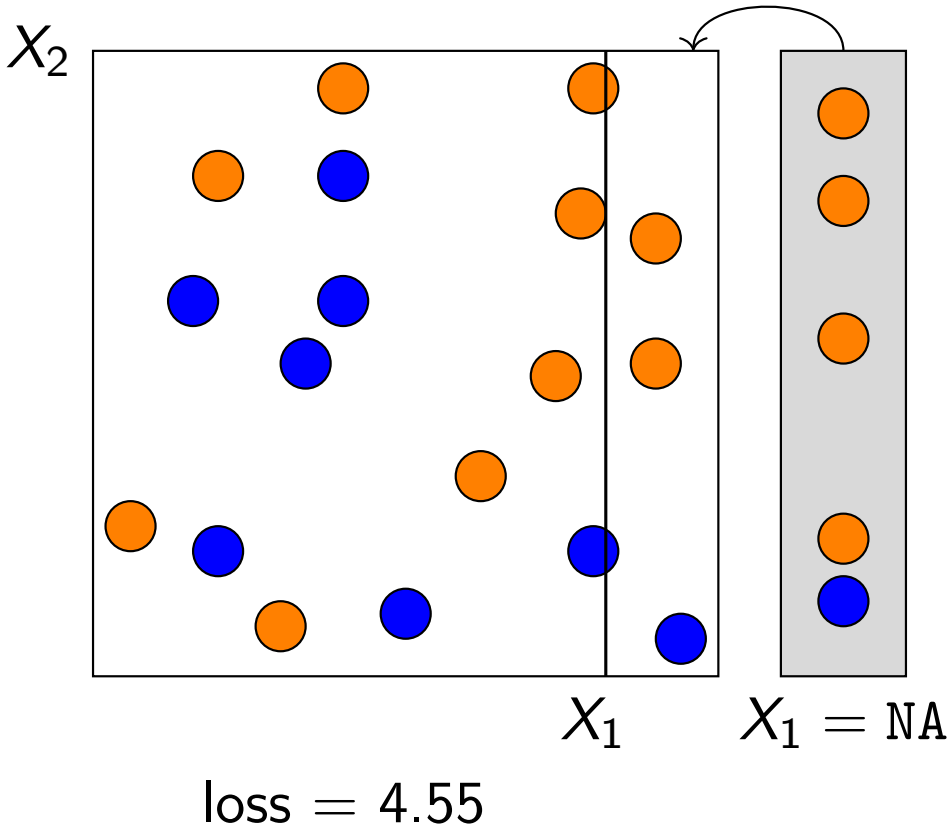


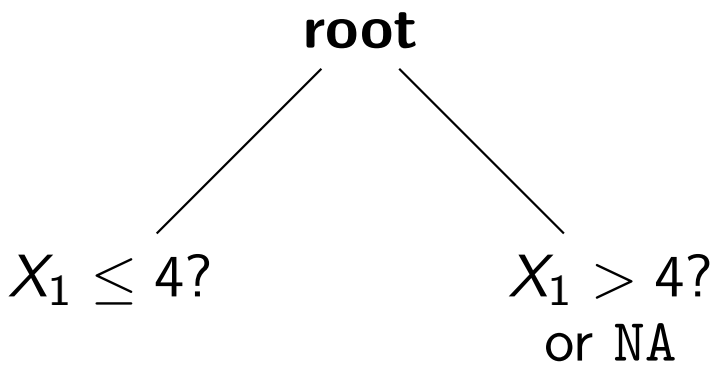
loss $= 4.95$



Missing incorporated in attribute

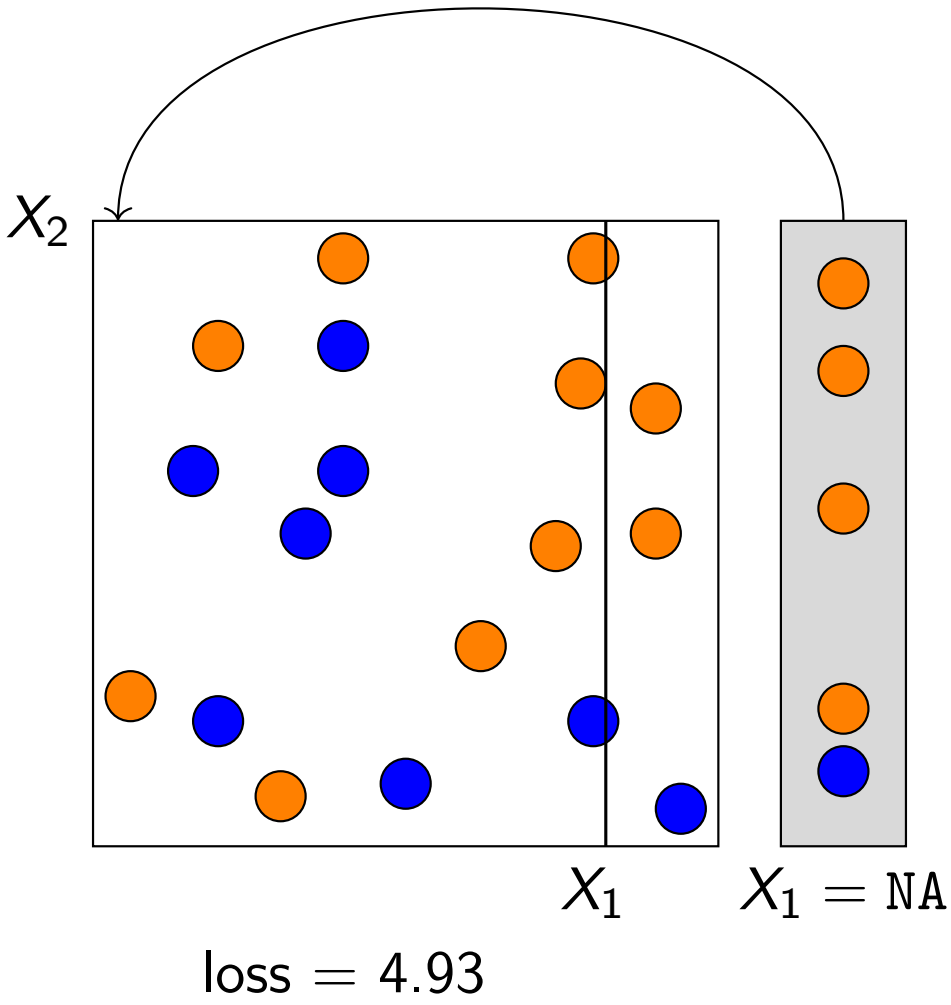
CART with missing values

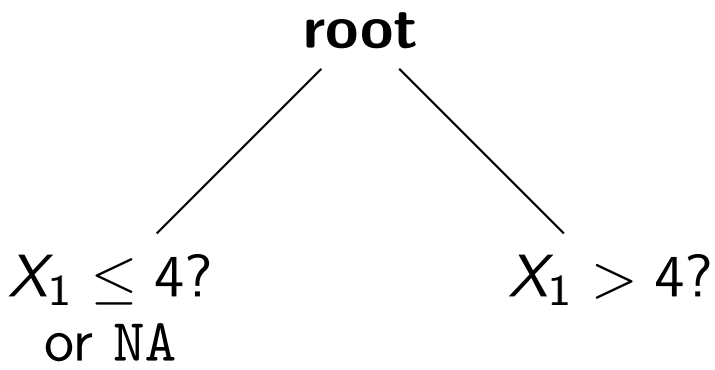




Missing incorporated in attribute

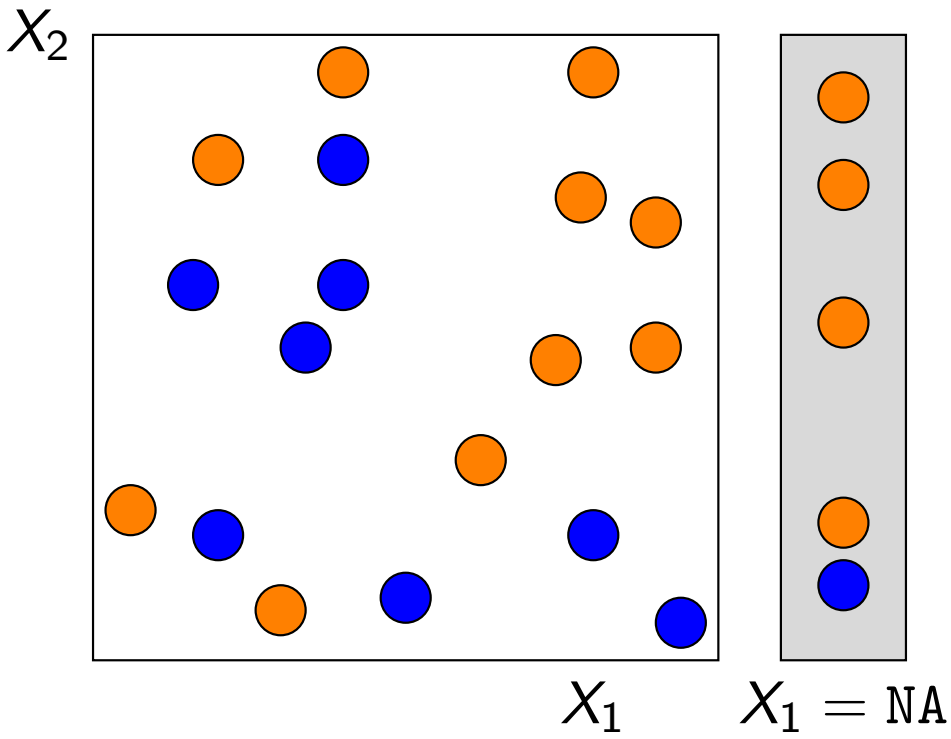
CART with missing values



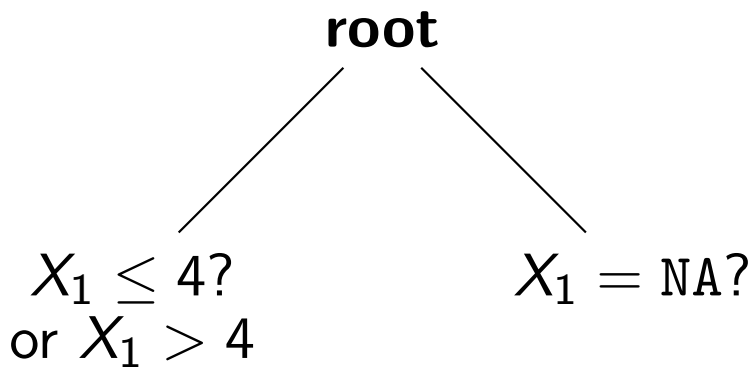


Missing incorporated in attribute

CART with missing values

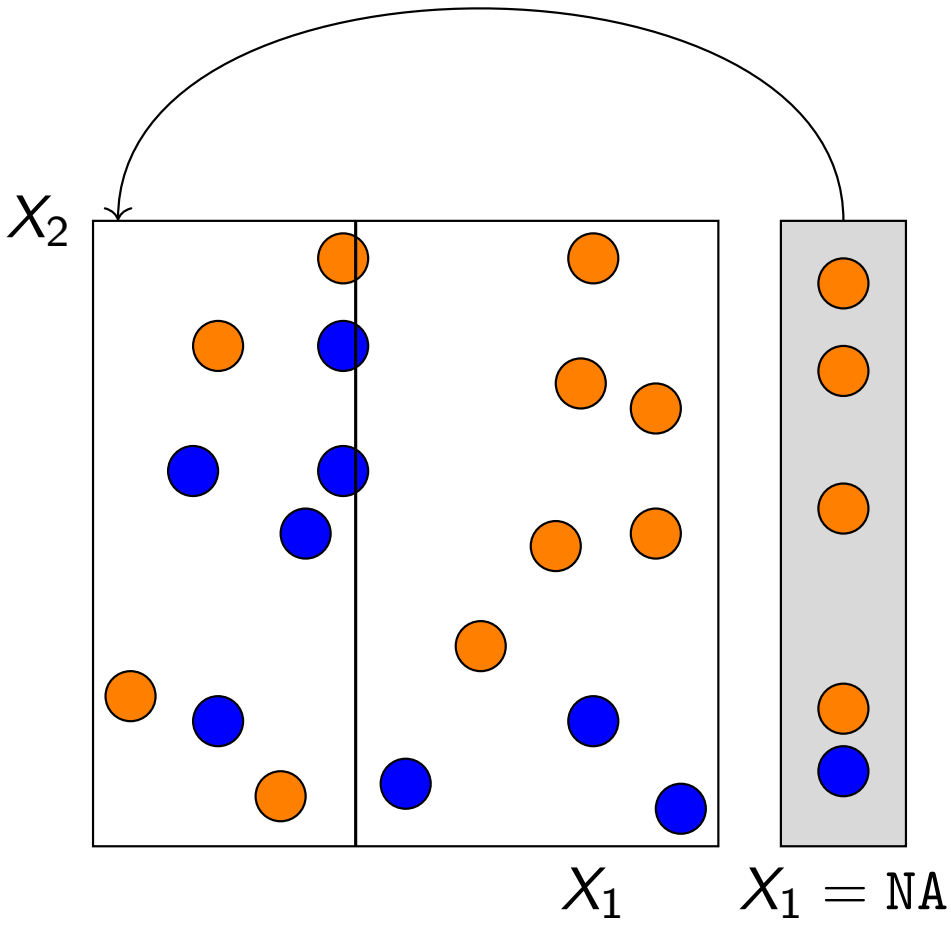


loss $= 4.90$

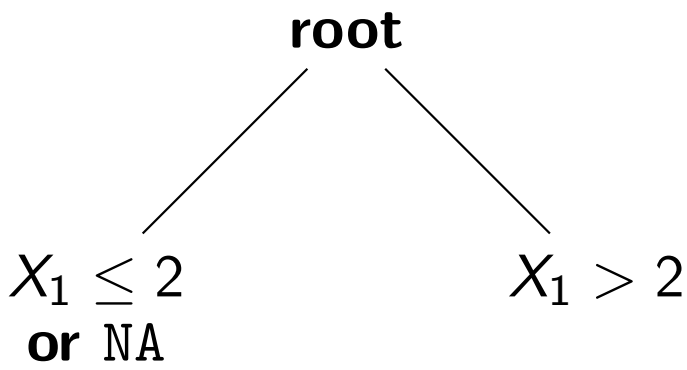


Missing incorporated in attribute

CART with missing values



loss = 4.53



Simulations

First experiment

Models

1. Quadratic: 3 gaussian features, $Y = X_1^2 + \varepsilon$.

Mechanisms

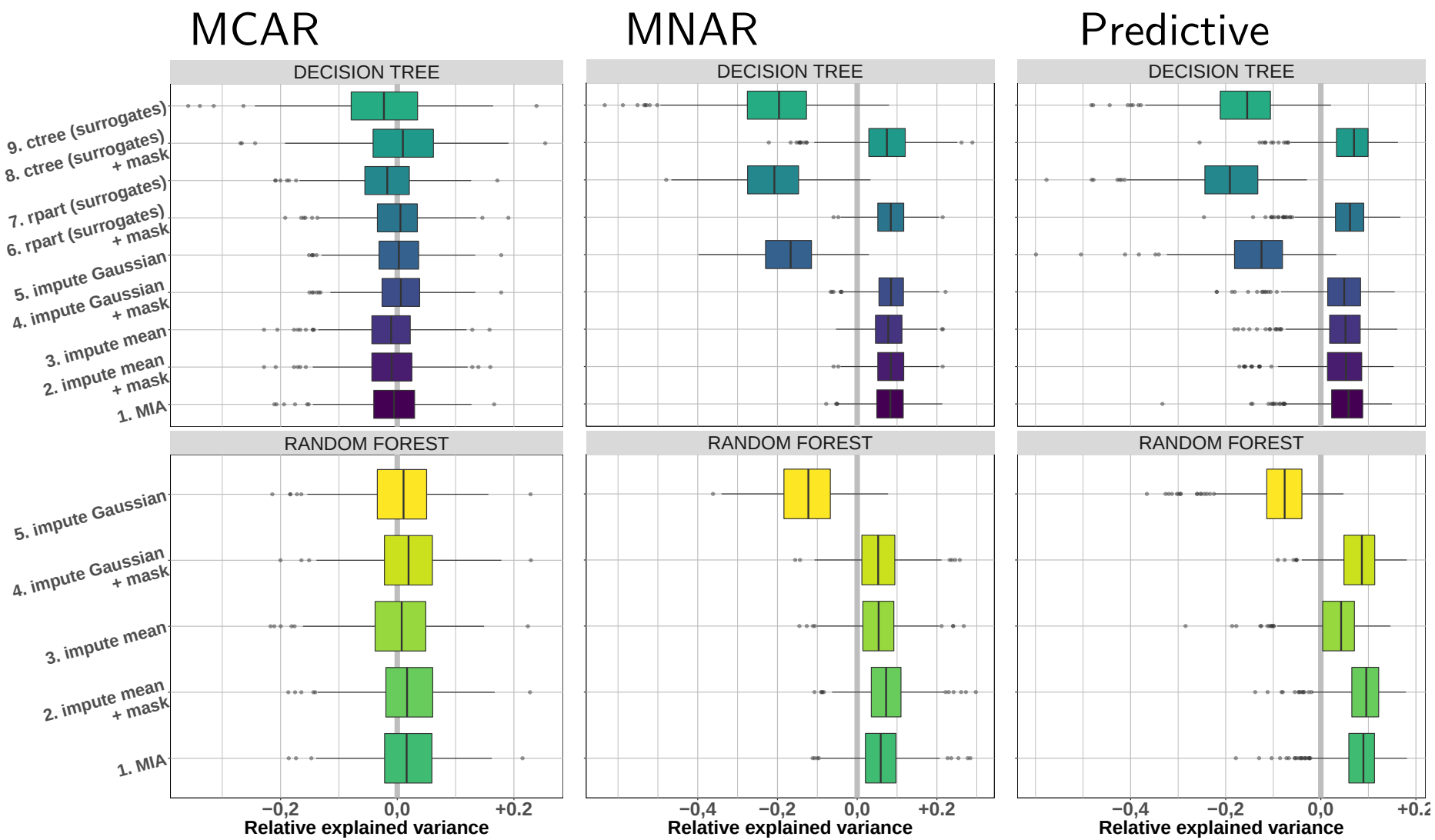
1. Missing completely at random on X_1
2. Missing on large values of X_1
3. Predictive: $\mathbf{M} \perp\!\!\!\perp \mathbf{X}$, $Y = X_1^2 + 3M_1 + \varepsilon$

Methods

1. CART or conditional trees with surrogate splits
2. MIA
3. Mean or gaussian imputation
4. Adding the mask

Relative scores, quadratic model, 20% missing values

First experiment



Simulations

Third experiment

Models

1. Linear: gaussian features, $Y = X\beta + \varepsilon$
2. Friedman: Gaussian feature, nonlinear regression[8]
3. Nonlinear: Nonlinear features, nonlinear regression.

$d = 10$

Mechanisms

1. Missing at random on all ten variables

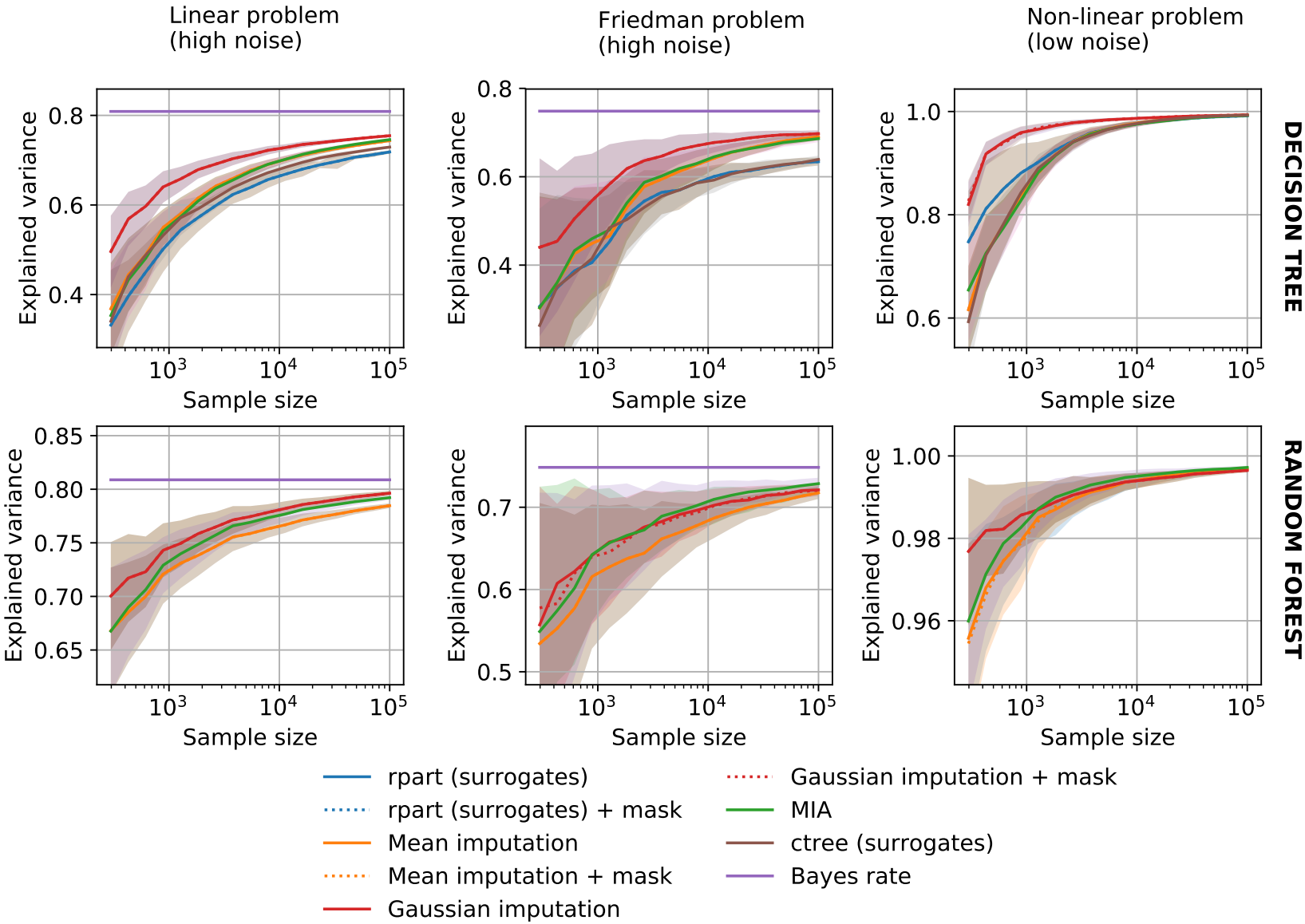
Methods

1. CART or conditional trees with surrogate splits
2. MIA
3. Mean or gaussian imputation
4. Adding the mask

[8] Jerome H Friedman. "Multivariate adaptive regression splines". In: *The annals of statistics* (1991).

Consistency, 40% missing values, MCAR

Third experiment



Conclusion

- Regarding variable selection, conditional inference trees are unbiased;
- this does not solve prediction;
- MIA allows empirical risk minimisation in the "incomplete" space;
- most missing-encoding methods are empirically equivalent in prediction;
- larger discussion on supervised learning with missing values: talk by Julie Josse tomorrow.

Thank you!